PRE-RATE MY PROFESSOR: PREDICTING COURSE RATINGS AND RESPONSE RATES FROM LMS ACTIVITY IN COLLEGE COURSES

Richard Scruggs

A DISSERTATION

in

Education

Presented to the Faculty of the University of Pennsylvania

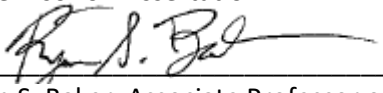
in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation



Ryan S. Baker, Associate Professor of Education

Graduate Group Chairperson

J. Matthew Hartley, Professor of Education

Dissertation Committee:

Manuel González Canché, Associate Professor of Education

Shane Dawson, Professor of Learning Analytics, University of South Australia

ACKNOWLEDGEMENT

Even when most of the work of a dissertation is conducted alone in a room, staring at code and data on computer screens, no dissertation could be completed without the help of many other people.

First, thank you to Stefan, for starting this whole thing off by suggesting I ask Ryan to become my advisor. Before then, I hadn't realized it was even a possibility, but it ultimately led to the Penn Center for Learning Analytics and working on a lot of great projects with wonderful people.

Also at the PCLA, thank you to Jaclyn for your suggestions at the initial proposal and general advice since. Thanks to all the other PCLA students – it's been such a treat working with and learning from all of you. I've really valued the atmosphere of collaboration and intellectual curiosity that we've all shared.

Thank you to Manuel, Shane, and Ryan for all your help and advice as committee members. In particular, thank you to Shane for access to the data, and also thank you, Srecko, for all your help with actually getting and explaining the data logs.

From the quantitative methods side of my background, thank you to Jess, Michael, Roland, and everyone else in the stat lab for your help with factor analyses. I really missed being able to bounce ideas off you in the lab over the last year and a half. Also, thank you to Paul for your suggestions and ideas on how to structure a factor analysis when the data is multiply nested.

Ryan, I could never thank you enough for being a fantastic advisor. Your thoughtful guidance on so many aspects of this dissertation – and many other projects –

has been truly invaluable. In my time working with you, I've learned so much about how to be a good scientist, how to ask good questions, and how to be a successful academic researcher. I cannot count the number of times I've told my wife, my family, and my friends how glad I am that you're my advisor.

My parents may not have understood all of my research, but they've nonetheless supported me my entire life. As long as you're still interested in hearing about my work, I promise I'll do my best to keep explaining it to you.

Finally, my deepest love and gratitude to my wife, Daniella. This dissertation never would have been finished without you. Of course, I'd never have finished my master's paper without you either.

ABSTRACT

PRE-RATE MY PROFESSOR: PREDICTING COURSE RATINGS AND RESPONSE RATES FROM LMS ACTIVITY IN COLLEGE COURSES

Richard Scruggs

Ryan S. Baker

College teaching is primarily assessed through the use of course ratings, which are expected to act as both summative and formative feedback. Considering the significance of teaching in academia and the amount of time professors spend on teaching and related activities, it is particularly important that ratings are effective formative feedback. In this study, methods from learning analytics and data mining are used in an effort to predict course ratings and response rates on ratings surveys from students' activity in the course learning management system, with the goal of making predicted ratings available to faculty early.

Regression and classification methods used in this study included linear and logistic regression, random forests, and gradient boosting and features were chosen based on their inclusion in earlier studies predicting individual student success and motivation. However, none of the models created for this study were not able to accurately predict either course ratings or response rates on either the entire data sample or a subset of classes with higher LMS activity. This may have been caused by difficulties with aggregating features or outcome variables to the class level, which was necessary due to the confidentiality of the student ratings. It may also result from the complexity of good

college teaching: unlike individual student grades or motivation, which have been successfully predicted, there are many successful and unsuccessful forms of teaching.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: INTRODUCTION

College teaching, despite occupying the majority of professors' time (Bentley & Kyvik, 2012), is an activity for which professors receive little training and inadequate feedback and assessment. Professors are overwhelmingly hired based on their research, with their teaching skills or training receiving relatively little attention (Ishiyama et al., 2014; Norton et al., 2013). Robinson and Hope (2013), surveying Florida faculty, found that 78% entered the classroom with no training in pedagogy at all.

Once in the classroom, professors' teaching performance is primarily assessed through student ratings,[1] which are often misinterpreted or analyzed with unsuitable statistical tests (Boysen, 2015; Boysen et al., 2014; Kitto et al., 2019; Miller & Seldin, 2014). In addition to their use as evaluations, ratings are also intended as formative feedback and are far more prevalent than other types of formative feedback for instructors in higher education (Ronald A. Berk, 2005; Marsh, 2007; Miller & Seldin, 2014). Given this situation, it is not surprising that the use of course ratings rarely seems to improve teaching and learning (Hammonds et al., 2017; Kember et al., 2002; cf. Centra, 2015; Marsh, 2007). Considering that teaching is such an important and time-consuming part of professors' jobs, it seems self-evident that they should receive better feedback in order to help improve their teaching, and thus improve student learning.

