Student Evaluation of Teaching in the Chinese Tertiary Education Sector:

Potential Biasing Factors

Yanjun Chen

A Thesis
in
John Molson School of Business
Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Management) at
Concordia University
Montreal, Quebec, Canada

August 2019

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Yanjun Chen

Entitled: Student Evaluation of Teaching in the Chinese Tertiary Education Sector: Potential Biasing Factors

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Management)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Jooseop Lim Chair

Ingrid Chadwick Examiner

Alexandra Panaccio Examiner

Stéphane Brutus Supervisor

Approved by:

____________________________Chair of Department or Graduate Program Director

____________________________Dean of Faculty

Date:_________________

ABSTRACT

Student Evaluation of Teaching in the Chinese Tertiary Education Sector:

Potential Biasing Factors

Yanjun Chen

Student evaluation of teaching (SET), which refers to students' feedback about and evaluation of their professors, is the most frequently used teacher assessment method in the world (Newton, 1988; Seldin, 1989). Despite its popularity—and the fact that it does have its advantages—SET has long been a target of criticism from scholars and educators alike. Since China has the biggest population in the world and its tertiary education sector has grown rapidly in recent decades (Government of China, 2016), the primary purpose of this study is to find out the potential factors that can lead to biases in teachers' SET scores in China. The research for this study was conducted in a middle-sized Chinese university. It involved 1,371 business department undergraduate students and a total of 13,154 evaluations. Pearson's correlation analysis and multiple regression analysis were applied to the data in order to explore the relationship between six different factors—course type, class size, course level, student gender, professor gender, and professor seniority—and SET scores. The results revealed that five out of these six factors (all but student gender) can bias SET scores, but that their ability to do so is highly limited. These results indicate that SET scores can legitimately be used in the Chinese tertiary education sector to improve course quality and teaching quality, but that they cannot, on their own, be used to justify the promotion of professors.

ACKNOWLEDGMENTS

It is not an easy process to complete a thesis, I have a lot of thanks that I want to say to all those people who have encouraged and supported me during these one and a half years period.

First, I would like to express my sincere gratitude to my supervisor Dr. Stéphane Brutus for his invaluable guidance, encouragement, and consistent availability. His positivity and enthusiasm for this thesis make me feel the entire process is very meaningful and not that hard. I have learned a lot of things that I will never expect to learn from the class.

Then I would like to thank Dr. Ingrid Chadwick and Dr. Alexandra Panaccio, members of my thesis committee, who have provided me insightful comments and support.

Last but not the least, I also want to thank my parents and my friends. This would not be possible without their support and encouragement.

A special thanks to Chandler for his patience and constant encouragement during my most stressful times. Also, a special thanks to Curry and Charlie, for their everyday company.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Student evaluation of teaching (SET) is commonly used across tertiary education for the purposes of course improvement and teaching quality assurance (Brockx, Spooren, & Mortelmans, 2011; Macfadyen, Dawson, Prest, & Gašević, 2016). Additionally, many universities also use SET as one criterion on which to base their faculty promotion decisions (Brockx et al., 2011). However, although SET is the most used teaching evaluation system in the world (Badri, Abdulla, Kamali, & Dodeen, 2006), it has been controversial ever since it was first introduced in the United States during the 1920s (Marsh, 1984). Since then, a great number of studies have challenged the validity, reliability and diagnostic power of SET (Kember & Leung, 2009). Moreover, beyond the possible problems associated with the evaluation instrument's design, there are rising concerns shared by many institutions about whether students take course evaluations seriously (Clayson & Haley, 2011). Recently, for example, the University of Southern California announced that it would no longer use SET for making promotion decisions because the school had found that SET tended to favor faculty members of specific genders and backgrounds (Flaherty, 2016). In addition to the aspects of gender and background, some empirical investigations have also found that some other external factors could also affect SET results (Badri et al., 2006; Brockx et al., 2011; Macfadyen et al., 2016). For instance, researchers have argued that SET was biased by the characteristics of the students who wrote the evaluations, such as student gender, school year, and grade, and also by the characteristics of the courses that were being evaluated, such as course type and course level (Beran & Violato, 2005). Hence, since SET has been given considerable weight in faculty promotion decisions (Beran, Violato, Kline, & Frideres, 2005), it is essential to investigate potential biases in the SET process that may influence the results of these evaluations and the degree to which they might affect them.

The use of SET around the world is chiefly driven by its use in the US education system. In the 1920s, SET was first implemented in the US tertiary education system (Mueller, 1951). Then, in the 1970s, SET was applied widely for formative purposes in the United States, such as improve and shape the quality of teaching (Hornstein, 2017). Since then, SET has become the primary tool used in North American universities and postsecondary institutions for summative teaching evaluation: it has been used in order to gain an overall picture of professors' teaching performance, and it has been used by tenure committees to make promotion decisions. Accordingly, most of the previous studies on SET were conducted in North America (Dev & Qayyum, 2017). However, since people's values and priorities vary from society to society, the factors that are influential in one culture may not have the same influence in a different culture (Tarman & Acun, 2010). As a consequence, the factors that can influence SET results in the West may have different impacts on SET results in the East. For example, according to a study conducted in France by Boring (2017), male students gave a significantly higher overall score to their male professors than to their female professors. Moreover, the score that male students gave to their male professors was much higher than the score that female students gave to both their male and female professors. The author explained these results by stating that it was male students were more likely than their female counterparts to assign the "excellent" score to their male professors. However, another recent study conducted in the Middle East had a totally different result. Thawabieh (2017) found that female students' evaluations were higher than male students' evaluations, and he argued that this result had two possible causes: either the female students were more motivated and more familiar with the learning environment than the male students, thus leading to them being more accurate their evaluation of their teaches, or the female students were more sympathetic than the male students, thus leading them to inflate their scores. This opposing finding indicates that different cultural contexts can possibly influence male and female students' perspectives on their professors, which can also impact SET results. Moreover, according to a study conducted at the United Arab Emirates University by Badri and colleagues (2006), the time and day when a course was taught can also influence SET scores. They pointed out that many students perceived the courses scheduled late in the afternoon or in the evening as being less effective because of the fact that they were usual tired after a long day at school. However, Koh and Tan (1997) found that, in a Singaporean university context, teachers of courses scheduled late in the afternoon or in the evening received better teaching evaluations. They speculated that the students in their sample might have perceived that there was a more relaxed atmosphere in the classroom at the end of a day. All of the above examples illustrate the likelihood that, in different cultural contexts, even the same factors can have the opposite effect on SET scores. Consequently, in order to apply SET as an evaluative tool in different cultural contexts, all the factors that can potentially influence SET scores should be considered beforehand.

For the present study, China has been selected as the target cultural context in which to study SET scores. It was chosen for three main reasons. First, there is a big cultural difference between China and the United States. Based on Hofstede's theory of cultural dimensions, Fernandez, Carlson, Stepina, and Nicholson (1997) used the multi-item scales that were developed by Dorfman and Howell (1988) to measure the cultural difference between China and the United States. They found that the United States can be considered as an individualistic society with a score of 13.41 (M = 11.55), while China can be considered as a collectivistic society with a score of 10.38 (M = 11.55). Additionally, China also had a power distance score that was above the mean (14.50 / M = 12.72), which indicated that power was distributed unevenly across Chinese organizations and across Chinese society more broadly. The United States, however, had a power distance score that was below the mean (12.70 / M = 12.72), which indicated that power was distributed relatively evenly across US organizations and US society more broadly.

Since China is a collectivistic society and has a larger power distance score than the United States, there is a considerable cultural distance between the two countries. This large cultural distance may lead students in China and the United States perceive and understand their professors' performance rather differently. Specifically, Liu, Keeley, and Buskist (2015) have discussed in their study that the larger power distance in China influences Chinese students' perceptions of what factors an excellent professor should possess. They found that, when compared to students from the US, Chinese students were less likely to endorse the importance of their professors' accessibility, flexibility, and respect, but they were more likely to care about their professors' competence in their field of expertise. Oppositely, since there is a smaller power distance in the United States, students in the United States are more likely to prioritize professors' accessibility, flexibility, as well as the qualities of promoting critical thinking and providing constructive feedback to determine their performance (Liu et al., 2015). This result indicates that professors in Chinese education are always put in the central place (which will be discussed in the next part), while the interaction and the primary role of professors are emphasized in the United States education sector. Consequently, in contrast with North America, where students and professors are usually in an equal position, professors in China are traditionally viewed as higher than their students. The large power distance in China makes people accept and also get used to being not close to their higher-level leaders. Therefore, Chinese students should not be too concern about their professors' accessibility, flexibility, and respect. Even though the big cultural distance between the United States and China may cause their students to have some different perceptions toward their professors' performance, no research has shown that the cultural issue can influence all student's cognition on their professors' performance.

Second, the cultural background of Chinese students is unique. Confucianism stresses absolute compliance with and respect for authority, and these tenets have had and continue to have a strong influence on the relationship between Chinese professors and their students (Ting, 2000). In such a cultural context, professors are often put on a pedestal, as it were, and placed at the center of the education system; arguing with a professor, would be considered rude and disrespectful in China, presumably even if the professor is mistaken (Guo, 1996; Louie, 1984;

Rao, 1996). In the United States, in contrast, students are placed at the center of the education system (Guo, 1996; Rao, 1996). Professors in the United States have been making extensive efforts to engage students with question-based learning in the classroom, since they believe that this approach can better help students develop scientific reasoning skills and the ability to better construct and defend arguments (DeBoer, 1991; Duschl, 1998). Therefore, it is not surprising that, unlike American students, Chinese students today are very humble, show little initiative, and obey professors without question (Guo, 1996). As a result, when students in China are asked to evaluate their professors, they may find themselves in a unique position: should they evaluate their teachers accurately or should they be deferential to their teachers and perhaps color their evaluation accordingly? This is an especially tough situation, given the fact that some students may need to collaborate with their professor on future projects and/or take future classes with him or her.

Last but not least, China is now seeing significant growth in its tertiary education sector. China implemented a curriculum reform in the fall of 2001, which was designed to overhaul the traditional education system, which had overemphasized knowledge delivery and “passive learning,” and introduce a new education system that would stress diversity in teaching and “initiative learning” (Liu & Teddlie, 2003). It is therefore important to realize that new teaching evaluation methods have only been under development for 18 years, and so it is fair to say that the use of SET in China has, therefore, not yet reached its full potential (Liu & Teddlie, 2003). It will take some time for lasting change to solidify. However, according to 2016 China Census, 28 more universities (colleges) were operated in 2015 compared to the previous year

(Government of China, 2016). Moreover, more than 7.38 million students were enrolled in tertiary education in 2015, which was 160,000 more than during the previous year. The number of school faculty and staff members in tertiary education in 2015 was 2.37 million, which was 33,600 more than in 2014 (Government of China, 2016). Considering the rapid growth of the Chinese tertiary education sector, a study investigating the use of SET in China can help Chinese college and university administrators apply the tool in the most effective way possible.

In summary, the primary purpose of this study is to clarify and discuss the factors that can influence Chinese college and university students' perceptions of their professors' teaching performance and teaching quality. The cognitive appraisal model designed by DeNisi (1996) will be majorly applied in this study to develop the relationship between each potential biasing variables and SET scores. Since SET studies have rarely been conducted in the context of the Chinese tertiary education sector, this study can not only fill the research gap in SET studies, but also help the development of SET practices in China.

Professor Performance

So far, a growing number of theoretical and practical scholars have begun to conceptualize professors' performance in SET terms (Cai & Lin, 2006). Commonly, job performance is defined as the degree to which an individual employee executes a particular role or responsibility that is based on certain standards (Nayyar, 1994). According to this definition, researchers have defined teaching performance as the execution of a set of pedagogical tasks or behaviours that are consistent with the educational objectives of a given course (Cai & Lin, 2006; Marsh, 1987). Since teaching is a complicated process and professorial behaviours are too, researchers have yet to come to a consensus regarding which professorial behaviours should be concretized or generalized in determining whether a university or college professor's performance is poor, adequate, or excellent (Cai & Lin, 2006).

