Student Ratings and Course Modalities: A Small Study in a Large Context

Charles Dziuban
Patsy Moskal
Annette Reiner
Adysen Cohen
University of Central Florida

Abstract

This article examines the impact of course modality on student evaluation of courses and professors. Data were collected through the *Student Perception of Instruction* end of course rating form at the University of Central Florida (UCF), which contains nine items and maintains student anonymity. The findings indicate that while course modality accounts for only 1% of the variance in student evaluations, there is strong internal consistency and reliability in the rating scale. The distribution of ratings showed a concentration of scores at the high end, resulting in a high variability coefficient. However, when the long tail of low ratings was removed, the mean increased and the distribution became more symmetric, affecting various psychometric indices. The correlation matrices among the items revealed a single factor solution for each modality, suggesting that students tend to rely on a general impression when rating their courses. The multidimensional scaling process identified underlying categories such as structure, course climate, engagement, and consideration, even though students did not explicitly differentiate these elements in their responses. The study concludes that course modality has minimal influence on overall student ratings, a finding consistent across different time periods, including the COVID-19 pandemic. Although a single factor captures students' general evaluations, underlying categories shape their responses. The article also presents a classification model that predicts student ratings based on the scale items. This research addresses the complex dynamics of student evaluations, highlighting the nuanced relationship between course modality, student perceptions, and the underlying factors influencing their ratings.

Keywords: digital learning, student ratings, course modality, asymmetry

Dziuban, C., Moskal, P., Reiner, A., & Cohen, A. (2023). Student rating and course modalities: A small study in a large context. *Online Learning, 27*(3), 70-103. DOI: 10.24059/olj.v27i3.4053

As long as there has been post-secondary education, students have critiqued their educational experience. In recent decades this has become a standardized course rating form resulting in high stakes data for faculty evaluation, student selection of courses, and administrators who make personnel and programmatic decisions. However, educational technology, the COVID pandemic, questions about the value of higher education, and other issues have altered the conversation toward rethinking the rating process. One symptom of the change is the website ratemyprofessor.com that uses course evaluation to frame a faculty member's national reputation. This is the metaphorical tip of the iceberg because now students evaluate their courses on social media sites including Twitter, YouTube, Facebook, Instagram, TikTok, and others where their opinions gain traction. As we scrutinize the factors that impact student ratings, it is important to remember the observer dependence of the process, the proliferation of course modalities, the instructor's role, and what the ratings can do to improve the teaching learning process. All parties have a stake, but the psychological contracts involved make the situation complex. Contemporary education is much more than the sum of its parts because it is interconnected, interactive, diverse, adaptive, self-organizing, and emergent.

The Foundational Research

According to Wang and others (2009), student ratings of instruction evoke contradictory and conflicting responses dating back to the beginning of course evaluation. For instance, an entire issue of the *American Psychologist* addressed the validity, reliability, stability, usefulness, and dimensionality of the ratings (Greenwald, 1997). Dating back even further to the 1970s, the Dr. Fox phenomenon characterized an instructor who feigns student empathy, eliciting high ratings with strategies that have little or no relationship to effective teaching practice (Wang et al., 2009 via Williams & Ware, 1977). Further work in this area was extensive, using measurement and psychometric procedures to model the rating process (Algozzine et al., 2004; Gump, 2007; Marsh & Roche, 1997; Pounder, 2007; Wachtel, 1998). Generalized factor analysis approaches addressed underlying dimensionality (Bangert, 2006; Clayson, 1999; Cohen, 2005; Feldman, 1976; Lannutti & Strauman, 2006; Marsh & Roche, 1997; Smith & Anderson, 2005). Hypothesis-based studies used confirmatory models while other investigators incorporated methods such as cluster analysis (Ginns & Ellis, 2007) and visualization techniques (Abrami & D'Apollonia, 1991; Apodaca & Grad, 2005; Cohen, 2005; Ginns et al., 2007).

Causal and predictive approaches applied path analysis and structural equation modeling (Chang, 2000; Ginns et al., 2007; Greenwald & Gilmore, 1997; Renaud & Murray, 2005; Rinderman & Schofield, 2001; Shevlin et al., 2000) that augmented regression and correlational analysis (Cohen, 2005; Davidovitch & Soen, 2006; Eiszler, 2002; Nasser & Fresko, 2006; Read et al., 2001; Renaud & Murray, 2005; Sheehan & DuPrey, 1999; Stapleton & Murkison, 2001). A body of research applied hypothesis-testing models such as analysis of variance (Crumbley et al., 2001; Maurer, 2006; Renaud & Murray, 2005; Riniolo et al., 2006; Smith & Anderson, 2005) and chi square contingency analysis (Howell & Symbaluk, 2001). Another approach involved deductive analysis typified by studies that used criticism techniques to clarify the roles of students and instructors (Gump, 2007; Kolitch & Dean, 1999; Oliver & Sautter, 2005; Pounder, 2007). Any attempt to summarize this body of research converges on defining elements that underlie students' conceptions of excellent and poor instruction. The high-stakes nature of evaluations impacting university decisions such as tenure and promotion caused instructors to take their ratings more seriously. Contemporary analysis focuses on the validity of the process, examining students' decisions to engage meaningfully as well as how these ratings interact with, and are confounded by, multiple characteristics in the educational environment.

Evolving Contemporary Research

Course Modality, Level, and Content

The main purpose of this study is to address the impact of course modality on student rating evaluations. There is evidence that students in online courses are marginally less satisfied than with the in-person modality (Brocato et al., 2015; Capa-Aydin, 2016; Filak & Nicolini, 2018; Lowenthal et al., 2015; Mather & Sarkans, 2018; Sellnow-Richmond et al., 2020; Turner et al., 2018). While online students respond well to flexibility, convenience, and autonomy, they feel impacted by diminished feedback and interaction. Other findings show that students are more critical of professors teaching quantitative courses in general. Larger classes receive lower ratings as do those with heavy workloads (Lowenthal et al., 2015; Royal & Stockdale, 2015; Turner et al., 2018; Uttl & Smibert, 2017). There is conflicting research—Yen et al. (2018) found no differences in student ratings based on course modality. The consensus, however, is that course modality does have an impact on how students evaluate their educational experience, but it is not an overriding concern.

Instructor Characteristics

Factors such as instructor personality, temperament, and demeanor influence course ratings. Investigators examined distinct roles teachers take on and how this may affect their evaluations (Badur & Mardikyan, 2011; Foster, 2023; Kim & MacCann, 2018; Wang et al., 2009). Influencing issues include whether instructors are addressed by their first name or title and last name, how well-prepared they are, interest they show in their students' learning, and the attitude they display. An additional consideration is how instructors respond to evaluation and how they use the results (Floden, 2017; Golding & Adam, 2016). Some professors use the data to improve their courses or their teaching style. Others, however, discount the end of course evaluation process believing the opportunity costs outweigh any added value, but Mandouit (2018) concluded that student feedback is an important contributing factor and powerful stimulus for instructor reflection.

Student Characteristics

Social issues impact a student's decision to complete their end of course evaluation. According to Ernst (2014), they consider a multidimensional environment when determining if they will engage in the process: anonymity, avoiding social scrutiny, and the amount of time required. In addition to deciding *if* they should complete an evaluation, other issues determine *how* they complete the process. As student ambivalence increases, so do the number of dimensions they use to evaluate their courses (Dziuban et al., 2012). Griffin (2016) found that autonomy in courses leads to higher satisfaction, thus resulting in higher evaluation results. One research study found a strong association between a student's seriousness and dedication and the ratings they assign to the course or professor (Gunduz & Fokoue, 2021).

Bias and Validity Concerns

Recent studies emphasize concern about bias and validity in the student rating process. According to multiple sources, female professors receive lower ratings compared to their male counterparts (Boring et al., 2016; Boring et al., 2017; Buser et al., 2022; Chatman et al., 2022;

Flaherty, 2019; Heffernan, 2021; Mengel et al., 2019; Mitchell & Martin, 2018; Ray et al., 2018). For example, Ray and colleagues (2018) found that women instructors are held to a higher standard and must work harder to be viewed as competent. Even when female instructors exhibit similar performance to their male counterparts, they are rated significantly lower (Chatman et al., 2022). Often language is rooted in student evaluations that can lead to gender and racial biases (Genetin et al., 2021).

Aside from the bias regarding race and gender, students' perceptions of their own achievement impact their ratings. Researchers have confirmed this (Boring et al., 2016; Buser et al., 2022; Flaherty, 2022; Kogan et al., 2022; Scherer & Gustafsson, 2015; Stott, 2016; Stroebe, 2016; Tejeiro et al., 2018). Additionally, there is an imprecision in the relationship between student evaluations and instructor quality (Esarey & Valdes, 2020). Students may not have formed a well-grounded construct of what constitutes good teaching and might rate their professors solely based on extraneous elements such as confirmation bias, misaligned expectations, or indifference (Kornell & Hausman, 2016).

Because student evaluation is used to make high-stakes decisions (Flaherty, 2018; Kogan et al., 2022; Stark & Freishtat, 2014; Stroebe, 2016), there are assertations that the process results in grade inflation (Stroebe, 2016). With research indicating that women and racial minorities experience most student evaluation equity issues, more questions about the process arise. Therefore, validity constitutes the overriding concern with the course evaluation process (Hornstein, 2017).