[1] The terms "course ratings" and "student ratings" are used interchangeably; the literature review discusses the terms at more length.

In this study, the intent was to help address the dearth of instructor feedback by offering a new form of feedback through analysis of student learning management system [LMS] data. The primary focus of this project was twofold: reliably detecting relationships between LMS interaction patterns and course ratings, and determining how much interaction data is necessary to produce reliable and valid detectors. Since low response rates are a common challenge to the use of course ratings, relationships between LMS interaction patterns and students' response rates on ratings surveys are also explored.

This analysis was based on the assumption that an instructor's actions, materials, and other aspects of their curriculum would lead to class-level differences in student interaction with an LMS. This study aimed to use those differences to build detectors of student response on course rating surveys. Early predictions of student ratings would be able to inform teachers, at a basic level, about how their classes are functioning.

If LMS data could be shown to be related to course ratings, feedback from such data would provide several advantages over the traditional end-of-course survey format. Most significantly, LMS data can easily be collected from all students in a course, not only those who respond to a survey. In addition, LMS data is available much earlier in the semester, meaning that instructors could receive feedback with more time to improve their courses before the end of the semester, helping the rating process become more formative.

Although no published work has explored the relationship between LMS data and course ratings, LMS data has been used to predict individual student success or dropout

since at least 2007 (Arnold & Pistilli, 2012). More recently, Babić (2017) used such data to predict students' scores on a psychometrically validated scale of academic motivation (Vallerand et al., 1992). Since motivation is among the student-level factors that have been shown to relate to ratings and rating response rates (Hoel & Dahl, 2019), this suggests that LMS interaction patterns may be affected by similar intrinsic characteristics as those which affect course ratings.

Even if there are no detectable relationships between course ratings and LMS interactions, identifying courses with low response likelihood would still be valuable. If department chairs and administrators knew that students in certain course sections were less likely to fill out course rating forms, they could either act to increase the response rate – add incentives to the surveys or emphasize the importance of the process to students in those sections – or add alternative methods of course assessment – peer ratings, self-evaluations, et cetera. These actions, while possibly too resource-intensive to apply to all courses, could address professors' criticisms that their ratings are based on the responses of only a few students.

CHAPTER 2: LITERATURE REVIEW

In order for predictions of a measure to be useful, the underlying measure which was being predicted should be sufficiently trustworthy that predictions, being less accurate than the original measure, are more of a help than a hinderance in the system in which they are eventually used.[2] This section will establish that course ratings are reliable and valid instruments of teaching quality, although ratings data are often misinterpreted or misused in practice, leading to poor formative utility. In addition, prior research in learning analytics and educational data mining that relates to course ratings is discussed.

Student ratings of courses have become one of the facts of life of modern higher education. Most ratings consist of students filling out standardized forms, rating courses on various criteria on a five-point scale – "The instructor helped me achieved my goals," "The course was organized in a way that helped me learn," "How would you rate the overall effectiveness of this course?" often with an open-ended question or two about the effectiveness of the course at the end (Sheehan & DuPrey, 1999; examples taken from https://teaching.berkeley.edu/course-evaluations-question-bank). Regardless of the precise form, however, course evaluations are used at nearly all higher education institutions in the United States (Hmieleski & Champagne, 2000; Miller & Seldin, 2014; Seldin, 1999).

[2] There is no precise cutoff for this – some predictions will have greater effects than others and will consequently need to be more accurate – but the future uses of predictions and the impact of both positive and negative predictions should be kept in mind.

Although course rating forms are used at nearly all institutions, there is little standardization in the precise format of rating forms. Therefore, it is difficult to compare findings across contexts with certainty (see Barre, 2015). Some forms, such as ETS' Student Instructional Report II (Centra, 1998) have been psychometrically validated; many have not. Studies of course rating data are often methodologically questionable, conflating arguments relating to individual rating instruments with arguments relating to the wider process of course ratings (L'Hommedieu et al., 1990).