Generally, there are two major categories of job performance: *task performance* and *contextual performance* (Bakker & Bal, 2010; Cai & Lin, 2006; Carson, 2006; Min, 2007). Task performance refers to employees' behaviours that are positively linked to their organizations' goals or objectives (Khan, Gul, Shah, & Khan, 2012), and these also include the technical tasks and activities that employees perform on a daily basis (Borman & Brush, 1993; Griffin, Neal, & Neale, 2000). For professors, task performance consists of teaching effectiveness, professor–student interaction, and teaching value (Cai & Lin, 2006; Khan et al., 2012). Teaching effectiveness mainly refers to the behaviours that show how professors prepare and organize their classes, as well as how they deliver their subject material. Professor–student interaction refers to the how professors communicate and interact with their students both inside and outside the classroom. Finally, teaching value refers to the positive outcomes that are achieved by students, such as grade improvements or a heightened interest in the course material, that comes as a direct result of their professors' efforts (Cai & Lin, 2006).

Contextual performance refers to employees' activities that do not directly contribute to but nevertheless support their organizations' goals or objectives by positively influencing the social and/or psychological environment of the workplace (i.e., the place where the goals or objectives are followed) (Borman & Brush, 1993). Professors' contextual performance consists of occupation morality, job dedication, and collegiality (Cai & Lin, 2006; Khan et al., 2012). Cai and Lin (2016) have explained that, for professors, occupation morality mainly refers to professors' behaviours that are consonant with the institution's code of conduct, that reveal a deep concern for their careers as educators, and that show a deep passion and enthusiasm for teaching and a willingness to be responsible for their students. Job dedication refers to the ways in which professors can reflect on their teaching methods and activities, the ways in which they continue to hone their teaching skills, as well as to the ways in which they keep up with changes in their fields of expertise. Collegiality (literally "assistance and cooperation") refers to how professors interact with their colleagues, with administrators, and with students' parents. In short, how well do they seek the collective well-being of all the stakeholders involved in the educational process?

In 1999, Conway stated that employees' task performance and contextual performance played equally essential roles in contributing to employees' overall performance. However, when it comes to the college or university classroom in China, professors' task performance, which has to with how well they communicate with their students and how well they organize the class, is more likely to be noticed by students than professors' contextual performance, which has to do with how well they reflect on their experience and how well they work with their professorial and professional colleagues. Consequently, as opposed to professors' task performance behaviours, which can be easily observed by students in the classroom, professors' contextual performance behaviours are more related to behaviours that they exhibit outside the classroom, which, for obvious reasons, cannot be easily observed by students (Cai & Lin, 2006).

Since SET as a tool is based on students' perceptions of their professors' performance, the incomplete observations of students may lead to some biases in their evaluations. Hence, in contradistinction to the traditional evaluation system in China, which focuses more on professors' contextual performance—measuring such things as morality (e.g. whether the professor has set a good model to the students), and diligence (e.g. whether the professor present at school during school hour of each day, and attend the required meetings) (Feng, 2002; Jiang & Zhang, 2003; Li, 2002; Li & Xuan, 2003; Ying & Fan, 2001)—the modern SET tool in China is more similar to the SET in North America, which pays more attention on professors' task performance behaviours, especially on measuring professor's teaching effectiveness (Liu & Meng, 2009; Wright & Jenkins-Guarnieri, 2012).

Overall, SET nowadays focuses more on professors' task performance than on their contextual performance not only in China, but also in North America. However, SET scores only represent a part of professors' performance which do not give the big picture. Therefore, in order to apply SET all over the tertiary education sector in China, the state ought to know just how effective the tool is and what its weakness are, or else face the possibility of a biased process and a waste of funds.

The Reliability and Validity of SET

Although SET has been used in North America since the 1920s (Mueller, 1951) and more than 2,000 relevant published articles on the topic are listed in popular online databases (Centra, 2003), some researchers still question its reliability and validity (Badri et al., 2006). In this section, both the reliability and the validity of SET will be discussed.

Reliability is defined as the "consistency" or "repeatability" of a measure (Haertel, 2013). In the context of teaching evaluations, when professors' ratings are consistent across various categories, the ratings are said to be reliable (Chen & Hoshower, 2003). "Are student ratings consistent both over time and from rater to rater?" is the major question that has been addressed by most reliability studies (Marlin & Gaynor, 1989; Nimmer & Stone, 1991). In his book on SET, Page (1974) pointed out that one of the major concerns about the tool's reliability was the uncertainty of whether or not students' judgments of their professors were objective and rational. Then, he explained that this uncertainty was reasonable because even the judgment of the most intelligent human could still be influenced by internal or external factors. In order to examine the reliability of students' judgments, Page (1974) reviewed the related SET studies from 1927 to 1970, and his results showed that students' judgments about their professors were generally reliable. However, Feldman (1983) questioned the reliability of SET, and he pointed out that these evaluations were negatively related to age and years of teaching experience. Then, Marsh and Hocevar (1991) conducted a longitudinal study of the changes in SET scores of 195 professors that had been evaluated continuously over a 13-year period. The result showed no systematic change in professors' SET scores, which also proved that students' ratings were consistent both over time and from rater to rater. Consequently, consistent with the result of Page's (1974) study, Marsh and Hocevar (1991) also suggested that SET conducted in their study was a reliable tool. Although most of the researchers have suggested the SET that they applied in their studies is reliable on measuring professors' performance, DeNisi (1996) has indicated in his cognitive appraisal model that the "nature of the rating instrument" can impact raters' evaluation process. Therefore, the result can be concluded that most SET is reliable on evaluating professor's performance; however, the different nature of the rating instrument can lead each SET system to have a different level of reliability.

As for validity, this concept refers to the extent to which a measure quantifies a construct (Messick, 1989). With regard to SET, if student ratings can reflect the instructional process and the instructional consequences, the ratings are said to be valid (Abrami, d'Apollonia, & Cohen, 1990). "Does SET measure teaching effectiveness?" and "Are student ratings biased?" were the major questions that have been addressed by most validity studies to date (e.g, Howard, Conway, & Maxwell, 1985; Tagomori, & Bishop, 1995). In 1979, Seibert stated that students usually preferred to give higher ratings to professors from whom they learned the most. This statement indicated that the higher ratings would be given to those professors who had excellent teaching performance and who could deliver more effective classes. Consequently, Seibert's (1979) study supported the notion that SET was valid with respect to predicting teaching effectiveness. Moreover, in 1981, Cohen used a meta-analysis to synthesize research on the relationship between student ratings of instruction and student achievements. The results of this study showed that students of professors whose classes were proceeding on-schedule, who used class time well, and who generally had the class well organized tended to learn more than students of professors whose classes were disorganized and who did not manage class time efficiently. Furthermore, the results of this study also suggested that there was a strong tendency for students to give a high rating to professors who they learned the most from. Therefore, Cohen's (1981) study provides strong support for the validity of SET as a means by which to measure teaching effectiveness. Moreover, Shevlin, Banyard, Davies, and Griffiths (2000) also pointed out that in their study, SET can significantly predict teaching effectiveness. Later, Marsh (2007) found evidence from a variety of different studies, which supported that SET can effectively measure professors' performance. Consequently, although SET is commonly recognized as an effective measurement of professors' teaching performance in some studies, researchers have pointed out that there are still potential response biases in the tool. As mentioned above, there are two major categories of teaching performance, and professors' task performance is much easier for students to observe than their contextual performance. This observational constraint may lead to the presence of some biases in SET scores, since students cannot completely observe all their professors' behaviours. Beran and Violato (2005) conducted a research study in Canada, and they found that, consistent with their previous results, student grade expectations, class attendance, and course categories could all influence SET scores. Moreover, in 2007, Pounder proposed an analytical framework for future SET researchers, which contained 11 potential biases: student gender; student academic level and maturity; low grade retribution; grading; class size; course content; class timing; professor gender; professor age, experience, and rank; professor's influencing tactics; and professors' behavioural traits. Consequently, even though many researchers have given an affirmative answer to the question "does SET measure teaching effectiveness?" researchers are still not in a position to confirm that such an evaluative tool is free from major sources of bias. And this is why the validity of SET still needs to be discussed.

In summary, SET has been applied for almost a century, and most research has supported its reliability. However, some concerns remain as to its validity and to the potential biases that can influence SET scores. Therefore, a further investigation into the factors that can cause those biases in different cultural contexts should be conducted to improve SET's validity.

SET in China

At the end of the first decade of the 21st century, the Chinese government published a report entitled "Outline of National Medium- and Long-term Program for Education Reform and Development (2010–2020)." The report stated that 2,852 universities or colleges, with 3.65 million full-time students in tertiary education institutions, existed in China in 2015 (Government of China, 2016). However, research has shown that the population of people with a tertiary level education (25 to 64 years old) in 2017, as a percentage of the total world population, is 36.9%, while the population with a tertiary education (25 to 64 years old) in China for the same year only made up 15.8% of the total Chinese population (OECD, 2019). These numbers indicate that only a small proportion of the total Chinese population has been enrolled in tertiary education at some point in their lives. As a consequence, the president of China has emphasized the need for the expansion of tertiary education offerings to improve the overall level of Chinese education (Xinhua News Agency, 2010). As of the fall of 2001, China has undertaken a major reform of its curriculum. The purpose of this reform was to change the traditional education system in China, which placed much emphasis on knowledge delivery and “passive learning” (Liu & Teddlie, 2003). As a result, in the new era of Chinese education, diversity in teaching and students’ active participation in the classroom have been and still are being emphasized. Before the Cultural Revolution (1966–1976), the traditional Chinese evaluation of professors had four major characteristics (Liu & Teddlie, 2003; Liu & Teddlie, 2005; Ying & Fan, 2001). First, the evaluation was designed to differentiate professors from one another. Specifically, decisions about rewarding or penalizing professors were made based on their individual performances. Second, the evaluation criteria were abstracted and not based on actual teaching situations, since they were overly dependent on students’ test scores. Third, the traditional evaluation methods in China emphasized numerical indicators, regardless of whether they were suitable for quantification. And fourth, the evaluators were primarily school leaders, especially principals. These four characteristics of the traditional Chinese system of professor evaluation persuaded the professors to only focus on improving students’ grades on examinations and on leaving a lasting positive impression on their superiors. These evaluations did not encourage professors to train students to ask questions and come to conclusions on their own. They did not encourage professors to teach students to be resilient and to think creatively, that is, “outside the box” (Liu & Teddlie, 2005). Hence, in the new curriculum reform, how to go about establishing a new evaluation system to replace the old one became a heated question for the Chinese educational community. Therefore, following the curriculum reform of 2001, several modifications were made to the evaluation system. First, professional development was identified as the main objective of the evaluation system. Second, the new evaluation system was based on assessing the professor's daily performance rather than students' test scores. Third, the evaluators were not restricted to school leaders, as other evaluators were included in the system. Finally, both qualitative and quantitative indicators were considered (Liu & Teddlie, 2005). Nowadays, the effective teaching indicators in China usually contain three, four, or five major domains (Liu & Meng, 2009). For instance, Jiang (2001) introduced a five-domain teaching evaluation system in his study. The domains are (1) teaching objectives (i.e., present clear objectives of the lesson); (2) teaching content (i.e., present the content of the lesson in a logical sequence); (3) teaching methods (i.e., create contexts for learning and promote the interests of the students, and provide timely feedback); (4) teaching processes or skills (i.e., deliver accurate knowledge, demonstrate a natural and elegant demeanour while teaching); and (5) teaching effects (i.e., secure students interest in the lesson). All of these five domains have as their major focus the measuring of professors' task performance and not the measuring of their contextual performance. Since SET is now widely implemented in Chinese universities, the tool has become one of the most common teacher evaluation systems in China (Liu & Teddlie, 2003). Consequently, it is essential to explore the potential biases that can influence the result of SET in China.

THEORETICAL FRAMEWORK

Cognitive Appraisal Model

In order to study SET in the Chinese tertiary educational sector, it is important to first understand the psychological behaviour of Chinese students in evaluating their professors. For many years, the majority of research studies on performance appraisal have focused on rating scale formats and rater training, which were intended to eliminate psychometric errors (DeNisi, 1996). In this context, most researchers assumed that, in comparison to the ratees, the raters in the evaluation process became the passive participants therein, since they were required to provide evaluations to others. Although raters were responsible for providing the evaluations, they were sometimes not motivated enough to provide accurate and truthful judgments. Therefore, it is essential to develop techniques that can motivate raters to give more accurate evaluations of their teachers (DeNisi, 1996). This implies that in real life, organizations view understanding how raters form impressions of and make inferences about other people in interpersonal and social environments as the most fundamental part of understanding performance appraisals (Govaerts, Van de Wiel, Schuwirth, Van der Vleuten, & Muijtjens, 2013). Similarly, in tertiary education, understanding how students rate their professors is also the most fundamental problem in understanding SET. Hence, it is significant to figure out precisely which factors can influence students' decision-making processes.