A Final Thought on Research

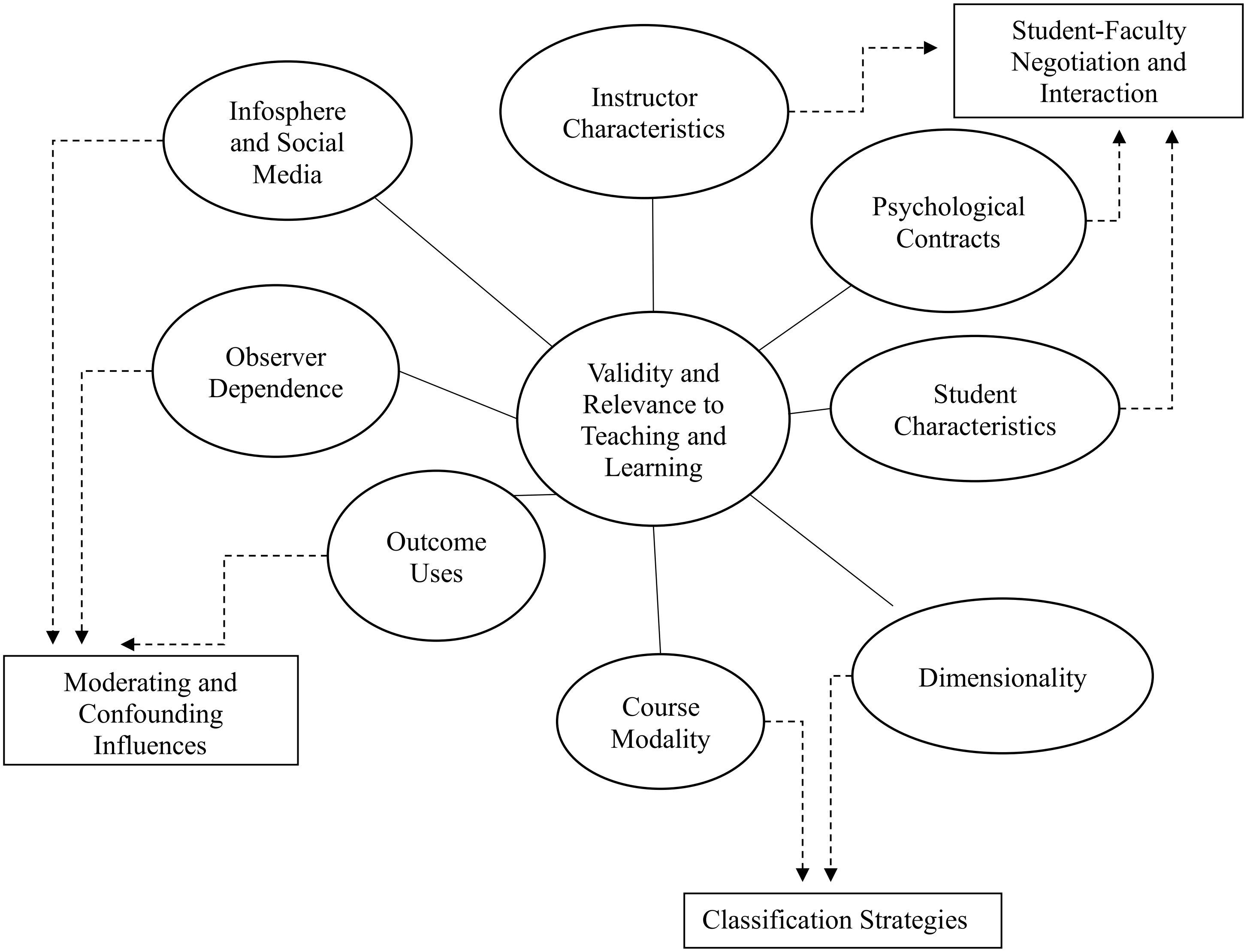
Like so many things in our accelerating educational culture, the student rating process has undergone significant reconfiguration. The initial research canon dealt with one modality, face-to-face instruction, where mitigating factors, such as class size, college, department, and discipline were bounded by the classroom walls and limited in the analysis methods that were available. However, in the decades after, there was an expansion of newly developed psychometric and multivariate techniques applied to the rating process. This was the psychometric period where excessive analysis sophistication may have obscured the end game in assessing meaningful teaching and learning. Most recently, the research emphasis traces social, cultural, and preconceived biases held by students toward instructors and courses. The digital age has changed the rules of the game and the boundaries for what is off limits. Apparently extraordinarily little is out of bounds. The reality is, however, the number of papers published on this topic is simply overwhelming. Consider ChatGPT's (2023) response to that number of articles question:

> A search on Google Scholar using the key words "student ratings" yields over 2.7 million results…using the key words "student ratings" AND "higher education" …yielded over 167,000 results.

By any stretch of the imagination, this makes a traditional review of research intractable and no longer realistic. Our current review is evidence of this. The references we have on the topic represent far less than 1% of ChatGPT's estimated 167,000. The chance we missed something important is certain. Kabudi, Pappas, and Olsen (2021) proposed another approach using surrogate large language models, or generative artificial intelligence, identifying prototype categories from which AI can organize a search that identifies research clusters and their interconnectedness. Of course, the constraining factor hinges on organizing the search parameters. A different organizational scheme will yield different results. For instance, one might accomplish it by data analysis methods used, topic, modality, or any other structure. With this process, a graphic result emerges—one that facilitates an investigator understanding the research environment. The platform manages the overwhelming and tedious workload; however, a similar result can be obtained with semantic intelligence, which is what the current authors have done. Each one designed their concept of the complex systems underlying our meager review of research. The composite result is presented in Figure 1, but note that all models are approximations to theories, constructs, and concepts.

The model is hub and spoke revolving around the validity core impacted by components that self-organize into student and faculty negotiation and interaction, classification strategies and moderating and confounding influences. Through a careful review we have assigned (using intuition and judgement) proximity vectors to the spokes. Closer elements indicate greater impact. We will discuss this with a thought experiment later, but this literature model is constrained by two dimensions and individual component interaction with the foundational base. Obviously, those influencing components (eight of them) can and do interact with each other in an extremely complex pattern that is difficult to deconstruct. We are confident they influence student ratings but unraveling the high order model is beyond the scope of this research. Therefore, we become apologists for conducting a small study of student end of course ratings.

Figure 1
An Underspecified Model of Elements That Impact Student Ratings


Student-Faculty Negotiation and Interaction
Infosphere and Social Media
Instructor Characteristics
Psychological Contracts
Observer Dependence
Validity and Relevance to Teaching and Learning
Student Characteristics
Outcome Uses
Dimensionality
Course Modality
Moderating and Confounding Influences
Classification Strategies

The Study and Research Methods

This study depicted in Figure 2 is small not because of the sample size that, accounting for missing data, surpasses 660,000 student responses to their courses at the University of Central Florida for the years 2017 through the 2022 fall semester (N = 664,473). The work is small because course modality is the primary independent variable (with a sidebar for COVID's impact), ignoring the remaining influences identified from the literature (Figure 1). There are reasons for this. The most significant is the complexity issue. The second justification for modality research is the number of studies on the topic. With the onset of the pandemic, multiple course contexts emerged in an extraordinary attempt to keep university doors open (Appendix B).

The "big data" approach in this study, however, renders concepts like statistical significance and standard error moot, reframing sampling, estimation, and hypothesis testing into modeling, small pattern recognition, and machine learning. The research model (Figure 2) defines a study where the student rating process undergoes examination across modalities to identify information, meaning, and outcomes that transcend analysis strategies. This is a weakness but a strength as well because it attempts to clarify whether, and how, course modality impacts students' framing of their educational experience. We made other decisions to prevent this article from becoming unwieldy by omitting the derivation and formulas for the analysis, but references for the reader are provided should they choose to pursue them. We have, however, included a rationale for each data analytic procedure.

The Data Collection Protocol

The end of course student rating form entitled *Student Perception of Instruction* was the source of data for this research (Appendix A). The scale resulted from a series of faculty, student, and administration groups working collaboratively to modify and improve the process. The instrument contains nine items. The final design was approved by the faculty and student senates and was first administered in 2013. In addition to the protocol redesign, the committees addressed the strengths and weaknesses of this approach and specified the ethical use of the data for faculty evaluation and professional development. The instrument is student-anonymous, preventing identification of individual respondents. Administration takes place online for all classes, irrespective of modality, managed by the university's division of information technology that also summarizes the results by course and presents the findings to the faculty members, augmented with normative data. Instructors and departments make individual determinations about data use, but these data are used in promotion and awards.

The Analysis Procedures (Figure 2)

1. Modality impact was assessed by summing the nine items with a maximum score of 45 and a minimum of 9 (5 = excellent, 1 = poor). The mean differences across course modalities were analyzed with a one-way linear model, discounting significance level in favor of the ETA squared effect size estimation (Richardson, 2011). The index gives the percentage of variance accounted for in the dependent measure (total score) by the independent variable (course modality). However, recommendations for interpreting ETA are rules of thumb so that judgment by the investigator about impact is required.
2. The impact on the total scores of the three COVID periods (pre-2017–2019, 2019–2020, and post-2021–2022) were determined with methods identical to those used with course modality described above.
3. The Alpha reliability coefficients, average item total correlation, skewness index, and coefficient of variation for the rating scale results for each modality were calculated – the classical measurement model (Crocker & Algina, 1986). In any study, a requirement is that the investigators become familiar with their data. What are the moments of the distributions? Are there missing data and if so, are they of consequence or can they be ignored and not appreciably impact the findings? What are the distributional characteristics? What are the measurement properties?
4. The domain sampling characteristics of the instrument were indexed using the measure of sampling adequacy (MSA). This is the measurement sampling issue.