Terminology

There have been many terms used to refer to the mostly-quantitative survey instruments that attempt to measure teaching and student experience in college classes. Spooren, Brockx, and Mortelmans (2013) conduct a literature review using the following terms: "SET, student evaluation of teaching, student ratings, student ratings of instruction, teacher evaluation, teaching effectiveness, teaching performance, higher education, and student evaluations" (p. 601). In addition to these terms, many researchers use the phrase "course evaluations," "student course evaluations," "course ratings," "course experience questionnaire," or the even farther-reaching "student evaluation of faculty" (see, e.g., Bendig, 1953; Degheri, 2017; Granzin & Painter, 1973; Gravestock & Gregor-Greenleaf, 2008; Haskell, 1998; Sauer, 2012; Talukdar et al., 2013). In this section, the terms "student ratings" and "course ratings" will be used indistinguishably.

It is also important to draw a distinction between *feedback* and *evaluation* and situate course ratings in relation to those two terms. Feedback, in education, generally relates to more formative information that is intended to affect future practice (Boud &

Molloy, 2013), while evaluation is typically more summative, used for performance assessments (C. J. Harrison et al., 2015). This study will focus on the formative purposes of course ratings – which is to say, their use as instructor feedback.

Nearly all of the above terms describe student ratings as summative instruments, but the literature also refers to ratings as feedback, with relatively little discussion of the differences (e.g., Denson et al., 2010; Kember et al., 2002; cf. Donovan et al., 2010; Flodén, 2017 who draw the distinction more explicitly). In the next section, this issue is discussed at more length, as the summative and formative purposes of course ratings affect their validity.

Validity of Course Ratings

Defining validity can be contentious, but I will follow the definition from *The Standards for Educational and Psychological Testing*: "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association et al., 2014). Thus, any discussion of the validity of course rating scores hinges on the purposes of those ratings. Marsh, a long-time scholar of ratings, offers the following purposes for gathering student ratings (2007; slightly modified from Marsh, 1987):

- "diagnostic feedback to faculty for improving teaching;
- a measure of teaching effectiveness for personnel decisions;
- information for students for the selection of courses and instructors;
- one component in national and international quality assurance exercises, designed to monitor the quality of teaching and learning; and

- an outcome or a process description for research on teaching (e.g., studies designed to improve teaching effectiveness and student outcomes, effects associated with different styles of teaching, perspectives of former students).” (p. 320)

Sheehan and Duprey (1999), writing from the more uncommon but likely more influential perspective of administrators, list only two purposes of ratings: providing data for performance appraisals and improving teaching. These two purposes are those most discussed in the literature; therefore, I will focus on them in this section and attempt to show that ratings can be used for each of these purposes.

In recent years, the quantitative nature of course ratings has lent them appeal as an easy metric of teaching quality, which has likely contributed to their ubiquity. Nearly all institutions that use course ratings use them to aid in making hiring, promotion, and tenure decisions (Emery et al., 2003; Haskell, 1998; Hornstein, 2017; Linse, 2017; Young, 1993), although some critics still argue that they should not be used (Hornstein, 2017; Stroebe, 2020), or should be used much more cautiously (Jones et al., 2014).

In addition, while ratings have been commonly used in higher education for decades, their formative utility is still debated, with some authors (e.g., Centra, 1998) stating that the use of ratings improves teaching effectiveness and critics (e.g., Emery et al., 2003) stating otherwise, with Stroebe (2020) going so far as to claim that ratings encourage *poor* teaching by rewarding lenient instructors and easy classes. Kember, Leung, and Kwan (2002) studied the use of ratings at university in Hong Kong for several

years, concluding that the instrument had no effect on teaching quality, as measured by the rating instrument.

Research suggests that information from ratings is more effective at improving teaching if ratings are collected midsemester, rather than at the end, and if ratings are used as part of comprehensive efforts to improve teaching (Centra, 2015; Marsh, 2007). Marsh (2007) cites a number of studies showing that ratings, particularly when used as part of an improvement program, led to higher student satisfaction and better student learning (see also Piccinin et al., 1999). Hampton and Reiser (2004), performing a multisection study with some teaching assistants receiving midterm course ratings and consultations, found that TAs who received the ratings used different practices, but there were no significant differences in student learning as measured by final exam scores.

Unfortunately, ratings are often used merely as alarm systems, with administrators only investigating courses with very low ratings (Edström, 2008; Andersen, 2018 also describes this practice at one university). The reduction of Likert-type data to means and the focus on one "overall" item may also contribute to the problem – instead of administrators and faculty seeing the actual distribution of rating scores, they may be reacting to the isolated complaints of only a few students.