That raters are the information processors and that the behaviour observation process plays a prominent role in the performance evaluation process are the central claims of the cognitive appraisal model (DeNisi, 1996; Feldman, 1981; Govaerts, Schuwirth, Van der Vleuten, & Muijtjens, 2011). The model argues that "performance appraisal is an exercise in social perception and cognition embedded in an organizational context requiring both formal and implicit judgment" (DeNisi, Cafferty, & Meglino, 1984, p. 362). The model also demonstrates that six major steps should occur in an evaluation process: behaviour observation, information interpretation, information storage, memory retrieval, information integration, and decision making (DeNisi, 1996). Although the model clearly explains the cognitive process of performance evaluation, some researchers have insisted that it is impossible for raters to observe ratees' every job-relevant behaviour in the first step. For instance, as mentioned above, it is easier for students to observe professors' task performance in class; however, it is hard for them to observe their professors' contextual performance (i.e., their behaviours outside of the classroom), because in China, only the representative of the class can have more access to the professors outside class. Therefore, DeNisi (1996) pointed out that the first step—observing a ratee's job-relevant behaviour—is the most critical for the accuracy of the entire evaluation. To reduce the bias caused by the behaviour-observation process, DeNisi and his colleagues (1984) proposed four major factors that can influence it: (1) preconceived notions; (2) the purpose of the appraisal; (3) time pressures; and (4) the nature of the rating instrument. Since this study is based on the students' feedback of a university-designed questionnaire, the factors of the purpose of the appraisal, time pressures, and the nature of the rating instrument are generally identical for each student. Hence, for the present study, preconceived notions represent the major factor that will be used in the development of its hypotheses.

Furthermore, since SET differs from traditional performance appraisal systems where students are asked to evaluate their professors instead of having supervisors assess their subordinates, studies have rarely applied the cognitive appraisal model to study students' behaviour in the evaluation process. According to the cognitive appraisal model, the evaluation process begins with the behaviour-observation process, while the outcome of this process can determine the nature of the final evaluation (DeNisi, 1996). However, researchers have also assumed that raters cannot observe the entirety of ratees' job-relevant behaviours, since conflicting demands on their attention could exist or various physical constraints might not be avoidable (DeNisi, 1996). This indicates that in SET students' evaluations of professors are usually influenced by professors' immediate behaviours, such as how well they organize the class and whether they can effectively communicate with their students, which cannot accurately and completely represent the professors' real performance. In addition, DeNisi (1996) also mentioned that, even though some raters will observe the same ratee behaviours, they may go through a different information-interpretation process based on their personal preference and intentions, and that this may lead them to have different perceptions about the same observed behaviour(s). In other words, in SET even though students can all observe their professors' behaviours, their personalities and emotional makeup may ultimately lead them to give different ratings to heir professors, depending on just how much they like or dislike them. Consequently, to reduce the bias in the behaviour-observation process, the present study will discuss several course-related factors, a student-related factor, and professor-related factors that may influence SET scores.

To begin with, Allen (2006) pointed out that gender was one of the factors that may impact a rater's information-interpretation process. Previous studies have shown that raters may have biases toward gender-role stereotypes that can lead them to categorize ratees' behaviours (Ashmore & Del Boca, 1979; Neisser, 1976; Taylor & Crocker, 1981). Robbins and DeNisi (1993) conducted a laboratory designed study to examine the gender issue in the performance evaluation process. The result indicated that if the raters possessed some biases toward gender-role stereotypes, then they would only pay attention to poor performance in women. Then, through their information-interpretation, information-storage, memory-retrieval, and information-integration processes, they would just retain the information that, indeed, women perform more poorly than men. In China, a country with a much longer history than the United States, male dominance is even more firmly entrenched (Chia, Moore, Lam, Chuang, & Cheng, 1994; Hofstede, 1980). Therefore, because of the gender stereotypes and gender discrimination that exist in China, student and professor gender are also discussed in the present study in order to determine whether gender can influence SET results in China. Similar to stereotypes about gender, stereotypes about people's seniority in China are another important bias to take note of. People in China are more willing to trust people who are a lot older than them because they think that older people always have more experience than the younger people. Older people's claims are therefore thought to be more convincing (Boduroglu, Yoon, Luo, & Park, 2006). These stereotypes about seniority can also lead to a bias in SET scores.

Overall, there are three potential kinds of factors that may influence the success of SET in China: course-related factors (course type, class size, and course level), a student-related factor (gender), and professor-related factors (gender and seniority). The impact of these six factors on SET will be discussed in the following section.

LITERATURE REVIEW AND HYPOTHESES

As mentioned in above, three major types of factors have been proposed that can influence the result of SET, which are course-related factors (course type, class size, and course level), student-related factor (student gender), and professor-related factors (professor gender, and professor seniority). In this section, the relationship between professors' SET scores and these six factors will be discussed in turn.

Course-Related Factors

Course-related factors mainly refer to the factors that relate to the different characteristics of the course. Pounder (2007) pointed out several typical factors that relate to course, which are grading, class size, course content, and class timing. For the present study, course type, class size and course level will be discussed in some detail.

Course type (mandatory course / elective course) has been considered to have a significant influence on SET (Brockx et al., 2011). According to Marsh (1987), professors in the United States who teach elective courses receive higher evaluation scores than those who teach mandatory courses. Previous research suggested that students' passion for the subject of their choice would be reflected in their evaluations for that course and its professor (Marsh & Dunkin, 1992). Based on this view, Dev and Qayyum (2017) found that in the United Arab Emirates, elective courses also received higher SET scores than mandatory courses due to students' preference for the subjects of elective courses and that this preference constituted a bias in the Emiratis SET scores. However, in contrast to these last few studies (and from most studies on the topic), Beran and Violato (2009) found no direct relationship between course type (mandatory course / elective course) and student ratings. They did, however, suggest that student engagement can mediate this relationship.

In 2001, China introduced its new curriculum, which allows students to have more freedom in the selection of the courses that they want to take. Still, some political-related courses are required and listed as mandatory courses: students have to both take and pass these courses in order to graduate. Moreover, all of the basic courses offered at Chinese universities are mandatory courses. As opposed to North America universities where students can choose their concentration in grade 3 or grade 4, Chinese students are already assigned for concentration when they enter the university leading to less flexibility; also the basic courses are mandatorily assigned to them. Furthermore, in North American, students can choose courses based on the professors who teach them and on the time at which the courses are offered. However, in China, before they begin their university careers, students are placed into different classes, and mandatory courses are automatically assigned to their curriculum. For each mandatory course, several professors are assigned to a section. As a result, students rarely have the chance to choose the professor or the time for their mandatory courses. However, more flexibility is allowed in China when it comes to the elective courses. In this regard, the situation at Chinese universities is similar to that at North American universities, where students can choose the fields that they are interested in and also choose the professors with whom they want to learn.

The cognitive appraisal model suggests that even though students may observe the same behaviours of the professor, they may have different intentions during the observation process. If they are really interested in the class, they may have a stronger tendency to observe the professor's positive behaviours in the class, since people in general are more willing to focus on the things that they are really interested in and are not willing to be forced into activities (Clark, 2003). However, if students are not interested in the class, they will more likely observe the negative behaviours of the professor, which will lead them to selectively remember negative information about the professor's performance (DeNisi, 1996). Because students in China have more opportunities to choose their elective courses over and against their mandatory courses, students may well be more interested in the topics covered in the elective courses than those covered in the mandatory courses. This means that as far as SET is concerned, students will rate professors of elective courses higher than they would professors of mandatory courses.

Hypothesis 1a: *Professors who teach elective courses will receive a higher SET score than those who teach mandatory courses.*

Additionally, a great number of SET studies also have addressed the influence of class size on SET scores (Bedard & Kuhn, 2008; Feldman, 1984; Krautmann & Sander, 1999). In 1984, Feldman conducted a systematic analysis of the many studies that had been conducted on class size at around that time. The data was taken from 52 different studies of colleges and universities in the United States and Canada that had related class size to SET. The results showed that only 2 (3.85%) out of 52 studies found a positive relationship between class size and SET scores, while 22 (42.31%) studies reported an inverse relationship between variables, 12 (23.08%) studies found a curvilinear relationship between class size and SET scores, and the rest (30.76%) posited no relationship between the two variables. More recently, Bedard and Kuhn (2008) used data from the University of California Santa Barbara from the fall of 1997 to the spring of 2004 to investigate the same relationship. Their result showed that there was a large, highly significant, and non-linear negative relationship between class size and SET scores. However, Krautmann and Sander (1999) and Ting (2000) reported a non-significant relationship between class size and SET scores.

While a great number of researchers have already looked into and discussed the impact of class size on SET scores in the United States, only a handful of looked into and discussed the impact of class size on SET scores in China. However, the development of the Chinese tutorial services market can serve to demonstrate the impact that class size has had on the education system. In China, many people realize the value of receiving a sound education. Extracurricular tutorials have therefore long been popular, but they have become even more so in the aftermath of the reform. According to the document entitled “Investigation Report on the Status Quo of Chinese Extracurricular Tutorial Industry and Extracurricular Tutorial Institutions (2016)” from the Chinese Society of Education, the market value of China’s primary and secondary school extracurricular tutorial institutions exceeded 800 billion Yuan (approximately 160 billion Canadian dollars). More than 137 million students attended extracurricular tutorials in 2016, which took up to nearly 70% of the students in several big cities such as Beijing, Shanghai, and Shenzhen (Chinese Society of Education，2016). Within the extracurricular tutorial industry, in 2009, one-on-one tutorials accounted for up to 29% of the market, and small size tutorials (fewer than 15 people) took up to 55% of the market. However, it was estimated in 2014 that one-on-one tutorials would take up to 33% of the extracurricular tutorial market. From this information, it can be concluded that one-on-one tutorials have been experiencing an upward growth trend and are highly sought after in China, since smaller class sizes are generally considered to be more efficient, and students can better focus on the lecture and the professor in a small class. This tendency is consistent with what the cognitive appraisal model suggests. The cognitive appraisal model, as presented above, explains that it is difficult for raters to observe all aspects of the performance of each rate, since there may be some simple physical constraints or conflicting attention demands that may prevent them from doing so (DeNisi, 1996). In a smaller class, students can often communicate more frequently with their professors and get more timely feedback than they could in a larger class. Conversely, in large classes, there is an increased likelihood that students will get distracted by other things or students since professors are unable to pay attention to each of them. Therefore, these distractions may lead to students fail to catch the class and obtain all their professors' behaviours. Moreover, feedback would be harder to get, and direct face-to-face communication would be harder to obtain. Consequently, in a large class, students' final decisions on their professors' teaching performance may only be based on what little of their professors that they have seen, and this may lead to an inaccurate SET score. Therefore, although professors' skills and abilities are not being considered, it is still possible for professors teaching small classes to receive a higher evaluation than professors teaching large classes.

Hypothesis 1b: *There is a negative relationship between class size and SET scores.*

Course level is another course-related factor that can influence SET scores (Badri & Abdulla, 2006; Liaw & Goh, 2003). Marsh (1981) established that no relationship can be observed between course level and SET scores in the United States. However, based on a later study by Marsh in 1984, higher-level courses received relatively better student evaluation scores than lower-level courses in the United States. Recently, Lewis and McKinzie (2019) conducted a study in the southwestern United States, and the result of this study was consistent with the previous contentions that course level can indeed impact SET scores and that lower level courses tend to receive lower SET scores. Therefore, since course level can influence students' decisions about their courses and their professors in the United States, one may reasonably predict that course level will also impact Chinese students' perspectives on their courses and their professors. According to the data from one of the biggest Chinese online learning platforms, Chinese University MOOC (2018), which contains different courses from 213 different Chinese universities, higher-level courses may receive lower feedback than lower-level courses—even those that are taught by the same professors. For example, the course "Advanced Mathematics" contains four different levels of online classes provided by Tongji University, which was ranked among the top 10 universities in China by both the QS World University Rankings and Times Higher Education World University Rankings in 2018.