Statistical sampling answers the question, “Do I have a good sample of subjects from an identified population? Domain sampling answers the question, “Do I have a good sample of items from a measurement domain in which I am interested?” It is the other sampling issue (Dziuban & Shirkey, 1974; Kaiser & Rice, 1974). Without verification the results can misrepresent the underlying measurement issues that are fundamental to valid research.

5. The latent components of the student responses to the rating scale were determined for each modality. The question was one of multidimensionality, and, if it existed by modality, what were the pattern differences. This was accomplished with the factor analytic model by examining the Eigenvalues of the item correlation matrices for each modality. As a criterion for dimensionality Eigenvalues greater than one are customarily used for factor retention. Once the factor(s) were removed from the system and the residual correlation matrix determined, the MSA was calculated to determine if what remained was random noise (Dziuban & Shirkey, 1993; Hill, 2011; Kaiser, 1968; Kaiser & Rice, 1974). The Eigenvalue criterion is another rule of thumb that is used extensively, but it remains to the researchers to determine if that method makes sense for the objectives of the study. This analysis technique bases itself on the proposition that the multiple relationships among the rating scale items can be explained by a smaller set of underlying constructs that are not directly observable. Should more than one factor or component result, the interpretation becomes more complex and relies on the knowledge, insight, and intuition of the investigator. There is subjectivity in the process because of the observer dependence phenomenon. There is a world of interpretation difference between one factor and more than one.
6. The Euclidean distances among the items for each modality on the instrument were derived and subjected to the multidimensional scaling procedure (MDS) to create a visual portrayal of the relationship of the items across teaching contexts. This was an augmented approach to assess how students characterize their educational experience (Borg & Groenen, 2005). In a metaphorical sense this involves examining student ratings at the “quantum level” where one can visualize what is not available to the naked eye. MDS initiates by identifying pairwise similarities between objects, in this case the items of the rating scale. Next, the distances among items are converted into coordinates that can be mapped into a lower dimensional space. The objective is to minimize the differences between the original similarities and those specified by the derived coordinate mapping.
7. Finally, a predictive model was developed for whether students assigned an overall rating of excellent to their courses using classification and regression trees (CRT) (Brieman et al., 1984). The variables assessed for productive power were college, department, course level, modality, term, and the remaining items on the rating instrument. The objective was to develop the simplest and most accurate decision rule for a student rating a course excellent. CRT recursively separates the data into smaller subsets determined by the predictors. At each step of the iterative process variables are selected that most efficiently sort the dependent measure into classes by reducing the variance. The splitting process continues until a predetermined stopping criterion is reached or variance reduction is no longer achieved.

Figure 2
A Hub and Spoke Model of the Research Process

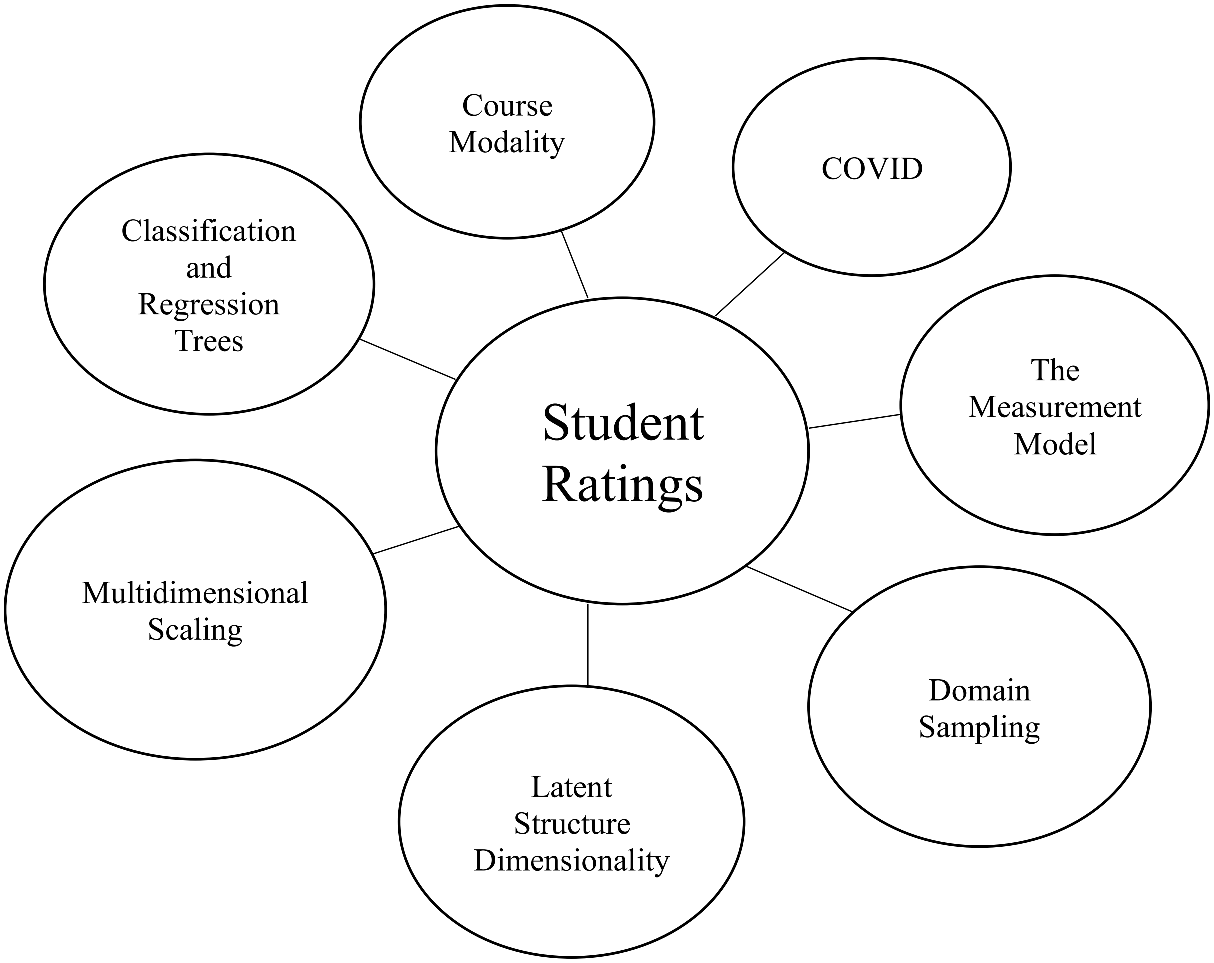


Table 1 and Figure 3 present the end of course rating scale total scores for each modality. The means show minimal variation with an overall value of 32.3 and a standard deviation of 9.0. The ETA squared effect size shows that modality accounts for 1% of the variance in how students evaluate their courses; however, the alpha reliability coefficient of .94 indicates strong internal consistency. The average item total correlation was .78 supporting the reliability finding. Examining indices for classical measurement models the results get a “pass.” However, the skewness indices in Table 1 show a “piling up” of the scores at the high end of the distribution with a summary value of -.64. Figure 3 confirms this visually. We are looking at the long tail with relatively few students using the low extreme of the scale. This level of asymmetry inflates the variance of the distribution producing a variability coefficient of 29—a value considered high. The high-end concentration creates a mean of 32.3 that is 72% of the total possible score. The median (34) represents 76% of the possible total and the mode of 4 is 80% of the highest possible assignable rating value of 5. But what if we cut off the long tail by removing scores in the first quartile? With the long tail gone the mean increased to 36.8, 82% of the possible total and the median increased to 37, 82% of that. The mode of 4 was unaffected by the shape of the distribution. Note that as the altered distribution became more symmetric the mean and mode converged. The effects on other indices were noteworthy. The standard deviation decreased to

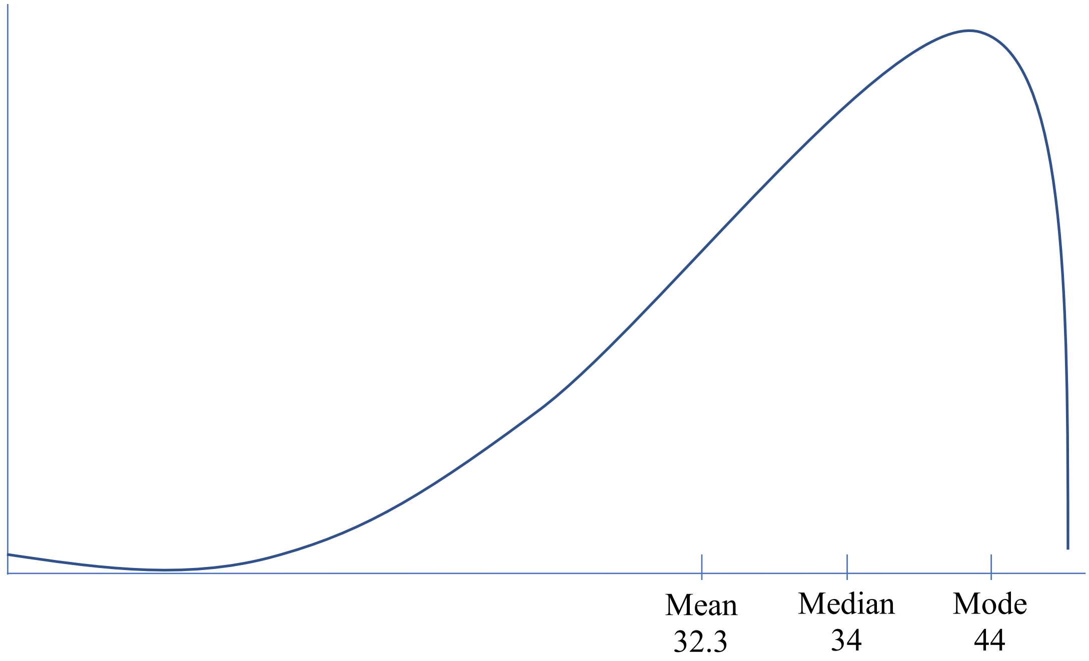
5.2, a drop of 42%. The skewness decreased to -.26, down 60%. Reliability was still high at .84 but showed a decrease from the original .94 and the item average item total correlation dropped to .54—a decrease of 31%. The coefficient variation decreased to 14 dropping 52% from the value in the asymmetric distribution. Distribution characteristics for student ratings of their courses influence the results and how they should be interpreted.