Students have their own views about the purposes of ratings. Chen and Hoshower (2003) found that students filling out ratings want the ratings to be used to improve the course and the instructor's teaching; fewer students saw personnel decisions as important. The importance of ratings as formative feedback was also noted by Brown (2008) and Abbott, Wulff, Nyquist, Ropp, and Hess (1990), who found that students strongly

preferred midsemester evaluations to the traditional end-of-course ratings. Finally, students are not certain how ratings are actually used, with only 30% in one survey believing that the ratings affect tenure decisions (Kite et al., 2015; see also Ernst, 2014). Shah et al. (2017) discuss how students may become frustrated with the ratings process as they see few changes despite repeatedly filling out surveys.

A great deal of research has explored the relationship between students' ratings of courses and student learning. Meta-analyses have generally supported the idea that ratings have a significant, positive relationship with student learning (Clayson, 2009; Cohen, 1981; Spooren et al., 2013; Wright & Jenkins-Guarnieri, 2012). However, Uttl, White, and Gonzalez (2017) strongly criticize the methodology of these older studies and perform their own analysis, which, while relying heavily on a large dataset of economics courses produced by Weinberg, Hashimoto, and Fleisher (2009), suggests that there is no relationship between student ratings and student learning. In a related study, Braun and Leidner (2009) examine the relationship between course ratings and students' scores on a self-assessment of competence, finding strong correlations but distinct constructs. Still more recently, Carpenter et al. (2020) note that students don't always understand their own learning processes, pointing out that subjects in a cognitive psychology study (Kirk-Johnson et al., 2019) poorly rate learning practices which require more mental effort, but were more effective than other, less effortful practices.

Linse (2017) argues that the entire issue of student learning is misleading, claiming that "student ratings have never been intended to serve as a proxy for learning." Indeed, although there have been numerous studies on the relationship between ratings

and learning (e.g., the meta-analyses of Clayson, 2009; Cohen, 1981; Uttl et al., 2017), researchers of student ratings tend to focus more on whether the ratings reflect teaching effectiveness (Marsh, 2007; McKeachie, 1997; Wachtel, 1998). Teaching effectiveness, as a term, is rarely well-defined but should generally lead to learning, if not always higher grades (Braga et al., 2014; Carrell & West, 2010; Linse, 2017; Pascarella et al., 2008).

Apart from the question of student learning, researchers have found that scores on course rating instruments correlate significantly with many other methods of evaluating teaching. Feldman (1989) synthesized studies comparing teacher self-ratings, colleague ratings, administrator ratings, trained observer ratings, and ratings from current and former students, finding that current student ratings highly correlated with former student ratings, as well as with colleague ratings and administrator ratings. More recent comparisons like Feldman's are scarce, although Roche and Marche (2000) find moderate relationships between student ratings and instructor self-concept. More research exists outside of higher education (e.g., van der Lans, 2018, who looks at secondary education).

Factor Analyses of Student Ratings

Factor analysis is a way to examine an instrument with many items by finding which items ask about the same underlying construct. Performing a factor analysis is one way to support the hypothesis that an instrument has construct validity (Thompson & Daniel, 1996). Many researchers have performed factor analyses on student ratings data. These analyses should, of course, all be considered with the caveat that the researchers

were not all working from the same course rating forms. In addition, many early analyses were conducted using principal components analysis (e.g., Granzin & Painter, 1973; McGhee & Lowell, 2003; Tetenbaum, 1977), which produces significantly inflated factor loadings in cases with low factor loadings or few variables [items] (Snook & Gorsuch, 1989). As few ratings instruments consist of many items, this is particularly worrisome; thus, factor analyses of student rating instruments should generally be conducted using common factor analysis, also known as principal axis factoring.

Few course rating instruments (cf. Centra, 1998; Marsh, 1982) were designed with an eye toward psychometric properties, but factor analyses of rating instruments generally have shown reasonable factor structures, suggesting that rating instruments are providing information about courses and teaching (Capa-Aydin, 2016; Marsh & Bailey, 1993; Marsh & Hocevar, 1991a; Zhao & Gallant, 2012).

Different factor structures may lead to different formative utility. If an instrument only has one factor, an instructor who scores low on that factor has very little direction as to how to improve their course. On the other hand, an instructor who scores low on Organization/Clarity knows that their students consider the organization of their course to be a problem. The instructor still does not have specific recommendations – although the students may offer some in response to the open-ended questions – but they have a place to start.

Reliability of Course Ratings

In order for a measure or assessment to be useful, its results should be both valid and reliable. Benton and Cashin (2014) discuss three aspects of reliability as it relates to

student ratings: consistency, stability, and generalizability. Consistency is the standard definition of reliability – whether observers' ratings agree, or whether items on a scale measure the same construct (see e.g., Shadish et al., 2002). Stability refers to whether ratings agree over time, and generalizability is the applicability of ratings to an instructor's other courses.