With the exception of second-level course, which were not scored, the first-, third-, and fourth-level classes received scores of 4.3 to 5, 4.4 to 5, and 3.4 to 5, respectively: it can be seen that the higher-level courses received lower scores. Based on the comments gleaned from each class, students wanted a more specific and effective teaching style in the higher-level courses than in the lower-level courses. Since the lower-level courses have relatively easy content, students were able to easily understand the class content irrespective of who was teaching the class and how well they behaved: this means that students gave relatively higher SET scores to the professors of the lower-level courses. However, for the higher-level courses, professors with poorer teaching skills, who cannot help students master the material, usually receive lower SET scores. In this case, the professors may receive lower SET scores for high-level courses than for low-level courses, since the SET scores of low-level courses are mainly based on the simplicity of the courses, while the SET scores of high-level courses are related more to professors' real performance and can more thoroughly reflect the professors' skills and professionalism. Moreover, according to the rating model derived from attribution theory (i.e., Kelley, 1967), DeNisi (1996) has suggested three major types of information that raters intend to observe in the first step of the cognitive appraisal process: the distinctiveness of the information; the consistency of the information; and the consensus regarding the information. The concept of distinctiveness of the information can be used to explain this situation. The distinctiveness of the information refers to the ratee's behaviour—whether he or she performs well on every task or only a specific task (Ruble & Feldman, 1976). For instance, a rater's decision about the ratee's performance of task B may be influenced by the latter's performance of task A. In the present study, distinctiveness of the information can be viewed as the connection between the tasks observed by the students and their view of their professors' competency. The question is: does it change from lower-level to higher-level courses because the course content is harder to understand, or does it change because the course content is harder to teach? Or is it both? Since higher-level courses are not only more difficult for professors to teach, but also more difficult for students to follow, the latter may attribute their slower learning progress to poor performance on the part of the professor. Hence students may suppose that their professor does not perform well in the higher-level courses compared to their performance in the lower-level courses, which may lead students to underrate their professor's performance in the higher-level courses. Consequently, it is possible that professors from a higher-level course will receive a lower SET score than professors from a lower-level course, and this result would go against the results of Marsh's research in the United States in 1984.

Hypothesis 1c: *There is a negative relationship between course level and SET scores.*

The Student-Related Factor

Student-related factors generally have to do with varied student-centered characteristics. Previous studies dealt with the influence of student gender on professors' SET scores. Moreover, a few studies also explored the impact of student year and maturity on SET scores (Pounder, 2007). The present study will focus on and discuss the issue of whether male students or female students are more likely to provide their professors with higher SET scores.

It is important to acknowledge that gender has always been considered as one of the most sensitive and inevitable elements of SET in the United States. However, it has not been studied a lot in the Chinese educational milieu. In 1975, Ferber and Huber found that, compared to their female counterparts, male students evaluated their female professors less favourably, and they explained that the same-gender preference effect might be the cause of this result. Later, Feldman (1983) reviewed 25 related studies published between 1932 and 1979, and he found that in seven studies female professors were rated higher than male professors on global items, in five studies female professors were rated higher than male professors on selected indices, and in 13 studies no gender differences were found. More recently, Centra and Gaubatz (2000) suggested that the different teaching styles among male professors and female professors—with female professors being generally viewed as better organized and more interactive—was the reason why female students were more likely to give a higher rating to female professors in the United States. However, a recent study has indicated that there were no significant differences between the SET scores provided by male and female students in the United States (Nowell, 2007). According to the research carried out by Hill and Motes (1995), males and females tend to process information differently, as females need more information in order to make decisions. In other words, during the evaluation process, female students may go through a longer behaviour-observation process and observe more professor behaviours than male students. Specifically, a high rating can be given by female students to a professor if and only if every behaviour of this professor is considered to be top tier, whereas male students are more likely to generalize their rating based on observing one instance of excellent or poor performance. As a consequence, female students hold professors to a higher standard when it comes to performance than do male students.

Moreover, in China, Carless (2009) argued that trust has a significant impact on evaluation process. Wang and Yamagishi (2005) investigated the gender differences related to trusting behaviour, since trust was (and still is) a critical component of social capital when dealing with interpersonal relationships. They conducted two experimental tests and concluded that 76% of the male students chose to trust their in-group members, while 33% of the female students chose to trust their in-group members. In addition, for the out-group members, 59% of the male students chose to trust them, while only 26% of the female students chose to trust out-group members, which suggests that male Chinese students are much easier to trust others than female Chinese students. Consequently, in consideration of the differences between male students and female students in China, it could be stated that female university students need to see more effort from their professors in order to be satisfied than do their male classmates. Hence, toward the same behaviours, male students will be more likely to provide higher evaluations to their professors than female students. Female students may require observing more nurturing behaviours from their professors than male students, in order to provide high assessments to their professors. Therefore, over and against the above-mentioned study results from the United States, male students in China are more likely to give a higher score than female students are to the same professor.

Hypothesis 2*: On average, male students will provide higher evaluations than female students.*

Professor-Related Factors

Professor-related factors refer to the factors that show the characteristics of professors. Generally, the central theme of professor-related factors is the effect of professor gender on SET (Pounder, 2007). Additionally, some researchers have also talked about professors' influencing tactics in SET studies, such as leniency and bringing food to class on the day of the evaluation. Professors' age, experience, and rank are also factors that have been focused on in SET studies. The present study will mainly talk about professor gender and professor seniority.

Since student gender appears to be a critical factor in SET, professor gender may also be thought to be one as well. In Western culture, the effect of professor gender on SET has been widely studied. The result of Mengel, Sauermann, and Zölitz's (2018) study showed that female professors received lower teaching evaluations than male professors from both male and female students in the Netherlands. They explained that this result was caused by the negative stereotypes of female professors, which made female professors look like they had a lack of confidence and appear more shy or nervous than their male counterparts. Furthermore, Anderson and Smith (2005) pointed out that the discrepancies of female and male professors' SET scores could be due, in part, to that students usually had higher expectations to their female professors. Specifically, the role congruity theory suggested that the backlash effects on evaluations lead students to require more feminine-stereotyped characteristics from their female professors, such as warmth and compassion. If female professors led in a stereotypically masculine style would be rated more negatively than male (Rudman & Phelan, 2008). For instance, researchers have shown that compared to male professors, female professors have to be more friendly in order to gain a higher ratings form their students, whereas the characteristic of friendly is not one of the reasons that influence students' evaluation of male professors (Kierstead, D'Agostino, & Dill, 1988). Similarly, Boring (2017) also found that male students gave a significantly higher overall score to their male professors than to their female professors. Statistical discrimination theory (Arrow, 1973; Phelps, 1972) suggests that evaluators may rely on stereotypes when assessing competence because they lack information on the evaluatees' actual performance (Altonji & Blank, 1999). Consequently, Boring (2017) found that, because it is hard for students to observe professors' actual performance—even after an entire semester—and because of gender stereotypes that existed in France at the time of his study, male professors were always perceived by both male and female students as being more knowledgeable and as possessing stronger classroom management skills than female professors. However, the study of Whitworth, Price, and Randall (2002) suggested that students usually perceived female faculty members in a better light than male faculty in the United States. Hence, for Western culture, gender stereotypes are one of the most controversial factors that can impact students' perception of their professors.

In 2015, Wang and Yamagishi designed a survey on gender issues in Chinese domestic academic institutions and received more than 1,600 effective replies from more than 40 such institutions. The results of this survey indicated that severe gender discrimination issues still exist in Chinese academic institutions. "There are quite a group of women who want to do academic research, but they have left the field because of the gender discrimination," according to the authors. There are preconceived notions in China that women should devote themselves more to their families than to their careers, and that women are relatively more sensitive than men and cannot perform as well as men in academic activities (Wang & Yamagishi, 2015). Consequently, Chinese students may also have these gender expectations and gender-role stereotypes when they perceive their professors' performance and when they evaluate it. Based on the cognitive appraisal model, because of these preconceived notions, students may tend to observe "poor performance" in their female professors during the initial behaviour-observation process, and this indicates that they will likely perceive female professors as generally performing worse than male professors. As a result, the present study suggests that, all things being equal, the scores that students give to their female professors may be lower than the scores that they give to their male professors in China, which is similar to most of the conclusions from North American studies.

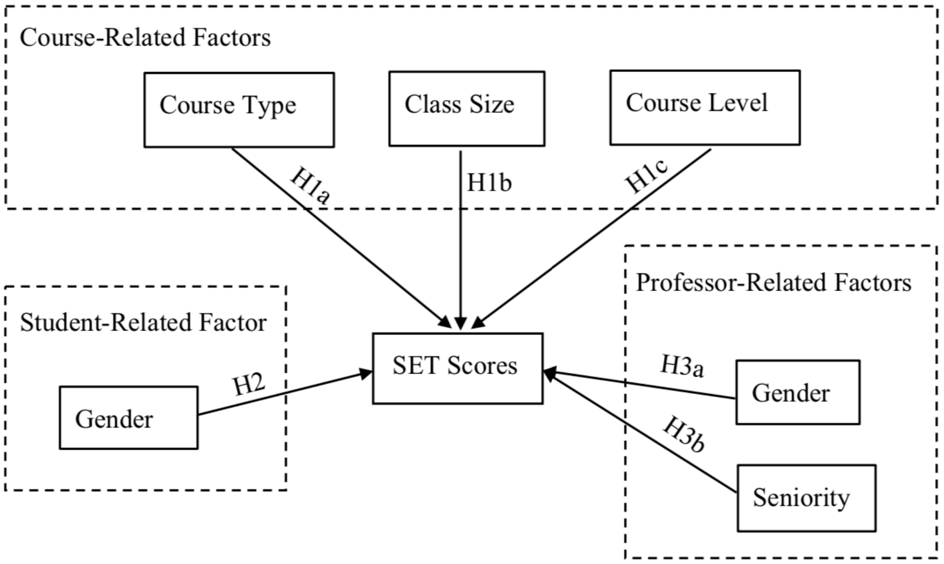
Hypothesis 3a: *On average, male professors will receive higher evaluations than female professors.*

Professor seniority in this study is a term that mainly refers to professor's status. However, in China, the promotion of professors is correlated with professors' age and teaching experience. Hence, the term professor seniority is also a combination of professors' age and experience. Generally, professors with a higher seniority usually elder and have more experience than professors who have lower seniority. The seniority of professors has also been shown to be a significant part of SET scores in the United States (Gokcekus, 2000). According to Remmers (1959), professors who had less than five years of experience at Purdue University received a lower assessment than those professors who had more than eight years of experience. Riley, Ryan, and Lifshitz (1969) pointed out that this differentiation was mainly due to the fact that professors with more experience were thought to have more knowledge of and interest in their specific subjects. However, they were also awarded the lowest rating for helpfulness and teaching attributes. Kinney and Smith (1992) suggested that professor seniority did have an impact on SET scores. They found that older and more experienced professors tended to receive more positive student evaluations than their younger and less experience colleagues. Clayson (1999), however, came to the opposite conclusion in his study: teacher seniority was negatively correlated with professor SET scores. He explained that teaching effectiveness did not improve with experience, but may actually have declined. Additionally, the results of Langbein's (1994) study showed a significant relationship between professors' experience and SET scores as well, but the relationship was non-linear, with experience having a positive influence on SET up to a certain point, beyond which the effect turned negative.

In traditional Chinese culture, the concept of seniority, not the concept of equality, is awarded a great deal of prestige. In general, most Chinese companies will stipulate some mandatory requirements for promoting employees that will involve seniority, regardless of the abilities of the candidates. Moreover, even with regard to employees who hold the same position, the one who is older is considered first for most employee welfare programs. Similarly, with respect to education, professors who are a lot older than their students are thought of as more experienced and gain more respect than those professors who just graduated and who are only a few years older than their students. Furthermore, professors who have a higher status or who have been assessed as having a higher rank among students are also more popular. According to Hofstede’s theory of cultural dimensions, China has a pretty low mark on the individualism level (Fernandez et al., 1997), which indicates that most Chinese people do not want to be thought of as being different from others. As a result, conformist behaviour is prevalent among Chinese people. Under this context, although some students may have a relatively negative view of elder professors’ performance, they still will not give them a lower score. Since China has these seniority-role stereotypes, people who are older or higher in rank or status are respected more. Hence, professors’ seniority will have a positive influence on SET scores, which is also indicated by the results of studies carried out in the United States.

Hypothesis 3b: *There is a positive relationship between professors’ seniority and SET scores.*

Figure 1. Research Model and Hypotheses



METHODOLOGY

Participants

A quantitative design was used to determine the relationship between course-related factors (course type, class size, and course level), a student-related factor (gender), and professor-related factors (gender and seniority), and professors' SET scores. The institution where this study was conducted is a middle-sized Chinese university with approximately 14,400 students and 1,017 faculty members. Each course has been evaluated by the students on a regular basis since 2007. The business department of the university is one of the largest departments in the university with 1,657 students and 86 faculty members. There are seven different undergraduate programs in the business department (most notable among which are the accounting, marketing, and finance programs).