Table 1
Total Score Summary Across Course Modalities

Modality	N	Mean	S.D.	Cronbach's Alpha	Average Item Tot. Cor.	Skewness	Coef. of Var.
P	313,306	32.3	9.0	.94	.77	-1.6	28
WW	176,440	32.9	8.7	.94	.77	-1.4	26
M	64,795	33.3	8.7	.94	.76	-1.4	26
RS	17,875	28.9	9.6	.95	.79	-0.6	33
RA	11,134	30.7	9.0	.94	.78	-0.8	29
V	16,750	30.6	9.3	.95	.79	-0.8	30
R	8,912	30.4	9.3	.94	.78	-0.7	31
RV	5,654	27.7	9.6	.95	.79	-0.3	35
V1	49,607	32.0	9.2	.94	.78	-1.1	29
Total	664,473	32.3	9.0	.94	.78	-0.64	29

*Eta-squared = .01

Figure 3
Student Ratings—The Long Tail



A comparison of the total score differences by COVID periods in Table 2 indicates a similar result to the modality analysis. Remember, we have toggled back to the asymmetric version of the distribution for the remaining results. However, in this case the ETA squared showed that none of the variance in total score course ratings were attributable to the pandemic related educational program adjustment at UCF.

Table 2
Total Score Summary Across COVID Periods

COVID Period	N	Mean	S.D.	Average Discrimination	Skewness	Coef. of Var.
Pre-COVID	287,770	32.4	8.9	.77	-0.6	27
During COVID	187,735	32.4	9.0	.76	-0.7	28
Post-COVID	189,038	32.1	9.1	.78	-0.6	28
Total	664,473	32.3	9.0	.78	-0.6	29

*Eta-squared = .00

Table 3 presents the results of the domain sampling characteristics of the rating scale items for each modality and for the overall cohort. The measures of samplings adequacy (MSA) were all in the mid .90s, which according to Kaiser & Rice (1974) comprise an excellent sample of items for the domain. MSA is known to be sensitive to the number of variables, sample size, and number of factors; however, apparently not impacted by distributional characteristics. The average correlations across all items for each modality were in the .60s. Both findings indicate that from a measurement and psychometric perspective these are satisfactory results.

Table 3

Domain Sampling for the Course Modalities

Modality	N	MSA	Avg. Correlation
P	313,306	.94	.64
WW	176,440	.94	.64
M	64,795	.94	.63
RS	17,875	.95	.67
RA	11,134	.95	.65
V	16,750	.95	.66
R	8,912	.94	.65
RV	5,654	.94	.66
V1	49,607	.95	.66
Total	664,473	.94	.64

Table 4, Figure 4, and Table 5 present the Eigenvalue summaries for the correlation matrices among the items for each course modality. The average correlation was calculated according to the Kaiser (1968) procedure. Remembering that the rule of thumb is to retain factors for transformation and interpretation corresponding to those values greater than 1, Table 4 demonstrates a Spearman case where there is only one factor. This finding has precedent in the literature (Dziuban et al., 2018). For each modality, the single factor accounts for approximately 70% of the total variance in the system. Figure 4 depicts an Eigenvalue graph suggested by Cattell (1966) for determining the number of factors to retain. He posited that extraction should be terminated at the point where there is a noticeable break in the curve. This procedure supports the one-factor solution. According to this analysis model, students discount the individual elements on the scale and simply "go with their general impression." The results in Table 5 show the MSA and average correlations for the residual matrix. All MSA values were in the .50 range, indicating that nothing but random variation (noise) remained. The average correlations confirm this with all values being 0 to the first decimal place. One factor cleaned out the system.

Table 4

Eigenvalues for the Course Modality Factor Solution

	P	WW	M	RS	RA	V	R	RV	V1	**Total**
1	6.12	6.12	6.04	6.32	6.21	6.29	6.21	6.29	6.26	6.15
	68%	68%	67%	70%	69%	69%	69%	69%	69%	68%
2	.73	.77	.77	.74	.71	.72	.73	.74	.71	.74
3	.48	.45	.49	.43	.45	.43	.44	.43	.43	.45
4	.45	.42	.46	.37	.38	.40	.40	.38	.42	.44
5	.34	.35	.35	.32	.33	.33	.34	.33	.33	.34
6	.30	.27	.29	.25	.28	.28	.27	.24	.29	.29
7	.24	.26	.25	.23	.26	.24	.24	.23	.24	.25
8	.21	.20	.21	.20	.22	.18	.21	.20	.19	.20
9	.14	.15	.15	.14	.15	.14	.16	.15	.14	.14

Figure 4
Eigenvalue Scree Test for the Number of Factors

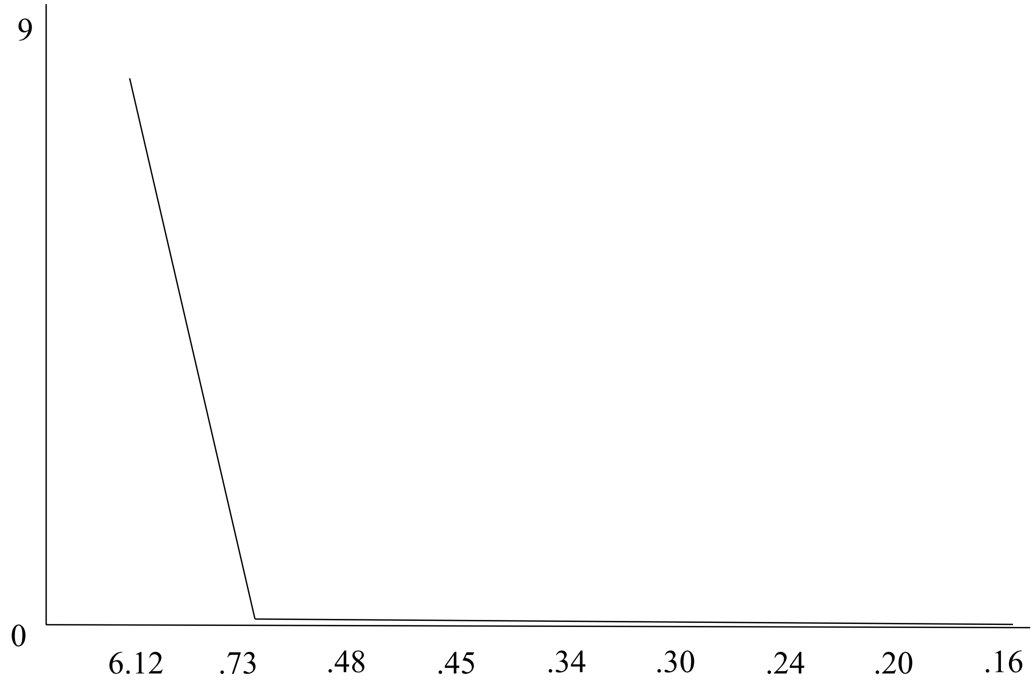


Table 5
Residual Correlation Domain Sampling Results

Modality	N	Residual MSA	Avg. Residual Correlation
P	313,306	.51	-.02
WW	176,440	.53	-.03
M	64,795	.52	-.05
RS	17,875	.58	-.02
RA	11,134	.54	-.03
V	16,750	.52	-.02
R	8,912	.59	-.02
RV	5,654	.57	-.04
V1	49,607	.51	-.03
Total	664,473	.55	-.02

Tables 6, 7, and 8 contain the coordinates for a two-dimensional multidimensional scaling of the items for each modality, the overall cohort with the stress on the system and the squared correlation between ordering with the Euclidian distances and those of the scaled solution. Table 6 shows close coordinate correspondence for the first dimension across modalities with an average correlation among them of .97 (94% variance accounted for). The same is true for Table 7 for the second dimension with an average correlation of .87 (76% variance accounted for). Table 8 confirms acceptable stress levels for each modality (approximately .10) and high squared multiple correlations (all in the mid .90s). The multidimensional scaling solutions for each course modality were close versions of each other. A forced two-dimension analysis is reasonable and facilitates interpretation.