Marsh and Roche (1997) state that the most important measure of student ratings' reliability is interrater agreement. They find that, while two individual students' ratings have little agreement, the overall interrater agreement is generally high (0.90 for a class of 20 students; they do not specify the length of the instrument). However, that agreement depends heavily on the number of students in a class; classes with more students will almost invariably have high interrater agreement (Cashin, 1995; Marsh & Roche, 1997; see also Tomes et al., 2019, who found that peer prediction was stable with fewer respondents than standard Likert scales).

Stability is also an important characteristic of reliability. Course ratings can be stable from an individual perspective, meaning that the same instructor receives the same ratings across different classes and different times, or they can be stable from a test-retest perspective, meaning that the same students gave the same class the same ratings at different times. Benton and Cashin (2014) argue that ratings of individual instructors are stable over time, citing Braskamp and Ory (1994) and Marsh and Hocevar (1991b). Hativa (1996) also concludes that ratings are stable, even across activities intended to improve teaching – which is to say the activities had no discernible effect on ratings. Carle (2009) conducted a multilevel growth modeling analysis, concluding that ratings

were generally stable. Drysdale (2010) conducted a test-retest analysis of ratings using one instrument, finding that the ratings were not stable over three-week retest periods.

In summary, quantitative end-of-course rating instruments can produce reliable and valid summative measurements of instructors' teaching. However, they generally do not appear to provide formative information that helps to improve teaching. They may be able to call attention to the worst teachers, but instructional consultants are better at providing formative feedback (Hampton & Reiser, 2004; Piccinin et al., 1999). In addition, collecting ratings at the end of courses limits their formative utility. Unfortunately, although course ratings can produce good measurements, they are often analyzed using inappropriate methods, leading to flawed or incorrect conclusions, as discussed later.

Sources of Bias

A great deal of criticism surrounding course ratings focuses on the idea that they are biased against some instructors. While bias is a term with an intuitive definition – unfair prejudice in favor or against something – scholars of course ratings add one significant element. Centra (2003) offers the following definition, also adopted by Marsh (2007) and Benton and Cashin (2014): "Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning" (p. 498). Centra (2003) argues, for example, that class size, which is known to influence course ratings, is not a source of bias as class size has also been shown to impact student learning.

For decades, instructors have claimed that course ratings are biased, often citing studies to support their points of view. However, more recent reviews conclude that most sources of bias are fairly small and controllable (Benton & Ryalls, 2016; Macfadyen et al., 2016; Sauer, 2012; Spooren et al., 2013), although the debate continues. Darwin (2021) cites studies showing bias on instructor race and gender, with students rating female instructors and non-white instructors lower. Fischer & Hänze (2019 examined at relationships between students' first impressions of instructors – as a potentially biasing measure – and students' eventual rapport with instructors, measured by observers and through self-report, finding significant relationships. This finding calls back to Centra's (2003) theories about the nature of bias in course ratings. Finally, Park & Dooris (2019) use decision trees to predict overall rating scores from other factors, but the algorithm they use does not split on either gender or race, leading them to conclude that their instrument does not reflect gender or race bias. Ultimately, the question of bias is likely to be argued for some time, but it should be reasonable to prefer the findings of reviews like those cited earlier to individual studies.

Supporters of ratings suggest simply adjusting course ratings for known sources of bias (Benton & Cashin, 2014; Macfadyen et al., 2016). If so doing, it may be advantageous to adjust ratings for all factors that affect ratings but are out of the control of the instructor – otherwise, ratings of instructors who are assigned to teach smaller classes will be incomparable to ratings of instructors who teach larger classes.

Sauer (2012) and Benton and Cashin (2014) review many potential sources of bias, concluding that while some factors (e.g., academic discipline and class size) are

related to course evaluation ratings, many other factors are not (e.g., instructor gender, student age, student GPA, time of course). Instructor personality has been shown to influence ratings; Benton and Cashin (2014) argue that personality relates to actual teaching effectiveness and therefore should not be controlled for.

Student grades are a particularly contentious issue here, with some scholars stating that students with lower grades tend to rate their instructors harshly (e.g., Matos-Díaz, 2012); Love and Kotchen (2010) and Stroebe (2020) theorize that the use of course ratings may contribute to grade inflation. Most recent work, however, tends to find that grades do not have a great impact on ratings (Benton & Ryalls, 2016; Centra, 2003). As will be discussed later, nonresponse biases are likely a greater concern in this study, particularly given that research suggests that students with higher grades are more likely to respond to course rating surveys (Adams & Umbach, 2012; Benton & Ryalls, 2016).