Procedure

The data collected pertained to students' evaluations of professors in the fall semester of the 2017–2018 school year.[1] For each semester, the university conducts its own SET process, and permission to collect the data for the present study was sought from the university's academic affairs office. Moreover, to preserve the anonymity of the students' accounts, a serial number was used to represent each student and professor instead of using their names or IDs.

[1] In China, there are only two semesters in each school year: the spring and fall semesters.

At the end of each semester, students are required to complete the school-designed electronic SET questionnaire. The university explains the purpose of the evaluation before students begin to complete it; the evaluation is used to help professors improve the quality of their teaching and to help the university evaluate professors' performance. Different from the commonly practiced North American procedure, the Chinese procedure is to require students complete their SET questionnaires, otherwise they will not receive their official grades and register for the next semester.

Measurement

The SET questionnaire used in the present study included four major dimensions, which were instructors' "teaching professionalism," "teaching quality," "teaching strategies," and "teaching effectiveness." The dimensions were evaluated using a 5-point Likert-type scale ranging from 1 = "Strongly Disagree" to 5 = "Strongly Agree." Each dimension included 2 to 3 items for a total of 10 items (see Appendix).

Course Type. All of the courses were categorized as either mandatory or elective courses by the university.

Class Size. In the present study, class size was measured by the actual number of students enrolled in each class.

Course Level. There were four different levels of courses. Level 1 courses refer to the lowest-level courses, while level 4 courses refer to the highest-level courses. Although most courses that first-year students were allowed to take were level 1 courses, not all the level 1 courses were taken by first-year students. Second-year, third-year, and fourth-year students were also allowed to take lower-level courses.

Gender. "1" refers to male students (professors), and "2" refers to female students (professors).

Seniority. In the present study, professor seniority was measured by a professor's current title. The four different titles of professors included graduate teaching assistants (TAs), lecturers, associate professors, and (full) professors, and these titles were represented by "1," "2," "3," and "4," respectively.

Data Preparation

First, four data sets were aggregated for the purpose of the present study. The first set of data comprised the SET scores reported for all the subjects that the business department's students had chosen in the fall semester of the 2017–2018 school year. The second set of data provided the information on the characteristics of each course—the course type (mandatory or elective), the course level (first, second, third, or fourth year), and class size. The third set of data comprised information about the student profile, which includes each student's gender and school year. The last set of data comprised the information about instructor profiles, which contains instructors' gender, age, seniority, and status.

A total of 1,371 business department students completed the SET questionnaire in the fall semester of the 2017–2018 school year, which yielded a total of 19,248 evaluations. Of these, 6,094 evaluations were excluded because the professors' information was not available. The final sample included 1,371 students with 13,154 evaluations. The demographics of these 1,371 students were as follows: 240 participants were male (17.50%), and 1,131 participants were female (82.50%). There were 335 first-year students (24.43%), 343 second-year students (25.02%), 354 third-year students (25.82%), and 339 fourth-year students (24.73%). A total of 202 courses were evaluated, with 157 mandatory courses (77.72%) and 45 elective courses (22.27%). The class size of these 202 courses ranged from 18 students to 196 students. The levels of the 202 courses were 69 level 1 courses (34.16%), 66 level 2 courses (32.67%), 45 level 3 courses (22.28%), and 22 level 4 courses (10.89%). Moreover, 109 professors were evaluated. The demographics of these 109 professors were 53 male professors (48.62%), 56 female professors (51.38%), with ages ranging from 25 to 66 years (M = 40.97, SD = 7.10). In terms of the seniority, the professors included 4 (full) professors (3.67%), 39 associate professors (35.78%), 60 lecturers (55.05%), and 6 graduate teaching assistants (TAs) (5.50%).

ANALYSIS AND RESULTS

Descriptive Statistics and Correlations

Table 1 gives an overview of the preliminary analysis, which includes the means, standard deviations, and correlations between variables.

Table 1. Descriptive Statistics and Correlations

	Mean	Std. Deviation	1	2	3	4	5	6	7
1. Student Gender	1.83	.38							
2. Course Type	.79	.40	-.00						
3. Class Size	69.68	42.42	.00	.12**					
4. Course Level	2.10	1.05	.05**	-.40**	-.27**				
5. Professor Gender	1.63	.48	-.01	.13**	-.03**	-.19**			
6. Professor Seniority	2.42	.59	.02**	.00*	.00	-.31**	.05**		
7. SET Scores	91.42	20.39	-.00	.03**	-.05**	-.18**	.02**	-.04	

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

- N = 13,154.

Tests of Hypotheses

To determine the relationships between the independent variables (course type, class size, course level, student gender, professor gender, and seniority) and the dependent variable (SET scores), simple linear regression analyses were used.

Hypothesis 1a states that professors who teach elective courses will receive a higher SET score than those who teach mandatory courses. A simple regression analysis was used to test whether the course type significantly predicted SET scores. A significant regression equation was found ($\beta = .03$, $F_{(1,\ 13148)} = 9.84$, $p < .05$) with an R^2 of .001.[2] The mean SET score for professors who taught elective courses was 91.53, while the mean SET score for professors who taught mandatory courses was 91.39. Consequently, hypothesis 1a was supported.

Hypothesis 1b states that there is a negative relationship between class size and SET scores. A simple regression analysis was used to test if the class size significantly predicted SET scores. A significant regression equation was found ($\beta = -.05$, $F_{(1,\ 13148)} = 28.62$, $p < .05$) with an R^2 of .002. Hence, hypothesis 1b was supported.

Hypothesis 1c states that there is a negative relationship between course level and SET scores. A simple regression analysis was used to test whether the course level significantly predicted SET scores. A significant regression equation was found ($\beta = -.18$, $F_{(1,\ 13148)} = 425.14$, $p < .05$) with an R^2 of .03. The mean SET scores for professors who taught level 1 courses, level 2 courses, level 3 courses, and level 4 courses were 91.26, 91.68, 91.51, and 91.41, respectively. Consequently, hypothesis 1c was supported.

[2] For the sake of precision, three decimal points were used to present R^2.

Hypothesis 2 states that, on average, male students will provide higher evaluations than female students. A simple regression analysis was used to test whether students' gender significantly predicted SET scores. There was no significant regression found between students' gender and SET scores (β = -.00, $F_{(1, 13148)}$ = .20, n.s.). Therefore, hypothesis 2 was not supported.

Hypothesis 3a states that, on average, male professors would receive higher evaluations than female professors. A simple regression analysis was used to test whether professor gender significantly predicted SET scores. A significant regression equation was found (β = .02, $F_{(1, 13148)}$ = 7.62, $p < .05$) with an R^2 of .001. The mean SET score for male professors was 91.38, while the mean SET score for female professors was 91.43. Consequently, hypothesis 3a was not supported and the opposite result was found to be significant: on average, female professors received higher evaluations than male professors.

Hypothesis 3b states that there is a positive relationship between professor seniority and SET scores. A simple regression analysis was used to test whether professor seniority significantly predicted SET scores. A significant regression equation was found (β = -.04, $F_{(1, 13148)}$ = 23.73, $p < .05$) with an R^2 of .002. The mean SET scores were 91.21 for graduate teaching assistants, 93.87 for lecturers, 91.49 for associate professors, and 91.38 for (full) professors. Therefore, hypothesis 3b was not supported, but the opposite result was found to be significant: there was a negative relationship between professor seniority and SET scores.

After evaluating the simple linear regression analysis, the results for course level and professor seniority indicated that there might be a curvilinear relationship between course level (professor seniority) and SET scores. Consequently, a quadratic component was added to the model to see whether the fit can be increased.

For course level, adding a quadratic component to the model produced a significant increase in fit. According to Table 2, in the complete model F $_{(1, 13147)}$ = 446.31, p < .05 with an R^2 of .06, while with the incremental fit F $_{(1, 13147)}$ = 452.87, ΔR^2 = .032, p < .05.

Table 2. Curvilinear Relationship between Course Level and SET Scores

Equation	Model Summary						
	R^2	R^2 Change	F	F Change	df1	df2	Sig.
Linear	.031	0.31	425.14	425.14	1	13148	.00
Quadratic	.064	0.32	446.31	452.87	2	13147	.00

For professor seniority, adding a quadratic component to the model also produced a significant increase in fit. According to Table 3, in the complete model F $_{(1, 13147)}$ = 16.80, p < .05, with an R^2 of .03, while the incremental fit F $_{(1, 13147)}$ = 9.86, ΔR^2 = .001, p < .05.

Table 3. Curvilinear Relationship between Professor Seniority and SET Scores

Equation	Model Summary						
	R^2	R^2 Change	F	F Change	df1	df2	Sig.
Linear	.002	.002	23.73	23.73	1	13148	.00
Quadratic	.003	.001	16.80	9.86	2	13147	.00

Moreover, in order to find out whether the impact of some factors is bigger than others, this study also conducts multilinear regression analysis. The result confirmed the significance of the multilinear regression model (F $_{(1, 13143)}$ = 98.91, p < .05, R2 = .04). However, according to Table 4, in this all-inclusive regression model, course type, class size, and course level all significant influence SET scores. However, course level had a dominant impact on SET scores among these three factors, while student gender, teacher gender, and seniority did not have a significant impact on SET scores.

Table 4. Multilinear Regression Model

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
	B	Std.Error			
(Constant)	105.92	1.41		74.97	.00
Course Type	-2.56	.48	-.05	-5.38	.00
Class Size	-.05	.00	-.10	-11.22	.00
Course Level	-4.45	.20	-.23	-22.41	.00
Student Gender	.41	.46	.01	.88	.38
Teacher Gender	-.67	.37	-.02	-1.81	.07
Seniority	.26	.32	.01	.81	.42

DISCUSSION

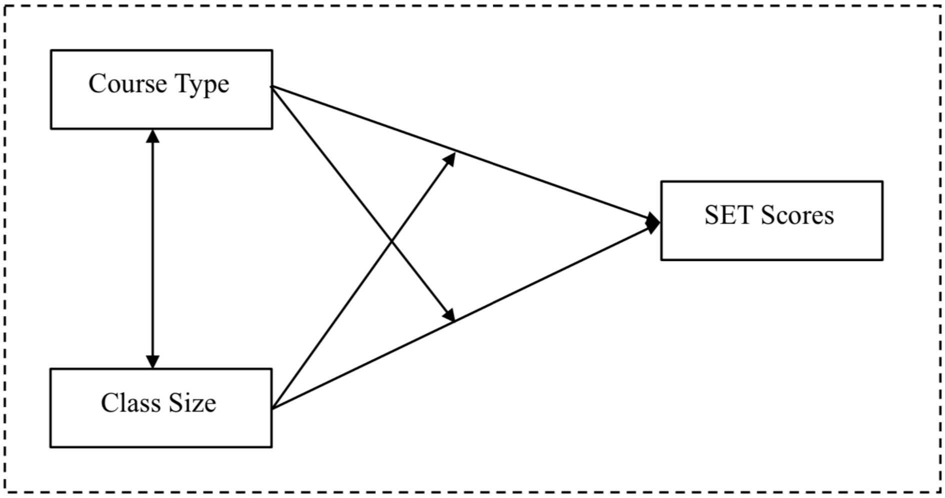
The primary purpose of the present study was to discuss the factors that can influence Chinese students' perceptions of their professors' performance levels in tertiary education. More specifically, this study investigated three major types of factors that have a potential impact on professors' SET scores, which are course-related factors (course type, class size, and course level), a student-related factor (gender), and professor-related factors (gender and seniority).

Course Type. In line with the previous research, a significant linear relationship between course type and professors' SET scores was found in this study, which suggests that professors who teach elective courses tend to receive higher SET scores than professors who teach mandatory courses. However, the changes of course type can only explain 0.1% of the changes in SET scores, which indicates that the difference between the mean SET score of professors who teach mandatory courses and professors who teach elective courses was quite small. Hence, although the course type can influence professors' SET scores, it is not the main factor that contributes to the difference in professors' SET scores. Generally, mandatory courses are more related to students' majors than elective courses are. Since most students choose their favorite field as their major, even though they may be more interested in their elective courses, they are also likely to enjoy their mandatory courses. Hence, in contrast to what was originally postulated, and most North American studies have suggested, course type can significantly impact professors' SET scores, however, its changes cannot explain a great number of the changes of SET scores. This result also supports the contention proposed by Marsh and Overall (1981) that the particular subject matter of a course has little effect on student rating.