Table 6

Coordinates for Dimension One of the Multidimensional Scaling of the Items

Items	P	WW	M	RA	RA	V	R	RV	V1	**Total**
Organization	1.7	1.9	1.8	1.7	1.8	1.6	1.5	1.9	1.8	1.9
Expectations	1.5	1.6	1.6	1.8	1.8	1.7	1.7	1.5	1.7	1.7
Communication	0.6	0.3	0.8	0.6	0.5	0.7	0.8	0.7	0.2	0.2
Respect/Concern	0.7*	0.6	1.0*	0.2	0.7	0.0	0.1	1.1*	0.4*	0.3*
Interest	1.7*	1.5*	1.4*	1.5*	1.4*	1.3*	1.8*	1.6*	1.7*	1.7*
Learning Environment	0.4*	0.5*	0.5*	0.7*	0.7*	0.2*	0.8*	0.7*	0.1*	0.1*
Feedback	0.6*	1.8*	0.8*	1.5*	1.7*	1.8*	1.1*	0.3*	1.4*	1.4*
Achievement	0.4*	0.4*	0.5*	0.5*	0.7*	0.4*	0.3*	0.3*	0.1*	0.1*
Instructor Effectiveness	0.0	0.2*	0.1	0.0	0.3*	0.2*	0.1*	0.1*	0.1*	0.2*

Average r = .97
*Denotes negative values

Table 7

Coordinates for Dimension Two of the Multidimensional Scaling of the Items

Item	P	WW	M	RS	RA	V	R	RV	V1	**Total**
Organization	0.6	0.0	0.2	0.9*	0.5*	1.2*	0.8*	0.7*	0.5	0.6
Expectations	0.4*	0.5	0.4*	0.0	0.4*	0.3*	0.0	0.3	0.4*	0.5*
Communication	0.4	0.7	0.3	0.1*	0.1*	0.3	0.0	0.6	0.7	0.6
Respect/Concern	2.3*	1.6*	2.0*	2.1	1.7	1.9	2.1	2.0	2.0*	2.0*
Interest	0.3	0.9*	0.1*	0.2	0.9	0.7	0.4	0.4*	0.1	0.0
Learning Environment	0.1*	0.2*	0.2	0.1*	0.2*	0.1*	0.1*	0.3*	0.2*	0.1*
Feedback	1.5	1.1	1.7	0.9*	0.9*	1.1*	1.3*	1.5*	1.3	1.3
Achievement	0.1*	0.2	0.2	0.1*	0.3*	0.2*	0.3*	0.1	0.0	0.0
Instructor Effectiveness	0.0	0.1	0.1	0.2*	0.1*	0.1*	0.0	0.2*	0.1	0.1

Average r = .87
*Denotes negative values

Table 8

Stress and R-Squared for the Multidimensional Scaling of the Items (Modality and Total)

	P	WW	M	RS	RA	V	R	RV	V1	**Total**
Stress	.11	.10	.10	.08	.08	.07	.08	.06	.12	.11
RSQ	.94	.94	.94	.97	.96	.97	.97	.98	.93	.94

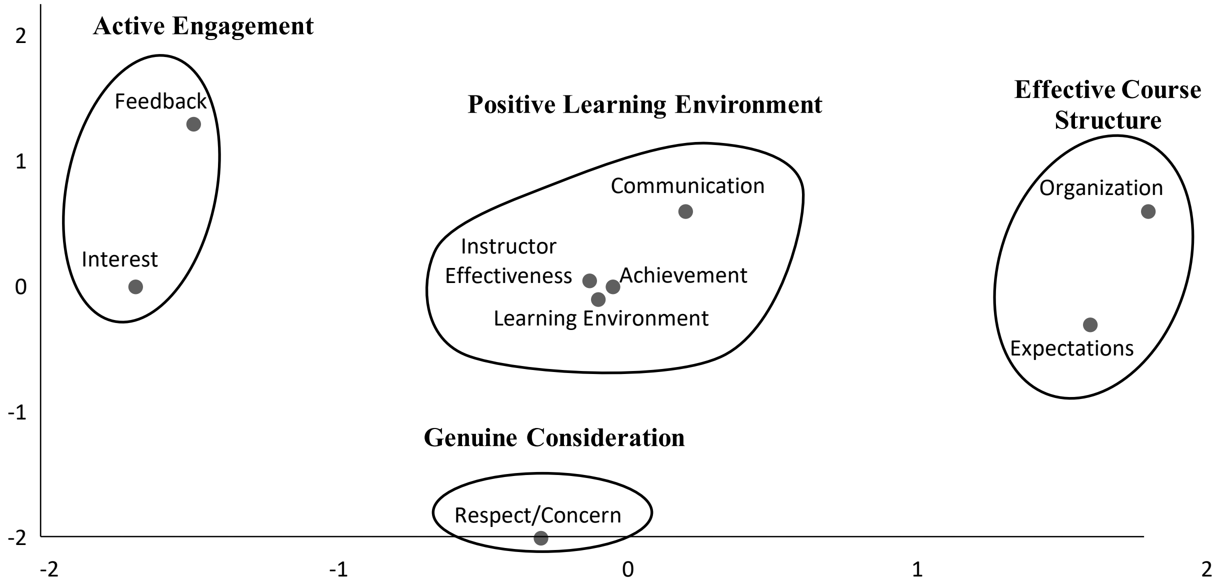
Figure 5 presents the rating scale items located in the two-dimensional space according to their coordinate values. According to the map, students respond to the quality of their educational experiences by:

1. Effective Course Structure
2. A Positive Learning Environment
3. Active Engagement
4. Genuine Consideration

This result corresponds to our literature and design in Figures 1 and 2, however, in this case the effective learning environment is supported by four different elements.

Figure 5

A Two-Dimensional Scaling of the Student Rating Items



Stress = .11
RSQ = .94

The results for the classification and regression tree are presented in Table 9. The dependent variable was whether students assigned an overall rating of excellent. The independent variables were characteristics of the educational environment—course modality, college, department, term, and level, plus the remaining 8 items on the end of course rating scale. When the analysis converged, the process eliminated all possible predictive variables except for two. If students believe that an instructor *achieved the course objectives* and *created an effective learning environment*, the probability of them specifying an excellent course is .82. Little else impacts their decision.

Table 9
A Decision Rule for a Faculty Member Receiving an Overall Rating of Excellent (n = 342,386) if a student responds

	Excellent	Very Good	Good	Fair	Poor
Achieve Objectives	√				
Learning Environment	√				

*The probability of an overall rating of Excellent = .82
Split half validation

Final Thoughts About the Results

This study sought to identify outcomes about student end of course ratings that were independent of analysis strategies. Unfortunately, the results only partially answer that question. Some findings are consistent, but some are not. From a measurement perspective, the ratings conform to quality specifications. They are internally reliable, produce small standard errors, feature items that are positively correlated with the total score, and present excellent psychometric (domain sampling) characteristics. However, except in relatively rare instances, students tend not to assign poor ratings to their classes. The upper end of the scale is used extensively creating noticeable asymmetry. This long tail circumstance creates measurement artifacts that when removed make meaningful assessment of teaching effectiveness difficult if not impossible because the ratings are so similar. If one were to grant the rating validity assumption, then most instructors are “very good” or” excellent” with a small percentage of poorly rated outliers.

Course modality exerts minimal influence on students’ overall ratings accounting for virtually none of the variation. The same finding was true for the three COVID pandemic periods. Further, when students rated their classes, a general component was identical for all modalities. The factor model was unable to identify any underlying response patterns. However, at a more granular level the scaling process was able to partition the single factor by structure, course climate, engagement, and consideration. Although students did not differentiate these elements directly in the data, they underlie their responses. This is the conundrum, only one factor but categories under the hood. Finally, the classification model produced a simple and clear prediction protocol. Nothing in the university or course organizational structure predicted this, but the rating scale items did.

Some Thought Experiments

Complexity

This study is about removing friction from the course evaluation system in a manner that helps us better understand teaching and learning. The process is complex because the whole exceeds the sum of the parts and is in constant flux. Consider our two hub and spoke models in Figures 1 and 2. Instead of constraining them into two dimensions, what if we cast them in three spaces with the spoke modules becoming orbiting satellites where the environment changes from moment to moment? This seems more reasonable and what Page (2009) described as dancing landscapes that are often unpredictable but at times self-organizing. If a university class is emergent (we believe it is) and its composition arises from the interactions among multiple individual elements such as instructor, students, curriculum, achievement, opportunity, technology, and modality then quality may not be captured by a single rating session. Taleb (2018) portrays complexity this way:

> The main idea behind complex systems is that the ensemble behaves in ways not predicted by its components. The interactions matter more than the nature of the units. Studying individual ants will almost never give us a clear indication of how the ant colony operates. For that, one needs to understand an ant colony as an ant colony, no less, no more, not a collection of ants. This is called an "emergent" property of the whole, by which parts and whole differ because what matters are the interactions between such parts. And interactions can obey very simple rules. (p. 69)

A class is a small world network where individuals are connected, and others are independent of each other. This contributes to complexity. There are both positive and negative feedback loops in a class—some reinforcing and some canceling. This is not amenable to a simple solution, but it is an important problem. If it were simple the answer would be linear and predicable, but a university classroom is unpredictable and nonlinear—this is not a new idea.

Psychological Contracts and Observer Dependence

Perhaps a useful way to conceptualize a class is through a series of psychological contracts that frame the expectations students have for their instructor and conversely the expectations an instructor holds for students (Dziuban et al., 2012). Effective teaching and learning require well-formed contracts built on positive relationships and mutual understanding. What if a class is not a unitary thing but a series of individually negotiated contracts between the instructor and students that are constantly renegotiated? If this assumption holds then each student is reacting to and evaluating a separate learning experience where data aggregation is not meaningful.