Other Challenges of Course Ratings

Course rating instruments face similar challenges to other survey instruments, namely low response rates and survey fatigue. Hoel and Dahl (2019) cite studies showing response rates between 30% and 70% on course rating surveys. Hoel and Dahl also found significant differences between students who usually respond to rating surveys and those who rarely or never do, with the frequent responders scoring higher on measures of autonomy and self-determination. These differences lend some credence to criticisms that ratings come from small, non-representative groups of students (Goos & Salomons, 2017; Nulty, 2008). This is a particular challenge to this study and will be discussed later.

Survey fatigue, as it affects course ratings, is somewhat more acute than the general case of respondents receiving too many solicitations for surveys. Although students are typically asked to fill out a course rating form for each course they take, they rarely see any results or improvement from the surveys (Y. Chen & Hoshower, 2003; Kite et al., 2015; Shah et al., 2017; Spencer & Schmelkin, 2002). Naturally, this harms students' motivation to participate in the process, leading to fewer, less-helpful responses (Y. Chen & Hoshower, 2003).

Another major problem with current course rating practice is incorrect analysis methods. Course rating instruments generally use 5- or 7-point Likert-type ordinal items. There are numerous disagreements in the literature as to whether data from such items or scales consisting of such items – note the distinction between individual items and scales comprised of groups of items – can be analyzed using statistical tests developed for interval data, or whether other statistical tests must be employed (Carifio & Perla, 2007; Harwell & Gatti, 2001; Mircioiu & Atkinson, 2017). However, recent research has used simulated data to demonstrate that data from Likert-type scales is usually suitable for analysis as interval data (Carifio & Perla, 2007; Mircioiu & Atkinson, 2017).

Many studies – and in practice, many uses – of course ratings do not analyze results as a scale, instead focusing solely on one "overall score"-type item from the instrument. Depending on the methods used, this can lead to erroneous conclusions. Clason and Dormody (1994) strongly criticize reducing the data from a Likert-type item to a single mean, or relying on statistical procedures that require the assumption of normality to analyze data from these items. Even authors such as Carifio and Perla (2007)

who support analyzing data from Likert scales as interval data argue that analyzing single items "should only occur very rarely" (p. 1151).

Considering that data from course ratings tends to be negatively skewed, with more high ratings than low, using mean values tends to overweight the impact of outlying low ratings. The use of weighted means can help with this problem. However, Harrison, Douglas, and Burdsal (2004), using data from one university, compared unweighted means from one "overall-rating" item with a variety of other measures, finding very little difference. They used data from a rating instrument with six first-order factors and two second-order factors: one, with data from four of the first-order factors, which was used as an overall evaluation of teaching quality (P. D. Harrison et al., 2004). They found that the mean of students' overall ratings correlated very highly (>0.8) with several weighted averages of first-order factor scores as well as the pertinent second-order factor which they considered the best overall measure (P. D. Harrison et al., 2004). Despite this evidence, I argue that using mean scores from single Likert-type items is overly reductive – even if scores correlate, there's no reason to use a single item when there are enough items to build a factor – and doing so contributes to the poor formative utility of ratings.

Course Ratings and Effective Teaching in Australia

As the data for this study comes from a large Australian university, it is important to discuss how course ratings are used and viewed in Australia, particularly as compared to the American context. Australia has taken more steps than the United States in formalizing student ratings as a metric of university performance, although the ratings that they use differ from the course ratings discussed here. The Course Experience

Questionnaire[3] (CEQ, now the Graduate Outcomes Survey) asks about recent graduates' experiences in their former university and program and has been nationally administered since 1993 (Barrie et al., 2008; QILT, 2019). The Student Experience Survey (SES), which asks all current students about their learning experiences, is conducted annually at every Australian university. Results from the SES are heavily used by university administrators to assess performance and the national government has used and plans to return to using scores on the survey to direct funds to universities and departments (Barrie et al., 2008; Minister for Education, The Hon Dan Tehan MP, 2019). Unlike course rating surveys, the SES focuses on students' larger educational experience within their program or university, an approach which has been criticized by some researchers. Marsh, Ginns, Morin, Nagengast, and Martin (2011) perform a multilevel structural equation model analysis, finding that the CEQ does not discriminate well between universities or departments. Marsh et al. (2011) warn against the common practice of using research on student ratings to argue for the validity of broader university experience surveys (e.g., Hirschberg & Lye, 2016; Talukdar et al., 2013), adding that while existing course rating surveys can discriminate between instructors, they are similarly ineffective at discriminating between departments and universities.