Class Size. Similarly, the present study showed that there existed a significant negative relationship between class size and professors' SET scores. Although the results of this study suggest that class size has a more substantial influence on professors' SET scores than course type, the changes for class size can only explain 0.2% of the changes in SET scores, which indicates that the correlation between class size and SET scores is still not large. Therefore, class size cannot be viewed as the dominant factor in determining professors' SET scores.

Although the results of the present study suggest that both course type and class size can influence students' perceptions of their professors, neither course type nor class size play a dominant role in determining professors' SET scores. According to Table 1, there was a significant correlation between course type and class size, where p < .05. Consequently, this study also tested the interaction effect between course type and class size. The result showed the existence of a significant interaction between course type and class size, where F (1, 13148) = 13.56, p < .05. This result confirmed that course type could moderate the relationship between class size and SET scores, and vice versa, class size can also moderate the relationship between course type and SET scores. However, in China, elective courses generally had a bigger class size than mandatory courses. Therefore, the interaction effect between course type and class size could weaken the influence of both course type and class size on SET scores, which can explain why course type and class size cannot explain a large portion of the changes of SET scores.

Figure 2. Relationship between course level, class size and SET scores



Course Level. This study also tested the influence of course level on professors' SET scores in two different models. The simple linear regression model suggested a negative relationship between course level and SET scores. Similar to course type and class size, the changes of course level could only explain 3% of the changes for SET scores, which was not large as well. After adding a quadratic component to the first model, course level could explain 6% of the SET's changes, which indicated that there was a significant curvilinear relationship between course level and professors' SET scores. The mean score for the professors at each level showed that professors who taught level 2 courses received the highest SET scores, while professors who taught level 1 courses received the lowest SET scores. Furthermore, the mean score of professors who taught level 2 courses as compared to that of professors who taught level 4 courses showed a decreasing trend.

Historically, considerable studies have suggested that course level was correlated to SET scores (Lewis & Mckinzie, 2019). Since students tended to provide their lowest evaluations during their freshman year (Aleamoni, 1989), professors who taught the freshman courses received the lowest SET scores compared to professors who taught the higher-level courses (Goodson, Miertschin, Faulkenberry, Stewart, & Johnson, 2007). Additionally, according to the study conducted by Macfadyen and colleagues (2016), students in level 1 courses completed SET questionnaires more frequently than students in higher-level courses, and with the increasing of course level the SET response rate dropped significantly. This result indicates that there was a decreasing willingness among students to evaluate higher-level courses. However, since the SET questionnaire used in the present study was mandated by the university, students had no choice but to complete the questionnaires for higher-level courses (otherwise, they would risk not completing the courses). Therefore, this unwillingness led to the result that, except for level 1 courses, a negative relationship was observed between course level and professors' SET scores.

Moreover, the confounding effect of students' academic level may also explain this curvilinear tendency. According to the previous research on the relationship between students' academic level and students' satisfaction, third-year and fourth-year students, in comparison to first-year and second-year students, demand more of their professors' time outside of class (Douglas, Douglas, & Barnes, 2006). This finding suggests that higher-level students had higher personal SET score criteria. If professors want a higher SET score from higher-level students, they have to put more time in than they do for lower-level students. Since each level course was primarily taken by students in the same year—for instance, level 1 courses were mainly taken by first-year students—most of the raters of level 1 courses were first-year students, and so on. As a result, in the present study, the higher requirements of higher-year students may also have caused the SET scores of professors who taught higher-level courses to be lower than the scores of professors who taught lower-level courses.

Student Gender. This study failed to detect whether male students were more inclined to give higher SET scores to professors than female students. In other words, student gender did not have a significant impact on professors' SET scores. According to table 4, this result may be caused by that male students gave more extreme ratings with more considerable variance (.45) to professors than female students, while female students gave more ratings with less variance (.20) to professors than male students. However, the mean scores that provided by male and female students were similar, which were 91.36 and 91.29 respectively. As a consequence, the gender of students did not play a significant role in professor's SET scores.

Table 4. Estimates of Male and Female Students

Student Gender	Mean	Std.Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Male	91.36	.45	90.48	92.23
Female	91.29	.20	90.89	91.68

However, as discussed in the literature review and the hypotheses sections above, the results of previous North American research suggest same-gender preferences in SET, and show that female professors have received higher evaluation ratings from female students six out of eight times, while male professors have received equal ratings from male and female students (Centra & Gaubatz, 2000). The researchers explained in their report that this result might be caused by the different teaching styles of male and female professors, where female professors are viewed as better organized and more likely to encourage discussion. Consequently, the mean score and standard deviation of each mixed situation (where student evaluator and instructor were of different genders) were also measured in the present study, as shown in Table 5. The results revealed a significant interaction effect of student gender and professor gender on SET scores ($F_{(1, 13148)} = 2.855$, $p < .05$). Table 6 shows the significant differences between the means of SET scores that female students gave to male professors and the SET scores that they gave to female professors ($p < .05$). However, Table 6 does not show any significant differences between the means of SET scores that male students gave to male professors and the means of SET scores that they gave to female professors ($p > .05$). As mentioned above, compared to male students, female students require more information to inform their evaluations of professors (Hill & Motes, 1995), which indicates that female students tend to pay more attention to professors’ performance than do male students. Instead of being more sensitive to professors’ gender, female students might be more sensitive to professors’ behaviours than male students. So, if Chinese students prefer the teaching styles of certain female or male professors, female students may show a stronger preference for those professors who exhibit those styles than male students. Hence, in contrast to the results from North America, same-gender preferences did not exist among male students in this study, but female students were more likely to give higher SET scores to female professors than to male professors, which was consistent with the result of hypothesis 3a.

Professor Gender. This study found that male professors and female professors tended to receive different SET scores. It was found that, on average, female professors received higher evaluations than male professors. This result supported the overall argument that professor gender influenced SET scores. This finding, however, was contrary to what has been hypothesized above. As mentioned in the discussion section about the student-related factor, different teaching styles of male and female professors were often viewed as one of the dominant factors affecting students’ different perceptions of their professors (Centra & Gaubatz, 2000). Hence, personal preference for different teaching styles might explain this result.

Table 5. Descriptive Statistics of Mixed Gender Situation

Student Gender	Professor Gender	Mean	Std. Dev.
Male	Male	90.49	21.60
	Female	92.22	18.43
Female	Male	90.85	21.64
	Female	91.72	19.85

Table 6. Pairwise Comparisons between Four Different Means

- Dependent Variable: SET score

Student Gender	Professor Gender (I)	Professor gender (J)	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
Male	Male	Female	-1.73	.90	.05	-3.49	.02
Female	Male	Female	-.87*	.40	.03	-1.66	-.08

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

* The mean difference is significant at the

In 2000, Kimmel stated that male and female professors were perceived differently in various ways by students, which reflected the stereotypically gendered expectations of communication and interaction patterns. The traditional social role theory describes the difference of labour between male and female as a specialization of men in task-oriented (or instrumental) behaviour, while women in socioemotional (or expressive) behaviour (Eagly & Wood, 2011). Moreover, based on these theories, researchers suggested that male professors were usually recognized as more knowledgeable, while female professors were seen as more sensitive and more respectful of student ideas (Basow, 1995; Starbuck, 2003). Consequently, researchers have argued that teaching styles might be different between male professors and female professors. More specifically, researchers have found that male professors' teaching styles were more dominant and exacting, while female professors' teaching styles were more informal and open toward students and their ideas, the latter being perceived by students to be more effective in creating a climate of participation in the classroom (Crawford & MacLeod, 1990; Lacey, Saleh, & Gorman, 1998). Furthermore, female professors were also more likely to use teaching techniques such as group discussions and student presentations, while male professors were more likely to apply personal approaches, such as lectures (Starbuck, 2003).

In 2006, Zhang conducted a study to discern the preferred teaching styles and thinking modes among university students in mainland China. Consistent with the preferences of students in the United States and Hong Kong, mainland Chinese students expressed a strong proclivity for creative teaching styles, which included collaborative work. Similarly, they also exhibited a strong dislike for the norm-conforming teaching styles that restricted students to working individually. This result indicated that mainland Chinese students might prefer female professors' teaching styles. Therefore, students' preferences of teaching styles might be the reason why female professors are more likely to receive relatively high SET scores than male professors in China.

Additionally, same-gender preference effect mentioned in this study can also explain the result. The discussion part of student gender suggested that female students were more likely to give higher SET scores to female professors than to male professors. Since the student gender distribution in this study is not very equally, where only 240 (17.50%) male students participated, while 1,131 (82.50%) female participants participated, the more substantial portion of female participants may also be one of the reasons that female professors receive relatively high SET scores than male professors in this study.

Professor Seniority. Besides professor gender, the present study also tested the influence of professor seniority on SET scores. First, a simple linear regression model was conducted to examine the relationship between professors' seniority and SET scores. Although the result showed that professor seniority could influence students' perception of their professors, it suggests the opposite view of what was hypothesized. The result indicates that there was a negative relationship between professor seniority and SET scores.

However, the simple linear regression model showed a minimal correlation between seniority and SET scores, so a quadratic component was added to the model in order to find a better fit. The result showed a curvilinear relationship between professor seniority and SET scores. According to the mean score of the different levels of professor seniority, lecturers received the highest SET scores, while graduate teaching assistants received the lowest SET scores. The result of graduate teaching assistants receiving the lowest SET scores was consistent with the related hypothesis. As mentioned in the theoretical background and hypothesis section, the small age gap may be the main reason for this result.

Moreover, the result also showed that associate professors received lower SET scores than lecturers, but their SET scores were higher than those of the (full) professors, a result that runs contrary to this study's related hypothesis on the issue. This result may be caused by the fact that all full professors involved in this study were teaching level 4 courses, while the associate professors involved in this study were teaching four different level courses. For the lecturers, most were teaching lower-level courses, as they made up 62.12% of level 2 courses and 69.57% of level 1 courses. Consequently, based on the prior assumption that professors who taught higher-level courses would receive lower SET scores, the different course levels they taught may explain the decrease in SET scores going from lecturers to professors.

Practical Implications

Although not all of the proposed relationships between the six factors and the SET scores were found to be true, the present study did offer some interesting insights as to what kinds of factors were more likely to lead to bias in SET. Moreover, since there have been few studies on SET in the context of East Asian culture, the findings of this study can help fill the gaps in the available related literature and they can help institutions in the Chinese tertiary education sector better apply SET.

The results of this study show that five out of the six proposed factors were proven to be correlated to SET scores in China, namely, course type, class size, course level, professor gender, and professor seniority. However, all of these factors showed small coefficients of determination, which indicated that these five factors did not have large impacts on the SET scores and cannot determine the results of SET evaluations. The minimal bias from these factors suggests that SET is a reliable tool for course improvement in China—a result that was consistent with the results of the SET studies carried out in North America (Wright & Jenkins-Guarnieri, 2012).

However, as mentioned above in the introduction, SET is applied widely not only for both the purposes of course improvement and quality assurance, but also for the purpose of professor promotion decision making (Brockx et al., 2011; Macfadyen et al., 2016). Even though the influence of course type, class size, course level, professor gender, and professor seniority are not enough to determine the SET scores, it is still unfair to apply a biased system to professor promotion decisions. Additionally, this study applied the cognitive appraisal model to find out the potential factors that can influence student perception of professors. This study suggests that one of the most common reasons that students perceive their professors' behaviours differently than professors perceive their own behaviours was that students cannot observe their professors' behaviours in any kind of complete or objective sense. Since the result of the multilinear analysis showed that course level and course type had a more noticeable impact on SET scores than other factors, students' course level and course type should be considered firstly on discussing the reasons that lead to the scenario mentioned above. Even though their professors may have performed well, some students may not have noticed or may not have liked the professors in the first place, which are situations that will result in poor evaluations. Consequently, in order to apply SET to professor promotions in China, there is a strong need to supplement this evaluation system with other measures and assessments as a means of obtaining a more dependable picture of a professor and his or her teaching.

Overall, this study suggested that it is beneficial for Chinese tertiary education sector institutions to apply SET for the purposes of improving course quality and teaching quality, but SET scores alone are not sufficient for professor promotions because of the existing biases therein. Moreover, this study also pointed out that when professors find students perceiving their performance different than what they viewed their own performance, they are suggested to firstly look at either students' course type or course level.