This frames observer dependance where the class is not a fixed construct but defined by student perception of it—for instance, ideas corresponding to quality, color, taste emotion, time perception, personal identity, memory, morality, political views, success, humor, and self-awareness. This concept formed the basis of Snygg and Combs (1949) work on symbolic interactionalism that is closely related to a phenomenal field where people create their personal meanings with subjective, rather than objective, experiences. Searle (1995) also referenced observer dependence—contending that qualities of an object (a class) depend on the perspective of the observers' assumptions or expectations. Pirsig (2006) concurs by examining the nature of quality and the fundamental difference between the subjective and objective experience arguing that quality cannot be fully understood with metrics. Is it possible that student ratings result from a series of individually negotiated contracts that are moderated by some degree of confirmation bias. Snygg and Combs (1949) provide a graphic example of the phenomenon:

> Several years ago, one of the authors was driving a car at dusk along a western road. A globular mass about two feet in diameter suddenly appeared directly in the path of the car. A passenger in the front seat screamed and grasped the wheel, attempting to steer the car around the object. The driver tightened his grip and drove directly into it. In each case the behavior of the individual was determined by his own phenomenal field. The passenger, an Easterner, saw the object in the highway as a boulder and fought desperately to steer the car around it. The driver, a native of the vicinity, saw it as a tumbleweed and devoted his efforts to keeping his passenger from overturning the car… the behavior of each was determined, not by the objective facts, but by his own phenomenal field. (p. 14)

Course Modality as a Treatment Effect

Multiple studies cited in this paper examined course modality as a treatment effect that impacts student ratings. Treatments assess the influence on dependent measures among groups that receive a particular intervention. To ensure outcome integrity confounding factors must be eliminated or controlled for statistically. However, modality of a college course is subject to uncontrollable effects such as availability of physical space (if required), scheduling issues, instructional design and technology support, curricular discipline, university, college and department policy, and student motivation and economic status. Each one of these is a confounding factor. Perhaps better concepts for course modality might be context, learning environment, pedagogical approach, boundary object, or idealized cognitive model—none of which can be reasonably considered treatment effects but rather nuanced notions of modality. Certainly, the COVID pandemic led to class definitions that are fluid, flexible, and in a continual churn. While some have been successful, some have not fared nearly as well and were quickly abandoned. Learning happened in an uncontrolled spontaneous environment outside the class and university making any attribution to outcomes based on modality virtually impossible. Our data indicate that modality accounts for virtually no part of the variation in student ratings. In our judgement, Rosch's (1973) prototype theory is the best characterization. These prototypes serve as benchmarks against which we evaluate other examples of a category that can be basic, superordinate or subordinate. The basic level prototype is the most generally accepted and acknowledged category—superordinate refers to a broad, general class that encompasses subordinate categories that fall within its domain. For example, for a typical online prototype designation, learning management system (LMS) superordinate categories might be Moodle, Canvas, and Blackboard, with subordinate categories:

1. Content Management Systems (CMS)
2. Assessment and Testing
3. Collaborative Learning
4. Adaptive Learning
5. Gamification
6. Mobile Learning
7. Analytics and Reporting

Within the context of this study the prototype is modality. The superordinate category is learning logistical arrangements and the subordinates are:

- Face-to-face (F2F)
- Mixed mode/blended with reduced F2F
- Blended with active learning
- Blended with no more than 20% F2F
- Lecture capture with live option
- Lecture capture
- Video-streamed
- Emergency remote instruction
- Fully online

Prototypes can clarify our understanding of classes and their definitional boundaries by providing specifications for subordinate course categories that are the operational versions of modality.

If we are correct that prototypes are the best descriptions of modality, we should be mindful that they are not fixed or universal. They are context dependent, have blurred boundaries, and may vary according to situational circumstances. Therefore, this does remove course modality as a candidate for a treatment effect. However, modalities as prototypes comprise a functional framework for helping understand contemporary teaching and learning.

Back to Asymmetry

In the results section we highlighted the asymmetric characteristics of student rating data that causes interpretation difficulties. We used an arbitrary shortening of the long tail to demonstrate the impact on skewness. This was a device for demonstrating the change in central tendency and variability. However, other outcomes were impacted by increasing the symmetry. The effect size went to zero. The average correlation among items dropped from .63 to .33. The MSA decreased to .81 and reliability decreased to .86. The factor model produced two components, communication and engagement, that were highly correlated. However, the multidimensional scaling produced identical coordinate maps for the both the symmetric and asymmetric data sets. The fact is, however, that cutting off the long tail invalidated the data because student ratings are markedly asymmetric. Parenthetically asymmetry is a significant contributor to inequality and unequal opportunity in fundamentally all forms of human endeavor, culture, and society (Andersen, 2018; Benjamin, 2020; Boghosian, 2020; Eubanks, 2019; Friedman & Laurison, 2019; Giridharadas, 2018; Gumbel, 2020; Isenberg, 2017; McGhee, 2021; Mlodinow, 2009; Mukherjee, 2016; O'Neil, 2018; Safir & Dugan 2021; Taleb, 2005; Taleb, 2007; Taleb, 2012; Taleb, 2018; Wilkerson, 2020). As we consider the student voice and end-of-course ratings in higher education, we should address the apparent asymmetry involved in modality prototypes.

This is particularly important at this time when so many aspects of higher education and faculty life are under assault and the current student generation has lost much of its enthusiasm for attending post-secondary education. Consider the following from Bryant (2022) paraphrased by Bush (2023):

> The last decade of social change, low birth rates, diminishing support from state governments, COVID-19, and student demands have slowly and severely weakened higher education's market value. Experts identified these events as the first death knell of a college enrollment crisis. The consequences look bleak--56% of high school graduates have no plans to attend college or are uncertain that they will ever attend. (p. 1)

Disappearing Class Boundaries

There can be no question that higher education is experiencing a revolution. Floridi (2014) explained it as the spoken word, the written word, the printed word, and the digital word that now encompasses the generative artificial intelligence word. Mukherjee (2016) framed it as the atom, the bit, and the gene. We are on the precipice of monumental understanding of the cosmos, information, and life. But what is so astounding is that both these revolutions are related, interconnected, and intimately bound up with each other. Scientific and linguistic boundaries are melting away, so we should not be surprised that what comprises a college class is undergoing a similar transition with its boundaries leaking, disappearing, and being absorbed into the information age. With learning technology making information instantly available, how we conceptualize education is radically altered.

In the digital age, students (meaning everyone) have access to a vast array of resources, platforms, and educational materials. They can learn beyond the confines of a single class or university curriculum, pursuing levels of knowledge and insight well beyond what a semester provides. Students can connect with likeminded people around the planet through online forums, discussion boards, virtual communities, and social media, expanding interconnectedness and creating an agile and community-based learning environment. However, there is a caveat. Although college classes may be escaping their evolutional boundaries, they still provide a learning framework for a systematic educational progression. The information age stretched the roots and twisted the vectors of traditional college organization. Although the bounded class is vanishing and boundary-crossing learning is becoming the norm, there is added value to structured learning and interaction; however, as Page (2006) describes complexity in the modern age, the learning landscape is dancing.

Conclusion

The Future of the Future

Teachers

Teachers are the human capital and reputational foundation of higher education. They shepherd information, knowledge, insight, and wisdom. Those of us who have worked at our craft know the joys, excitement, and pleasure that come from the "classroom" but also know how exhausting, frustrating, and fragile it can be. Nothing feels better than teaching well. Nothing feels worse than doing it poorly—and we have all done it poorly at some point in our careers. There is an anonymous aphorism, "I thought I understood it until I tried to teach it" and another attributed to Thomas Edison, "There were days of such discouragement that I ached to give it all up." Teaching is a demanding profession. Faculty members contribute so many resources: expertise and knowledge, research and innovation, communication skills, mentoring and guidance, intuitional reputation, community service, critical thinking, networking, diversity, advising, thought leadership, alumni relations, and human kindness. The list is long, but it only scratches the surface. Understanding the breadth and depth of what faculty undertake year after year is vitally important. The most important outcome of teaching, long-term impact, is the real measure of how effective an instructor has been over her or his career. Good teachers have bad days, so context becomes a parceled-out covariate.

Change Happens

The resiliency, dedication, and creativity of university faculty over the past decades has been remarkable. In addition to the pandemic and its yet-to-be-determined long-haul impact on America's higher education system, advances in educational technology have been dramatic. A brief list might include online and blended learning, open educational resources, mobile learning, microcredentials, adaptive learning, gamification, active learning, large language models, internationalization, student centered learning, and cloud-based learning. However, like all our lists in this article this one does not pretend to be comprehensive. But consider this quote (Gelsinger, 2018):

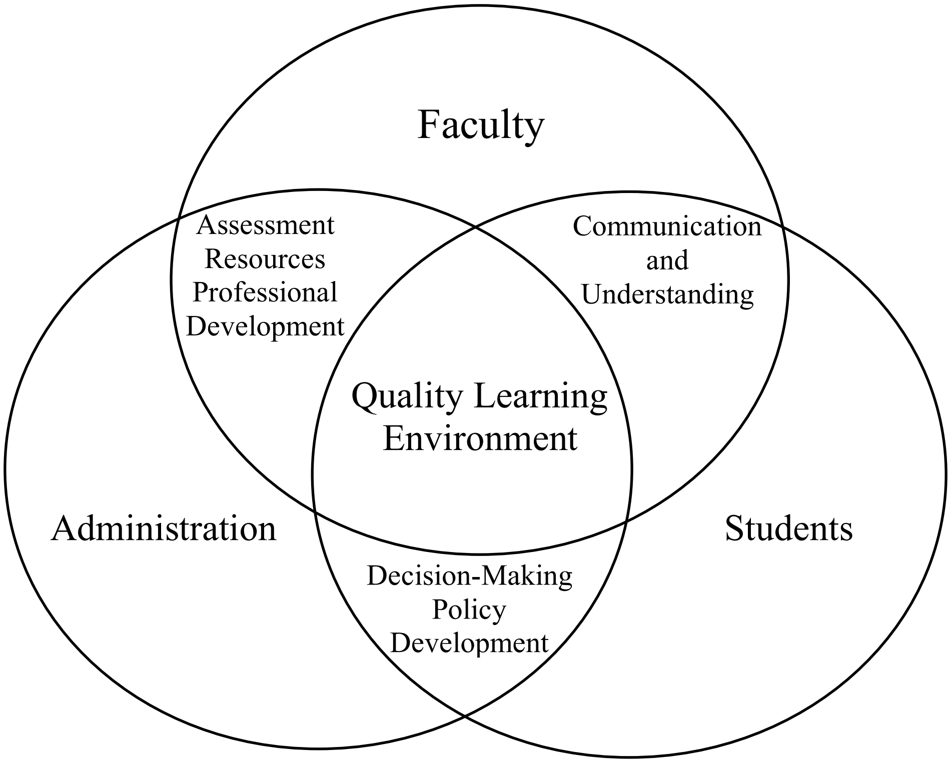
> It may feel like the pace of technology disruption and change these days is so dizzying that it could not possibly get any more intense. Yet here's the science fact: the pace of change right now is the absolute slowest it will be for the rest of your life. Fasten your seatbelts. It's going to be a fascinating ride. (p. 7)

