In the United States, college and university professors are evaluated to varying degrees on research productivity, service, and teaching effectiveness. These dimensions are often used for high-stakes administration decisions (e.g., hiring, retention, promotion, pay, and tenure; Hornstein, 2017; Stark & Freishtat, 2014; Spooren et al., 2013; Stoebe, 2020). Depending on the institution, a major failure of one these areas could jeopardize a professors’ position within the department; thus, evaluating research, service, and teaching is of the utmost importance. Focusing on evaluating educator on teaching effectiveness; however, is both difficult and costly. Indeed, the vast majority of the 9,000 professors polled by the American Association of University Professors shared that teaching needs to be taken as seriously as research and service (Flaherty, 2015).

*Teaching effectiveness* can be defined as the degree to which student achievement is facilitated (i.e., how much have students learned in a particular course; Cohen, 1981). Additionally, the assessment of this construct often comes from the students and their evaluations goes by many names with some student assessments focusing on the instructor (e.g., “Student Opinion of Instruction”, “Student Evaluations of Teaching”, “Students Opinion of Teaching Effectiveness”, “Students Evaluation of Faculty”); further, other measurements focus on the course (e.g., “Overall Course Ratings”, “Instruction Rating”; Cohen, 1981; Flaherty, 2020). Both reliability and validity need to be demonstrated for a measurement tool to have utility. If high-stakes decisions are to be made regarding administration decisions, then the natural question arises: are students’ evaluation of the course and/or instructor reliable and valid measures of teaching effectiveness?

**Validity**

Sheehan (1975) reviewed the literature and listed several factors of bias that exist within the student level (e.g., Student’s sex, class, age, GPA); class-level (e.g., subject matter, class size, if the course was required); and instructor-level (e.g., sex of the instructor and interaction with sex of the student, academic rank of the professor). Fifty years ago, serious concerns were raised if students’ assessments provided useful information to decide on personnel decisions. Important to note, Cohen (1981) is an often-cited work with the most optimistic estimate that ratings of instruction are moderately related to student achievement; however, the meta-analysis, others, and literature reviews of the topic have undergone serious criticism (e.g., Stroebe, 2018; Uttl et al. 2017). Fifty years later; these concerns remain (e.g., Boring et al., 2016; Dunn et al., 2016; Hornstein, 2017; Uttl et al. 2017).

The sources of bias listed by Streehan (1975) have been replicated and/or extended in length. Systemic reviews and recent studies underscore that sexism (e.g., MacNell et al., 2014; Mitchell & Martin, 2018), racism (e.g. Smith & Hawkins, 2011), and general bias pervades students’ evaluations of traditional courses and possibly exist for online ones as well (e.g., Heffernan, 2021; Rovai et al., 2006; Sullivan et al., 2013; Zheng et al., 2023). Individual factors may also yield some influence, including instructors’ cultural background (e.g., Fan et al., 2019), attractiveness (e.g., Felton et al., 2008; Wright, 2008), position ranking (e.g., Johnson et al., 2013), and students’ expected grade from the course (e.g., Chen et al., 2017; Crumbly et al., 2001; Marks, 2000). Others suggest biasing factors of students’ ratings include the volume of the instructor’s voice and how legible their instructor’s writing is (Becker et al., 2012) Stroebe (2018) underscores the possible danger of an incentive system that is tied to student ratings; instructors may be then incentivized to be a less effective teacher (e.g., grade leniently, choose to teach courses based off student interest) rather than challenge students critically.

What if we steel manned students’ ratings and assumed ideal conditions? Recent computational simulation work that assumes student ratings are valid, are unbiased, and employ the optimistic inter-student reliability estimates of instructor’s effectiveness. In other words, ideal assumptions – find that ‘known’ high quality instructors may be at the lowest percentile of student ratings and vice-versa. One of the most commonly used proposed solutions is that using multiple evaluations of teaching effectiveness (e.g., subject-matter sit-ins on lecture, peer reviews of course curriculums) may be the solution (e.g., Benton & Young, 2018; Berk, 2018; Esarey & Valdes, 2020; Kornell & Hausman, 2016); however, the cost of implementing a more accurate multi-pronged approach may be more than most universities can afford to pay. With all critiques of students’ ratings, it is one of the most affordable measurements of teacher effectiveness in both time and money. The utility of students’ ratings may extend out so far as to aid other students when choosing which classes to pursue and with whom (e.g., Stankiewicz, 2015); unfortunately, even if instructors believe such ratings to be an inappropriate, it may influence their self-efficacy as an educator regardless (Boswell, 2016).