Limitations and Future Research

In order to further interpret the findings of the present study, some limitations must be discussed. First of all, this study relied on a university-designed questionnaire to obtain the data. This poses several accuracy issues with the data collection process, as the questionnaire asked students to retrieve their memories about professor behaviour that they had observed during the previous semester. Research suggests that memory retrieval of past events could be inaccurate or even faulty (Bryman, Bell, Mills, & Yue, 2011), which indicates that relying only on memory-based data might introduce a degree of bias into the study results. Additionally, there are also some anecdotal notions that the real value of some courses can only be realized when students truly enter their careers. Moreover, in the Chinese tertiary education context, replying to the school's SET questionnaire was mandatory and related to the course selection process for the following semester, so even though there was a high response rate, the response quality needs to be improved. Therefore, for future SET studies, it is highly recommended to reconsider how to measure SET in a more appropriate way. There are two major questions that should be looked at: "Is it better to measure professors' behaviours at several different points in time over the course of the entire semester, or even asking students to measure their professors several years later, after they real enter their career?" and "Can students be motivated to complete the SET questionnaire in a more serious and less superficial manner?"

Second, the university setting in which the present study took place had certain limitations. The university setting for this study is a liberal arts university. According to the Chinese Ministry of Education, there are thirteen different categories of universities in the country, three of which are comprehensive universities, engineering universities, and liberal arts universities (Wu, Lv, & Guo, 2002). Previous studies have shown that liberal arts students and engineering students pay more attention to professors' teaching behaviours, while science students pay more attention to professors' teaching effectiveness (Li et al., 2018). Since students from different majors might have different perceptions of professors' performance, future researchers would be better served to study SET in different categories of Chinese universities in order to get more generalized results. The present study was conducted in a single middle-sized Chinese university with approximately 14,400 students and 1,017 faculty members. In order to pursue a student population with broader demographics, this study chose the business department as its target because it was supposed to have the most balanced gender distribution. However, surprisingly, male students only made up 17.50% of the total population while female students made up 82.50%. Although female Chinese students choose business-related majors more often than do male students, the gender imbalance here still cannot be ignored. Furthermore, the seniority distribution of this school's professors was not balanced either. Associate professors (35.78%) and lecturers (55.05%) made up most of the professoriate, while professors and graduate teaching assistants (TAs) only made up 3.67% and 5.50%, respectively. As a result, in order to reach more generalizable conclusions, future researchers would be better served to conduct SET studies in Chinese universities of various sizes. Analyzing the SET results of students from various departments and/or faculties will also contribute to achieving more generalizable results.

Third, due to data accessibility restrictions, the present study only focused on six factors that could potentially contribute to bias SET scores. Previous researchers have suggested that there are some other factors that can influence SET scores in the Western (e.g., Badri et al., 2006; Pounder, 2007). For instance, in 1997 Chye Koh and Meng Tan found that the time of the course influenced student perceptions of professors' behaviours. In their study, professors who taught courses that were held in the later part of the week received higher SET scores because of the more relaxed atmosphere that existed at the end of the week. Moreover, previous researchers have also found a significant positive relationship between student course grades and SET scores (McPherson, 2006; McPherson and Jewell, 2007; McPherson, Jewell, & Kim, 2009). Although most researchers have attributed this positive relationship to professors' grading leniency, some researchers still advocate that it was effective teaching practices that led to this result. For professor-related factors, previous Western studies have also suggested that professors' behavioural traits can influence SET scores. In 1986, Cardy and Dobbins advocated that a student's "liking" of the professor had significant positive associations with teaching evaluations, and Abrami and colleagues (1982) argued that a student's favourable opinion of his or her professor even played a more significant role in SET than how well their professor knew his or her field of expertise. Later, in 1999, Jackson and colleagues conducted a quantitative study, and the results supported the views of Abrami and colleagues (1982) by further explaining that university professors' ability to "get along" with their students (rapport) overlapped with more education-related factors, such as professor enthusiasm for the subject, breadth of subject coverage, group interaction, and learning value. Consequently, it will also be worthwhile for future researchers not only to look at more potential variables that are associated with SET in general, but also to look at whether certain variables can cause a bias in SET scores in different contexts. Moreover, conducting some meta-analysis on the topic of comparing different factors' varied influence on SET scores in different countries can also contribute to strengthening the application of SET both in general and in specific.

CONCLUSION

With the purpose of improving course content, assuring teaching quality, and helping determine professor promotions, SET is widely applied across the tertiary education sector across the globe. However, only a few studies have been conducted on SET in the context of East Asian culture. The present study chose China as its target and discussed the influence of six different factors on SET in the country's tertiary education sector, which are course type, class size, course level, student gender, professor gender, and professor seniority. Even though the results of this study show that, with the exception of student gender, all of the above factors had a significant impact on SET scores, all the correlations between these five variables and SET scores were rather small. These results suggest that SET scores should be taken into consideration for the purposes of course improvement and teaching quality assurance, but not for the sole purpose of professor promotions (it could certainly play a part, however). By improving upon the methodology and design of the present study, researchers can further explore the potential biases that can influence SET scores in China. Moreover, because of the huge culture gap that exists between Western countries and Eastern countries, it is also crucial for researchers in the future to develop more systematic and complete SET systems for China and other East Asian countries.

REFERENCES

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219–231. doi:10.1037/0022-0663.82.2.219

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research*, *52*(3), 446–464. doi:10.2307/1170425

Aleamoni, L. M. (1989). Typical faculty concerns about evaluation of teaching. In L. M. Aleamoni (Ed.), *Techniques for evaluating and improving instruction*. San Francisco, CA: Jossey-Bass.

Allen, T. D. (2006). Rewarding good citizens: The relationship between citizenship behavior, gender, and organizational rewards. *Journal of Applied Social Psychology*, *36*(1), 120–143. doi:10.1111/j.0021-9029.2006.00006.x

Altonji, J. G., & Blank, R. M. (1999). Race and gender in the labor market. In O. Ashenfelter and D. Card (Eds.), *Handbook of labor economics*, *Vol. 3C* (pp. 3143–3259). Amsterdam: North-Holland.

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184-201.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton, NJ: Princeton University Press.

Ashmore, R. D., & Del Boca, F. K. (1979). Sex stereotypes and implicit personality theory: Toward a cognitive–social psychological conceptualization. *Sex roles*, *5*(2), 219–248. doi:10.1007/BF00287932

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management*, *20*(1), 43–59. doi:10.1108/09513540610639585

Bakker, A. B., & Bal, M. P. (2010). Weekly work engagement and performance: A study among starting teachers. *Journal of Occupational and Organizational Psychology*, *83*(1), 189–206. doi:10.1348/096317909X402596

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*(4), 656–665. doi.org/10.1037/0022-0663.87.4.656

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of teacher effectiveness. *Economics of Education Review*, *27*(3), 253–265. doi:10.1016/j.econedurev.2006.08.007

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education*, *30*(6), 593–601. doi:10.1080/02602930500260688

Beran, T., & Violato, C. (2009). Student ratings of teaching effectiveness: Student engagement and course characteristics. *Canadian Journal of Higher Education*, *39*(1), 1–13. ERIC No.: EJ849729

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

Boduroglu, A., Yoon, C., Luo, T., & Park, D. C. (2006). Age-related stereotypes: A comparison of American and Chinese cultures. *Gerontology*, *52*(5), 324–333. doi:10.1159/000094614

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. doi:10.1016/j.jpubeco.2016.11.006

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance*, *6*(1), 1–21. doi:10.1207/s15327043hup0601_1

Brockx, B., Spooren, P., & Mortelmans, D. (2011). Taking the grading leniency story to the edge: The influence of student, teacher, and course characteristics on student evaluations of teaching in higher education. *Educational Assessment, Evaluation and Accountability*, *23*(4), 289–306. doi:10.1007/s11092-011-9126-2

Bryman, A., Bell, E., Mills, A.J., & Yue, A.R. (2011). *Business research methods*. Don Mills, ON: Oxford University Press.

Cai, Y., & Lin, C. (2006). Theory and practice on teacher performance evaluation. *Frontiers of Education in China*, *1*(1), 29–39. doi:10.1007/s11516-005-0004-x

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of applied psychology*, *71*(4), 672.

Carson, R. L. (2006). *Exploring the episodic nature of teachers' emotions as it relates to teacher burnout* (Unpublished doctoral dissertation). Purdue University, West Lafayette, IN.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education 44*(5): 495–518.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71*(1), 17–33. doi:10.2307/2649280Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71–88. doi:10.1080/02602930301683

Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education*, *34*(1), 79-89.

Chia, R. C., Moore, J. L., Lam, K. N., Chuang, C. J., & Cheng, B. S. (1994). Cultural differences in gender role attitudes between Chinese and American students. *Sex Roles*, *31*(1–2), 23–30. doi:10.1007/BF01560275

Chinese Society of Education (2016). Zhongguo fudao jiaoyu hangye ji fudao jigou jiaoshi xianzhuang diaocha baogao [Investigation report on the status quo of the Chinese extracurricular tutorial industry and extracurricular tutorial institutions]. Chinese Society of Education.

Chinese University MOOC. (2018). Retrieved from: https://www.icourse163.org/

Chye Koh, H., & Meng Tan, T. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4), 170–178. doi:10.1108/09513549710186272

Clark, R. E. (2003). Fostering the work motivation of individuals and teams. *Performance Improvement*, *42*(3), 21–29. doi:10.1002/pfi.4930420305

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21*(1), 68–75. doi:10.1177/0273475399211009

Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review*, *21*(2), 101–112. doi:10.2753/MER1052-8008210201

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3): 281–309. doi:10.3102/00346543051003281

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84*(1), 3–13. doi:10.1037/0021-9010.84.1.3

Crawford, M., & MacLeod, M. (1990). Gender in the college classroom: An assessment of the “chilly climate” for women. *Sex Roles*, *23*(3–4), 101–122. doi:10.1007/BF00289859

DeBoer, G. E. (1991). *A history of ideas in science education: Implications for practice*. New York, NY: Teachers College Press.

DeNisi, A. S. (1996). *A cognitive approach to performance appraisal: A program of research*. New York: Routledge.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance*, *33*(3), 360–396. doi:10.1016/0030-5073(84)90029-1

Dev, S. S., & Qayyum, N. (2017). Major factors affecting students’ perception towards faculty evaluation of teaching (SET). *Journal of Social Studies Education Research*, *8*(3), 149–167.

Dorfman, P. W., & Howell, J. P. (1988). Dimensions of national culture and effective leadership patterns: Hofstede revisited. In R. G. Farmer & E. G. McGoun (Eds.), *Advances in international comparative management*, *Vol. 3* (pp. 127–150). Greenwich, CT: JAI Press.

Douglas, J., Douglas, A., & Barnes, B. (2006). Measuring student satisfaction at a UK university. *Quality assurance in education*, *14*(3), 251–267. doi:10.1108/09684880610678568

Duschl, R. A. (1998). Research on the history and philosophy of science. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 443–465). Dordrecht, The Netherlands: Kluwer.

Eagly, A. H., & Wood, W. (2011). Social role theory. *Handbook of theories in social psychology*, *2*, 458-476.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied psychology*, *66*(2), 127–148. doi:10.1037/0021-9010.66.2.127

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, *18*(1), 3–124.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*(1), 45–116.

Feng, Z. (2002). On the transition of teacher evaluation mechanism in the context of quality education. *Journal of Shengyan Teacher College*, *26*(2), 10–12.

Ferber, M. A., & Huber, J. A. (1975). Sex of student and instructor: A study of student bias. *American Journal of Sociology*, *80*(4), 949–963.