At times, change is forced upon us, and we must adapt or be passed by. However, there is another amorphism: "The more things change the more they stay the same." That seems to be true with student ratings of instruction. We can find articles that date as far back as 60 years ago about designing a good end of course evaluation form. Since then, there have been creative approaches to making evaluations of teaching part of the educational culture, but the emphasis remains on the form. Perhaps we should start with fundamental questions. Why do we do it and how will the results be used? A starting point might be Muller's (2018) checklist:

1. What is it that you really want to assess and are there any valid indicators available?
2. How will this information be useful? Does it have potential harmful effects?
3. Does the process involve metrics and if so, how many will there be?
4. Do you need standardized measures?
5. Will the process be transparent?
6. What are the opportunity costs?
7. Who is initiating the evaluation process?
8. Who will do the developmental work?
9. Will the metrics become goals and no longer be metrics?

These are difficult questions, but their answers provide a framework for thinking reflectively and critically about the evaluation of teaching. Figure 6 presents a possible organizing paradigm.

Figure 6
The Future of Teaching Effectiveness



To implement the process in Figure 6, several things are required: first, faculty members, students, and administrators need to develop a definition or prototype for effective teaching and learning. This requires a variety of feedback mechanisms with a comprehensive framework, safe environments, and more than one assessment that enables continuous progress. Make the best possible use of technology and create a combination of recognition and accountability. This is a very tall order but one that is long overdue. Predicting and designing the future is difficult, however. Consider the protagonist Clay's response to predicting the future in the book *Mr. Penumbra's 24-Hour Bookstore* (Sloan, 2013):

> World government…no cancer…hover-boards.
> Go further. What's the good future after that?
> Spaceships. Party on Mars.
> Further.
> *Star Trek*. Transporters. You can go anywhere.
> Further…
> I pause a moment, then realize: I can't. We probably just imagine things based on what we already know, and we run out of analogies. (p. 60)

The end of the course rating form has been the standard for an exceptionally long time. We need a thoughtful national conversation about good ideas for change. Johnson (2011) tells that to do that, we need three things: first, identify the adjacent possible—the next reasonable thing we can accomplish; second, commit to a slow hunch, meaning a long-term commitment; finally, build a liquid supportive network. A good place to start the network would be the Online Learning Consortium (OLC). In summer 2023, they held a symposium on blending learning, reinvigorating that modality. Why not for evaluation of effective teaching?

Declarations

The authors declared no conflicts of interest.

The authors declared no funding sources.

Permission to collect data from human subjects was granted by the University of Central Florida, USA.

References

Abrami, P. C., & D'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of "N = 1" research: Comment of Marsh. *Journal of Educational Psychology, 83*(3), 411–415.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134–141.

Andersen, K. (2018). *Fantasyland: How America went haywire: A 500-year history*. Random House.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723–748.

Badur, B., and Mardikyan, S. (2011). Analyzing teaching performance of instructors using data mining techniques. *Inform. Educ. 10*, pp. 245–257.

Bangert, A. W. (2006). Identifying factors underlying the quality of online teaching effectiveness: An exploratory study. *Journal of Computing in Higher Education, 17*(2), 79–99.

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new jim code*. Polity.

Boghosian, B. (2020). *The inescapable casino. The Best Writing on Mathematics 2020*. Princeton University Press.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Boring, A., Ottoboni, K., & Stark P. (2016). *Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors*. LSE Impact Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2016/02/04/student-evaluations-of-teaching-gender-bias/

Boring, A., Ottoboni, K., & Stark, P. B. (2017). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, (01), 1. https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Brieman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. *Chapman and Hall/CRC.*

Brocato, B. R., Bonanno, A., & Ulbig, S. (2015). Student perceptions and instructional evaluations: A multivariate analysis of online and face-to-face classroom settings. *Education and Information Technologies 20*(1), 37–55.

Bryant, J. (2022, October 7). *High school graduates are saying no to college: Here's why*. Best Colleges. https://www.bestcolleges.com/news/analysis/why-high-school-grads-are-saying-no-to-college/

Buser, W., Batz-Barbarich, C., & Hayter, J. (2022). Evaluation of women in economics: Evidence of gender bias following behavioral role violations. *Sex Roles, Vol. 86.* p. 695–710.

Bush, M. G. (2023). *Why high school graduates are saying no to college and the consequences of their decisions.*[Unpublished manuscript]. Rosen Foundation.

Capa-Aydin, Y. (2016). Student evaluation of instruction: comparison between in-class and online methods. *Assessment & Evaluation in Higher Education, Vol. 41*(1), 112–126.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chang, T. S. (2000, April). *An application of regression models with student ratings in determining course effectiveness*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED455311)

ChatGPT. (2023, April 26). How many articles have been published on student ratings in higher education [Response to a question]. OpenAI. https://openai.com/

Chatman, J., Sharps, D., Mishra, S., Kray, L., & North, M. (2022). Agentic but not warm: Age-gender interactions and the consequences of stereotype incongruity perceptions for middle-aged professional women. *Organizational Behavior and Human Decision Processes, Vol. 173*.

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

Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education, 30*(2), 123–136.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Holt, Rinehart, and Winston Inc.

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education, 9*(4), 197–207.

Davidovitch, N., & Soen, D. (2006). Using students' assessments to improve instructors' quality of teaching. *Journal of Further and Higher Education, 30*(4), 351–376.

Dziuban, C., & Shirkey, E. (1974). When is a correlation matrix appropriate for factor analysis? *Psychological Bulletin, 81*(6), 358–361.

Dziuban, C., & Shirkey, E. (1993). A sequential psychometric criterion for the dimension of a rescaled covariance matrix. *Annual Meeting of the Florida Educational Research Association,* November 10–13, 1993.

Dziuban, C., Moskal, P., Kramer, L. & Thompson, J. (2012). Student satisfaction with online learning in the presence of ambivalence: Looking for the will-o'-the-wisp. *The Internet and Higher Education*. doi: 10.1016/j.iheduc.2012.08.001

Dziuban, C., Graham, C. R., Moskal, P. D., Norberg, A., & Sicilia, N. (2018). Blended learning: The new normal and emerging technologies. *International Journal of Educational Technology in Higher Education*, *15*(1). https://doi.org/10.1186/s41239-017-0087-5

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483–501.

Ernst, D. (2014). Expectancy theory outcomes and student evaluations of teaching, *Educational Research and Evaluation, 20(*7–8), 536–556. https://doi.org/10.1080/13803611.2014.997138

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education, 2020*(1), 1–15.

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*(3), 243–288.

Filak, V. F., & Nicolini, K. M. (2018). Differentiations in motivation and need satisfaction based on course modality: A self-determination theory perspective. *Educational Psychology*, *38*(6), 772–784. https://doi.org/10.1080/01443410.2018.1457776

Flaherty, C. (2018, May 22). Most institutions say they value teaching but how they assess it tells a different story. Inside Higher Ed. https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story

Flaherty, C. (2019, May 20). Fighting gender bias in student evaluations of teaching, and tenure's effect on instruction. *Inside Higher Ed*. https://www.insidehighered.com/news/2019/05/20/fighting-gender-bias-student-evaluations-teaching-and-tenures-effect-instruction

Flaherty, C. (2022, January 19). Study: Grade satisfaction a major factor in student evals. *Inside Higher Ed*. *https://www.insidehighered.com/news/2022/01/19/study-grade-satisfaction-major-factor-student-evals*

Floden, J. (2017). The impact of student feedback on teaching in higher education. *Assessment & Evaluation in Higher Education, 42*(7), 1054–1068.

Floridi, L. (2014). *The 4th revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Foster, M. (2023). Instructor name preference and student evaluations of instruction. *PS: Political Science & Politics, 56*(1), 143–149. doi:10.1017/S1049096522001068

Friedman, S., & Laurison, D. (2019). *The class ceiling: Why it pays to be privileged*. Policy Press.

Gelsinger, P. (2018). Mind-blowing to mundane: How tech is reshaping our expectations. *MIT Technology Review, 121*(2), 7.

Genetin, B., Chen, J., Kogan, V., & Kalish, A. (2021, December 1). *Mitigating implicit bias in student evaluations: A randomized intervention*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/epdf/10.1002/aepp.13217?saml_referrer

Ginns, P., & Ellis, R. (2007). Quality in blended learning: Exploring the relationships between online and face-to-face teaching and learning. *The Internet and Higher Education, 10*(1), 53–64.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603–615.