In recent years, both student ratings of individual classes and student surveys of their larger experiences in their programs of study have been used in efforts to improve university teaching in Australia (Talukdar et al., 2013). While student ratings of classes

[3] Note that the term 'course' here does not follow the American usage: rather than being a single class, 'course' refers to a student's entire course of study in their degree program.

are not nationally mandated, the culture of assessment fostered by the SES and CEQ means that scores on class rating surveys are very important to departments and universities (S. Dawson, personal communication, November 21, 2019). This culture of assessment has led to more research on the actual impacts of course ratings in Australia (e.g., Darwin, 2017) and on how the rating process might be improved (e.g., Darwin, 2021; Shah et al., 2017), although there are still concerns about inadequate formative utility and ignoring or downplaying the student voice in favor of the quantitative side of the surveys (Darwin, 2021; see also Golding & Adam, 2016, with similar views from the context of New Zealand). As this study focuses more on formative assessment, program- or university-level surveys like the SES and CEQ will not be discussed, keeping instead to the individual class level student ratings.

Ideal Use

Both proponents and detractors of course ratings argue that ratings are useful, but should not be used as the sole yardstick of instructional – or worse, faculty – quality (Benton & Cashin, 2014; Emery et al., 2003; Hativa, 2000; Marsh, 2007; Xu, 2012). Researchers are divided on the best statistical methods to analyze rating data. Kitto et al. (2019) offer a hierarchical Bayesian model that predicts the proportion of scale responses, but admit that it is not easily understood by non-statisticians; others support the use of scale scores (Linse, 2017) or even mean item scores (P. D. Harrison et al., 2004). Many authors favor the use of teaching portfolios, which would contain rating scores in addition to other materials (Berk, 2018; Braskamp & Ory, 1994; Centra, 1993; Paulsen, 2002; Seldin, 1999).

Most course rating researchers, however, believe that ratings can serve formative purposes as well as summative, with instructors gaining information about the student experience which they can then use to improve their courses. Unfortunately, partly due to the quantitative focus of most rating instruments, this idea rarely comes to fruition in practice – Sheehan and DuPrey (1999) neatly summarize the disconnect: "It is assumed that faculty will heed the information provided in their evaluations, making appropriate changes in either or both their instructional style and the content of their lectures" (p. 189). How faculty are meant to identify appropriate changes, of course, is rarely explained. Researchers have found that faculty rarely know how to react to student responses to open-ended questions (Edström, 2008; Lutovac et al., 2017), so it seems unlikely that they would know how they should modify their courses based on quantitative feedback. Andersen (2018) describes the situation at one Scandinavian university where faculty look at their qualitative feedback but mostly ignore the quantitative.

The concept of feedback also deserves some attention here. Although the literature on feedback in education focuses on feedback as information given to students, some scholars have studied feedback more generally, showing that it can improve teaching. Shute (2008), in an excellent review of research on formative feedback, presents many guidelines that have strong evidence of efficacy, including recommendations to keep feedback focused on the task, present feedback in manageable units, and give feedback soon after the task if the task is new or difficult. Brinko (1993), specifically writing about the role of feedback in improving teaching, also emphasizes

that feedback should be given soon after the performance. Studies on course ratings have echoed this as well, suggesting that giving ratings to instructors earlier in the semester can lead to higher student satisfaction and better student learning (Centra, 2015). Given that some studies have found that ratings given earlier in the semester strongly correlate with ratings given at the end (Clayson, 2013), it might be better to collect course ratings in the middle of the semester.

Blash et al. (2018), Taylor et al. (2020), and Veeck et al. (2016) describe midterm course rating processes which, while being more structured than the simple written survey, are more formative and provide more actionable information to instructors. However, all of these processes involve more class time spent gathering student feedback. In addition, the strategies in Blash et al. (2018) and Taylor et al. (2020) require external facilitators – other instructors or educational developers – coming into the classroom to gather students' feedback. Implementing midterm ratings as described by these authors would necessitate a greater investment of time and effort into the rating process, which may not be practical at many institutions.

Although there are fewer studies on simply moving ratings earlier in the term without changing their format, it is likely that this would also increase the formative utility of the ratings. It would improve adherence to Brinko's (1993) recommendation on giving feedback close to the performance. Instructors would also know that improvements were attributable to their changes, rather than a new group of students who may have been more receptive to their teaching style. In addition, students would have the chance to see their participation making an impact on the class, helping response rates

and motivation to participate in the rating process. It would not be a magic bullet – instructors would still have to work with making sense of students' comments or quantitative ratings – but it has the potential to significantly improve the rating process.