Fernandez, D. R., Carlson, D. S., Stepina, L. P., & Nicholson, J. D. (1997). Hofstede's country classification 25 years later. *Journal of Social Psychology*, *137*(1), 43–54. doi:10.1080/00224549709595412

Flaherty, C. (2016). Bias against female instructors. *Inside Higher Ed.*, January 11. Retrieved from: https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching

Gokcekus, O. (2000). How do university students value economics courses? A hedonic approach. *Applied Economics Letters, 7*(8), 493–496. doi:10.1080/13504850050033238

Goodson, C., Miertschin, S., Faulkenberry, L., Stewart, B., & Johnson, C. (2007). *Integrating technology: Our culture, our students*. Paper presented at the 2007 ASEE Annual Conference & Exposition, Honolulu, HI. Retrieved from: https://peer.asee.org/2056

Govaerts, M. J., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education*, *16*(2), 151–165. doi:10.1007/s10459-010-9250-7

Govaerts, M. J., Van de Wiel, M. W., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education*, *18*(3), 375–396. doi:10.1007/s10459-012-9376-x

Government of China. (2016). Gaodeng jiaoyu [Higher education report]. Retrieved from: http://www.gov.cn/guoqing/2016-07/09/content_5089882.htm

Griffin, M., Neal, A., & Neale, M. (2000). The contribution of task performance and contextual performance to effectiveness: Investigating the role of situational constraints. *Applied Psychology, 49*(3), 517–533. doi:10.1111/1464-0597.00029

Guo, S. (1996). Adult teaching and learning in China. *Convergence*, *29*(1), 21–33.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student scores*. Educational Testing Service. Retrieved from:

Hill, J., & Motes, W. H. (1995). Professional versus generic retail services: New insights. *Journal of Services Marketing, 9*(2), 22–35. doi:10.1108/08876049510085991

Hofstede, G. (1980). Culture and organizations. *International Studies of Management & Organization*, *10*(4), 15–41. doi:10.1080/00208825.1980.11656300

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), Article 1304016. doi:10.1080/2331186X.2017.1304016

Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, *77*(2), 187–196. doi:10.1037/0022-0663.77.2.187

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement*, *59*(4), 580–596. doi:10.1177/00131649921970035

Jiang, F. H. (2001). *Modern educational evaluation: Theory, technology, and methods*. China: Guangdong People Press.

Jiang, F. H., & Zhang, Q. L. (2003). Some thoughts on teacher evaluation in the context of the curriculum reform for basic education. *Educational Guiding Journal*, 5, 40–43.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, *80*, 342-344.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation, Vol. 15* (pp. 192–238). Lincoln: University of Nebraska Press.

Kember, D., & Leung, D. Y. (2009). Development of a questionnaire for assessing students' perceptions of the teaching and learning environment and its use in quality assurance. *Learning Environments Research*, *12*(1), 15–29. doi:10.1007/s10984-008-9050-7

Khan, S., Gul, S., Shah, I. M., & Khan, A. (2012). Teachers' stress, performance & resources: The moderating effects of resources on stress & performance. *International Review of Social Sciences and Humanities*, *2*(2), 10–23.

Kimmel, M. S. (2000). *The gendered society*. New York: Oxford University Press.

Kinney, D. P., & Smith, S. P. (1992). Age and teaching performance. *Journal of Higher Education*, *63*(3), 282–302. doi:10.1080/00221546.1992.11778363

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*(1), 59–63. doi:10.1016/S0272-7757(98)00004-1

Lacey, C. H., Saleh, A., & Gorman, R. (1998). *Teaching nine to five: A study of the teaching styles of male and female professors*. Paper presented at the Annual Women in Educational Leadership Conference, Lincoln, NB. Eric No.: ED442334

Langbein, L. I. (1994). The validity of student evaluations of teaching. *Political Science & Politics*, *27*(3), 545–553. doi:10.2307/420225

Lewis, V. J., & McKinzie, K. (2019). Impact of industry and teaching experience, course level, and department on student evaluations. *Quarterly Review of Business Disciplines, 5*(4) 335–356.

Li, G., Hou, G., Wang, X., Yang, D., Jian, H., & Wang, W. (2018). A multivariate generalizability theory approach to college students' evaluation of teaching. *Frontiers in Psychology, 9*, 1065. doi:10.3389/fpsyg.2018.01065

Li, R. (2002). New ideas on teacher evaluation. *Modern Elementary and Secondary Education, 6*, 53–55.

Li, R., & Xuan, L. (2003). Limitation and transcendence on teacher evaluation. *Educational Science Research, 3*, 22–24.

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37–43. doi:10.1108/09513540310456383

Liu, S., Keeley, J., & Buskist, W. (2015). Chinese college students' perceptions of characteristics of excellent teachers. *Teaching of Psychology*, *42*(1), 83–86. doi:10.1177/0098628314562684

Liu, S., & Meng, L. (2009). Perceptions of teachers, students and parents of the characteristics of good teachers: A cross-cultural comparison of China and the United States. *Educational Assessment, Evaluation and Accountability*, *21*(4), 313–328. doi:10.1007/s11092-009-9077-z

Liu, S., & Teddlie, C. (2003). The ongoing development of teacher evaluation and curriculum reform in the People's Republic of China. *Journal of Personnel Evaluation in Education, 17*(3), 243–261. doi:10.1007/s11092-005-2982-x

Liu, S., & Teddlie, C. (2005). A follow-up study on teacher evaluation in China: Historical analysis and latest trends. *Journal of Personnel Evaluation in Education*, *18*(4), 253–272. doi:10.1007/s11092-007-9029-4

Louie, K. (1984). Salvaging Confucian education (1949–1983). *Comparative Education, 20*(1), 27–38. doi:10.1080/0305006840200104

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education*, *41*(6), 821–839. doi:10.1080/02602938.2015.1044421

Marlin, J. W., & Gaynor, P. E. (1989). Do anticipated grades affect student evaluations? A discriminant analysis approach. *College Student Journal, 23*(2), 184–192.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707–754. ERIC No. EJ307773

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253–388. doi:10.1016/0883-0355(87)90001-2

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 143–233). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*(4), 303–314. doi:10.1016/0742-051X(91)90001-6

Marsh, H. W., & Overall, J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. *American Educational Research Journal, 18*(1), 103–112. doi:10.3102/00028312018001103

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *Journal of Economic Education, 37*(1), 3–20.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88*(3), 868–881.

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37–51.

Mengel, F., Sauermann, J., & Zölitz, U. (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566. doi:10.1093/jeea/jvx057

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi:10.2307/1175249

Min, Z. (2007). Adaptive performance: New development of teachers' adaptive performance structure. *Research in Higher Education of Engineering*, *2*, 235–242.

Mueller, F. J. (1951). Trends in student ratings of faculty. *American Association of University Professors' Bulletin, 37*, 319–324. doi:10.2307/40220817

Nayyar, M. R. (1994). Some correlates of work performance perceived by first line supervisor: A study. *Management and Labour Studies*, *19*(1), 50–54.

Neisser, U. (1976). *Cognition and reality: Principles and implication of cognitive psychology*. San Francisco: Freeman.

Newton, J. D. (1988). Using student evaluation of teaching in administrative control: The validity problem. *Journal of Accounting Education*, *6*(1), 1–14. doi:10.1016/0748-5751(88)90033-4

Nimmer, J. G., & Stone, E. F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning. *Research in Higher Education*, *32*(2), 195–215. doi:10.1007/BF00974437

Nowell, C. (2007). The impact of relative grade expectations on student evaluation of teaching. *International Review of Economic Education, 6*(2), 42–56. doi:10.1016/S1477-3880(15)30104-3

Organisarion for Economic and Cooperation and Development (OECD) (2019). Adult education level (indicator). doi:10.1787/36bce3fe-en. Retrieved from: https://data.oecd.org/eduatt/adult-education-level.htm#indicator-chart

Page, C. F. (1974). *Student evaluation of teaching: The American experience*. London: Society for Research into Higher Education.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*(4), 659–661.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education*, *15*(2), 178–191. doi:10.1108/09684880710748938

Rao, Z. (1996). Reconciling communicative approaches to the teaching of English with traditional Chinese methods. *Research in the Teaching of English, 30*(4), 458–471.

Remmers, H. H. (1959). *The appraisal of teaching in large universities*. Ann Arbor: University of Michigan Press.

Riley, J. W., Ryan, B. F., & Lifshitz, M. (1969). *The student looks at his teacher.* Port Washington, NY: Kennikat Press.

Robbins, T. L., & DeNisi, A. S. (1993). A cognitive look at sex bias in the performance appraisal process. *Journal of Management, 19*, 113–126.

Rockey, S. (2014). Women in biomedical research. National Institutes of Health, August 8. Retrieved from: https://nexus.od.nih.gov/all/2014/08/08/women-in-biomedical-research/

Ruble, D. N., & Feldman, N. S. (1976). Order of consensus, distinctiveness, and consistency information and causal attributions. *Journal of Personality and Social Psychology*, *34*(5), 930–937. doi:10.1037/0022-3514.34.5.930

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in organizational behavior*, *28*, 61-79.

Seibert, W. F. (1979). Student evaluations of instruction. In S. C. Ericksen (Ed.), *Support for teaching at major universities*. Ann Arbor, MI: Center for Research on Learning and Teaching.

Seldin, P. (1989). Using student feedback to improve teaching. In D. DeZure (Ed.), *To improve the academy, Vol. 16* (pp. 335–346). Stillwater, OK: New Forums Press.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures. *Assessment & Evaluation in Higher Education*, *25*(4), 397-405.

Starbuck, W. H. (2003). Turning lemons into lemonade: Where is the value in peer reviews? *Journal of Management Inquiry*, *12*(4), 344–351. doi:10.1177/1056492603258972

Tagomori, H. T., & Bishop, L. A. (1995). Student evaluation of teaching: Flaws in the instruments. *Thought & Action*, *11*(1), 63–78. ERIC No. EJ506864

Tarman, B., & Acun, I. (2010). Social studies education and a new social studies movement. *Journal of Social Studies Education Research, 1*(1), 1–16.

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario Symposium*, *Vol. 1* (pp. 89–134). Hillsdale, NJ: Lawrence Erlbaum.

Thawabieh, A. M. (2017). Students’ evaluation of faculty. *International Education Studies*, *10*(2), 35–43. ERIC No. EJ1130380

Ting, K. F. (2000). A multi-level perspective on student ratings of instruction. *Research in Higher Education, 41* (5), 637–661.

Wang, F., & Yamagishi, T. (2005). Group-based trust and gender differences in China. *Asian Journal of Social Psychology*, *8*(2), 199–210. doi:10.1111/j.1467-839x.2005.00167.x

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business*, *77*(5), 282–289. doi:10.1080/08832320209599677

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

Wu, S. L., Lv, J., & Guo, S. L. (2002). 2002 zhongguo daxue pingjia [2002 China university evaluation]. *Science and Management of Science and Technology*, *23*(5), 42-45. doi:10.3969/j.issn.1002-0241.2002.05.011

Xinhua News Agency (2010). Guojia zhongchangqi jiaoyu gaige he fazhan guihua gangyao [Outline of national medium- and long-term program for education reform and development (2010–2020)]. Retrieved from: http://www.hprc.org.cn/gsgl/zggk/guijiajg/xiangguanwenjian/201007/t20100730_56482.html

Ying, P. C., & Fan, G. R. (2001). Case study on teacher evaluation patterns: On the limitations of traditional pattern of teacher evaluation and exploration of a new pattern. *Theory and Practice of Education, 21*(3), 22–25.

Zhang, L. F. (2006). Preferred teaching styles and modes of thinking among university students in mainland China. *Thinking Skills and Creativity*, *1*(2), 95-107.

APPENDIX

I. Teaching Professionalism (20%)	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1. The instructor is well organized and prepared for every class (e.g., handouts, courseware, etc.). (30%)	○	○	○	○	○
2. The instructor was never late, never left the classroom early, and did not change the class time or cancel the class without a valid reason. (35%)	○	○	○	○	○
3. The instructor spoke clearly and had a moderate speaking speed so the students could easily take notes and follow the concepts being taught. (35%)	○	○	○	○	○

II. Teaching Quality (30%)	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1. The instructor clearly and effectively explained the concepts and students could easily understand the major points made in each class. (50%)	○	○	○	○	○
2. The instructor provided an appropriate balance between theory and practice. (50%)	○	○	○	○	○

III. Teaching Strategies (30%)	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1. The instructor used a variety of instructional methods to reach the course objectives (e.g., group discussions, student presentations, media presentations, etc.). (50%)	○	○	○	○	○
2. The instructor challenged students to do their best work, and focused on helping students to develop their analysis and problem-solving skills. (50%)	○	○	○	○	○

IV. Teaching Effectiveness (20%)	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1. There were active and positive interactions between the instructor and students, which likely stimulated students' interest in the class. (30%)	○	○	○	○	○
2. The instructor strict enforced class discipline, and students were able to learn and study in a quiet and secure atmosphere. (35%)	○	○	○	○	○
3. The instructor provided clear and constructive feedback, and the feedback was provided within the stated time frame. (35%)	○	○	○	○	○