Giridharadas, A. (2018). *Winners take all: The elite charade of changing the world*. Knopf. Golding, C., & Adam, L. (2016). Evaluate to improve: Useful approaches to student evaluation. *Assessment & Evaluation in Higher Education*, *41*(1), 1–14, http://dx.doi.org/10.1080/02602938.2014.976810

Greenwald, A. G. (Ed.). (1997). Student ratings of professors [Current Issues]. *American Psychologist, 52*(11), 1182–1225.

Greenwald, A. G., & Gilmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Griffin, B. (2016). Perceived autonomy support, intrinsic motivation, and student ratings of instruction. *Studies in Educational Evaluation. 51*, 116–125.

Gumbel, A. (2020). *Won't lose this dream: How an upstart urban university rewrote the rules of a broken system.* The New Press.

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30*(3), 55–68.

Gunduz, N., & Fokoue, E. (2021). Understanding students' evaluations of professors using non-negative matrix factorization. *Journal of Applied Statistics, 48*(1), 2961–2981.

Heffernan, T. (2021). Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education.* https://doi.org/10.1080/02602938.2021.1888075

Hill, B. (2011). The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A monte carlo simulation. *Graduate College of the Oklahoma State University.*

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016.

Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology, 93*(4), 790–796.

Isenberg, N. (2017). *White trash: The 400-year untold history of class in America.* Penguin Books.

Johnson, S. (2011). *Where good ideas come from: The natural history of innovation.* Riverhead Books.

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled Adaptive Learning Systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100017

Kaiser, H. (1968). A measure of the average intercorrelation. *Educational and Psychological Measurement, 28,* 245–247.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kim, L. E., & MacCann, C. (2018). Instructor personality matters for student evaluations: Evidence from two subject areas at university. *British Journal of Educational Psychology, 88,* 584–605.

Kogan, V., Genetin, B., Chen, J., and Kalish, A. (2022). Students' grade satisfaction influences evaluations of teaching: Evidence from individual-level data and an experimental intervention. (EdWorkingPaper: 22-513). *Annenberg Institute at Brown University.* https://doi.org/10.26300/spsf-tc23

Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education, 24*(1), 27–42.

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology, 7*, 570. https://doi.org/10.3389/fpsyg.2016.00570

Lannutti, P. J., & Strauman, E. C. (2006). Classroom communication: The influence of instructor self-disclosure on student evaluations. *Communication Quarterly, 54*(1), 89–99.

Lowenthal, P., Bauer, C., & Chen, K. (2015). Student perceptions of online learning: An analysis of online course evaluations. *American Journal of Distance Education, 29*(2), 85–97.

Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research, 26*(5), 755–769.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Mather, M., & Sarkans, A. (2018). Student perceptions of online and face-to-face learning. *International Journal of Curriculum and Instruction*, *10*(2), 61–76.

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*(3), 176–179.

McGhee, H. (2021). *The sum of us: What racism costs everyone and how we can prosper together.* One World.

Mengel, F., Sauermann, J., & Zolitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535–566.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics, 51*(3), 648–652.

Mlodinow, L. (2009). *The drunkard's walk: How randomness rules our lives*. Penguin Books.

Mukherjee, S. (2016). *The gene: An intimate history.* Scribner.

Muller, J. (2018). *The tyranny of metrics*. Princeton University Press.

Nasser, F., & Fresko, B. (2006). Predicting student ratings: The relationship between actual student ratings and instructors' predictions. *Assessment & Evaluation in Higher Education, 31*(1), 1–18.

Oliver, R. L., & Sautter, E. P. (2005). Using course management systems to enhance the value of student evaluations of teaching. *Journal of Education for Business, 80*(4), 231–234.

O'Neil, C. (2018). *Weapons of math destruction how big data increases inequality and threatens democracy*. Penguin Books.

Page, S., (2009). *Understanding complexity*. The Great Courses.

Pirsig, R. (2006). *Zen and the art of motorcycle maintenance: An inquiry into values*. William Morrow Paperbacks.

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.

Ray, B., Babb, J., & Wooten, C. A. (2018). Rethinking SETs: Retuning student evaluations of teaching for student agency. *Composition Studies*, *46*(1), 34–194.

Read, W. J., Rama, D. V., & Raghunandan, K. (2001). The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business, 76*(4), 189–192.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46*(8), 929–953.

Richardson, J. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 135–147.

Rinderman, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education, 42*(4), 377–399.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology, 133*(1), 19–35.

Rosch, Eleanor (Eleanor Heider). 1973. Natural Categories. *Cognitive Psychology, 4*, 328–350.

Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education, 4*(1). https://doi.org/10.5430/ijhe.v4n1p217

Safir, S., & Dugan, J. (2021). *Street data: A next-generation model for equity, pedagogy, and school transformation*. Sage Publications Inc.

Scherer, R., & Gustafsson, J. E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: an application of multilevel bifactor structural equation modeling. *Frontiers in Psychology, 6.* http://dx.doi.org/10.3389/fpsyg.2015.01550

Searle, J. (1995). *The construction of social reality*. The Free Press.

Sellnow-Richmond, D., Strawser, M. G., & Sellnow, D. D. (2020). Student perceptions of teaching effectiveness and learning achievement: A comparative examination of online and hybrid course delivery format. *Communication Teacher, 34*(3), 248–263.

Sheehan, E. P., & DuPrey, T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology, 26*(3), 188–194.

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.

Sloan, R. (2013). *Mr. Penumbra's 24-hour bookstore*. Picador Paper.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino professors. *Journal of Latinos and Education, 4*(2), 115–136.

Snygg, D., & Combs, A. W. (1949). *Individual behavior: A new frame of reference for psychology.* Harper.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*(3), 269–291.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stott, P. (2016). The perils of a lack of student engagement: Reflections of a "lonely, brave, and rather exposed" online instructor. *British Journal of Educational Technology, 47*(1), 51–64.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science, 11*(6), 800–816. https://doi.org/10.1177/1745691616650284

Taleb, N. N. (2005). *Fooled by randomness: The hidden role of chance in life and in the markets.* Random House.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. Random House.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder.* Random House.

Taleb, N. N. (2018). *Skin in the game: Hidden asymmetries in daily life.* Random House.

Tejeiro, R., Whitelock-Wainwright, A., Perez, A., Urbina-Garcia, M. A. (2018). The best-achieving online students are overrepresented in course ratings. *European Journal of Open Education and E-learning Studies, 3*(2), 43–58.

Turner, K. M., Hatton, D., & Theresa, M. (2018). Student evaluations of teachers and courses: Time to wake up and shake up. *Nursing Education Perspectives*, *39*(3), 130–131.

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ*. http://dx.doi.org/10.7717/peerj.3299

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief overview. *Assessment & Evaluation in Higher Education, 23*(2), 191–212.

Wang, M. C., Dziuban, C. D., Cook, I. J., Moskal, P. D. (2009). Dr. Fox Rocks: Using Data-mining Techniques to Examine Student Ratings of Instruction. In M. C. Shelley II, L. D. Yore, & B. Hand (Eds.), *Quality research in literacy and science education: International perspectives and gold standards*. Springer.

Wilkerson, I. (2020). *Caste: The origins of our discontents.* Random House.

Williams, R. G., & Ware, J. E., Jr. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer. *American Educational Research Journal, 14*(4), 449–457.

Yen, S.-C., Lo, Y., Lee, A., & Enriquez, J. M. (2018). Learning online, offline, and in-between: Comparing student academic outcomes and course satisfaction in face-to-face, online, and blended teaching modalities. *Education and Information Technologies*, *23*(5), 2141–2153. https://doi.org/10.1007/s10639-018-9707-5

Appendix A

Student Perception of Instruction

Instructions: Please answer each question based on your current class experience. You can provide additional information where indicated.

All responses are anonymous. Responses to these questions are important to help improve the course and how it is taught. Results may be used in personnel decisions. The results will be shared with the instructor after the semester is over.

Please rate the instructor's effectiveness in the following areas:

1. Organizing the course:
a) Excellent b) Very Good c) Good d) Fair e) Poor

2. Explaining course requirements, grading criteria, and expectations:
a) Excellent b) Very Good c) Good d) Fair e) Poor

3. Communicating ideas and/or information:
a) Excellent b) Very Good c) Good d) Fair e) Poor

4. Showing respect and concern for students:
a) Excellent b) Very Good c) Good d) Fair e) Poor

5. Stimulating interest in the course:
a) Excellent b) Very Good c) Good d) Fair e) Poor

6. Creating an environment that helps students learn:
a) Excellent b) Very Good c) Good d) Fair e) Poor

7. Giving useful feedback on course performance:
a) Excellent b) Very Good c) Good d) Fair e) Poor

8. Helping students achieve course objectives:
a) Excellent b) Very Good c) Good d) Fair e) Poor

9. Overall, the effectiveness of the instructor in this course was:
a) Excellent b) Very Good c) Good d) Fair e) Poor

10.What did you like best about the course and/or how the instructor taught it?

11. What suggestions do you have for improving the course and/or how the instructor taught it?

Appendix B

UCF Modality Codes

Code	Modality
P	Face-to-face
M	Mixed mode/blended with reduced F2F
RA	Blended with active learning
RS	Blended, with no more than 20% F2F
RV	Lecture capture with live option
V	Video streamed
R	Lecture capture
V1	Emergency remote instruction
WW	Fully online