**Reliability**

Past investigations and reviews utilizing large samples concluded student ratings are reliable and stable (e.g., Arubayi, 1987; Marsh & Roche, 1997). More recently, a review found that students’ ratings within the same class tend to be internally consistent (when teaching effectiveness was assessed through several items), reliable across students within the same class, and reliable across the same instructor across multiple courses (Benton & Cashin, 2014). Benton and Ryalls (2016) argue that calls for the complete removal students’ voices from the process is the wrong course of action. Both pay careful attention – and rebuttal - common arguments against the removal of students’ ratings.

Arguments made in defense of student ratings (while not necessarily using them for administrative decisions) are rooted in empirical studies and concrete recommendations for the utility of such ratings are put forth to readers that do not involve administrative decisions (Benton & Ryalls, 2016). Of note, certain arguments debunk certain points against student ratings. Hedged within a section titled “flagrantly false claims” and relevant to the current study: *SRI* [Student Ratings of Instructions] *Are Unreliable* largely based off the previously mentioned review (Benton & Ryalls, 2016). These claims regarding reliability are not new. Others shared similar thoughts that student ratings are highly reliable decades ago (e.g., Costin et al., 1971; Marsh, 1987; Overall & Marsh, 1980; Theall, 2003); however, we re-iterate these reviews have received extensive critiques (e.g., Stroebe, 2018; Uttl, et al. 2017). The question of reliability may need to be further investigated before making such strong claims.

**Perceived fairness**

Extant research tends to confirm that instructor evaluations are influenced by students’ grades, possibly pressuring some instructors into reducing the rigor of their course for the sake of attaining higher evaluation ratings (Greenwald & Gillmore, 1997; Marks, 2000). However, as pointed out by Wright (2008), students’ expectations of their final grades may not affect ratings nearly as much as their perceived fairness of the grading process. Professors who are consistent, representative, accurate, unbiased, and correctable in their grading may receive high evaluation ratings regardless of how much a student learns or what his/her final grade turns out to be (Horan et al., 2010; Leventhal, 1980). Thus, grades may predict evaluation ratings only so much as students perceive their grade *and* the processes by which they were determined as fair (Tata, 1999).

Additionally, the different facets leading into a final grade’s calculation may play on each other as students consider fairness in their evaluations. For example, Tripp and colleagues (2019) found that students’ perceived fairness of their instructors’ grading processes affected their perceived fairness of their assigned grade, which then translated to their evaluation ratings of teacher effectiveness. Perceived fairness of the course workload and difficulty may also be inversely related to perceived fairness of the grading process as a challenging professor may be thought of as less fair (Marks, 2000). Access to grading criteria, frequency of feedback, and proactive instruction are other aspects of grading known to explicitly affect perceived fairness (Pepper & Pathak, 2008); in turn, the fairness of these aspects must be factored in as well. Taken together, students’ perceived fairness of grading may be more akin to comprehensive assessments of the instructor rather than face-value judgements of their grade.

Though these interactions are notable, what they bode regarding the reliability and validity of measures of instructor effectiveness is relatively unknown. It is obvious that perceived fairness of grading is an integral part of these measures. However, given how intrinsic perceived fairness is to one’s instructor, interruptions in the continuity of student evaluation opinions may be more apparent in assessments of different instructors rather than between course types and/or across semesters. Moreover, the capacity of perceived fairness to measure teacher effectiveness may prove to be entirely irrelevant. If “outstanding” instructors in the eyes of students are those who consistently award good grades to all, their evaluations may yield little actual insight into their effectiveness as teachers.

**The current study**

The current study is similar in scope to recent work (e.g., Boring et al., 2016; Fan et al., 2019) in its calibration of teacher evaluations collected over an extensive period. Boring and colleagues’ (2016) investigation on both French instructors and U.S. teaching assistants’ gender ranged across five years; similarly, Fan and peers (2019) investigated the topic across seven. Their utilization of multi-sections has been described as the gold standard for researching students’ ratings. Thus, we aimed to follow their lead by analyzing the reliability of students’ ratings provided the same or different instructor, course type, and/or semester in addition to testing reliability over time. In doing this, we sought to explore the following research questions:

1. What is the reliability of instructor evaluations?
2. Are instructor evaluations reliable across time?
3. Is the average level of perceived fairness of the grading in the course a moderator of reliability in instructor evaluations?
4. Does the average variability in instructor fairness rating moderate reliability of instructor evaluations?

(Highlight: Amount of courses (6 vs, # of students, duration (5 year time frame vs ) our focus (gender vs on time and fairness)