This dissertation attempts to offer an alternative to moving the ratings earlier in the semester by using methods from educational data mining [EDM] and learning analytics [LA] to predict the ratings, with the goal of offering the resulting predictions to instructors. Early feedback generated in this way would have significant advantages over existing midterm ratings: predictions could be generated more often, allowing feedback to be given almost immediately after the performance; predictions would use data from all students, rather than only the subset who responded to a survey; and automated predictions would not require any class time at all to generate. In the next section, I will present related work from EDM/LA to show the feasibility of using these methods to produce such feedback.

Educational Data Mining and Learning Analytics

Educational data mining [EDM] and learning analytics [LA] are relatively new fields that focus on drawing conclusions through computer analysis of large educational data sets (see Baker & Siemens, 2014 and Siemens & Baker, 2012 for more discussion of the precise similarities and differences between the two, which are not germane to this dissertation). In this section, I will discuss prior research that has used methods from these fields to study either course ratings or data from course management systems

[CMSs; not to be confused with the broader "content management system"] or learning management systems [LMSs].

In education, the growth in electronic course management systems and their attendant data (referred to as log data or trace data) has received a great deal of attention from EDM/LA researchers. Morris, Finnegan, and Wu (2005), working with interaction data from online courses, were able to create a model that explained 31% of variance in student achievement. In an influential early article, Romero, Ventura, and García (2008) discuss how to use data from Moodle and outline many common data mining methods that can be applied to such data. Another early work that used CMS data was the Course Signals system at Purdue University, first used in 2007 (Arnold & Pistilli, 2012).

Course Signals used students' interactions with Blackboard to measure students' "effort", which was input, along with other factors, to an algorithm that predicted their likelihood of success in the particular course (Arnold & Pistilli, 2012). Course Signals' algorithm was not automatically calculated, however; instructors chose when to run it and generate new probabilities for students (Arnold & Pistilli, 2012). In addition, feedback generated by Course Signals was summative – students were not told what they should do differently, and actions were not suggested to administrators or faculty (Arnold & Pistilli, 2012; Gašević et al., 2015).

Following these early studies, many researchers have built systems to predict student success, or to identify students at risk of dropping out (Ferguson et al., 2016). For example, Joksimović, Gašević, Loughin, Kovanović, and Hatala (2015) explored relationships between student achievement and student interactions with various aspects

of interaction in a learning environment, finding that different interaction patterns affect students' grades. (Gašević et al., 2016) found that predictions from log data explained vastly different amounts of variability in students' grades depending on the precise context of the course. Gašević et al. (2016) also note that while trace data is predictive, it is not always actionable, although they hypothesize that a system like theirs could help determine whether a course is working as designed.

Fewer learning analytics researchers have investigated topics beyond student success (Papamitsiou & Economides, 2014). Valsamidis, Kontogiannis, Kazanidis, Theodosiou, and Karakos (2012), using unsupervised Markov clustering, created clusters based on students' interactions with an LMS, then grouped courses which contained students who had similar interaction patterns according to the clustering. They suggest that the methods and metrics they use could be employed to rank courses, telling instructors whether their students use the LMS for their course more or less than for other courses (Valsamidis et al., 2012). Kontogiannis, Valsamidis, Kazanidis, and Karakos (2014) work with supervised naïve Bayes classifiers to determine students' opinions from any text students enter in a course. They suggest using this process as part of the "course evaluation process" (Kontogiannis et al., 2014). Finally, Krull and Leijen (2015) discuss the feasibility of using learning analytics to offer formative feedback to student teachers. They conclude that while LA could provide valuable feedback, teaching is a complicated process that is difficult to model (Krull & Leijen, 2015).

Particularly relevant to this project is a study conducted by Babić (2017), predicting students' academic motivation from their LMS activity. Babić used data from

129 education students at a Croatian university, constructing seven features that described the students' interactions with different aspects of the LMS. Models using four of the features (a correlation analysis led to three features' removal) were able to accurately classify students with above-average academic motivation scores on a psychometrically validated instrument (Vallerand et al., 1992). Howard and Schmeck (1979), noting that student motivation to take a course influences their course rating, find that measurements of motivation taken at the end of the course are sufficient to control for this influence. Although more recent studies discuss the relationship between students' initial interest in a course and their rating (e.g., Feistauer & Richter, 2018), few have looked for a relationship involving students' overall academic motivation. However, it seems reasonable to expect that students who are more motivated will be more likely to rate a course and to be more careful and thoughtful in their rating.

Although many EDM researchers have studied CMS data, there are additional methodological challenges when working with course rating data. As ratings are anonymous, it is impossible to link individual students' Moodle activity with their answers on the course rating instrument. In addition, student response rates on course ratings are generally low, so it is necessary to draw conclusions about classes based on the available responses. These challenges mean that researchers must work at the class level, looking for relationships between the CMS activity of all students in a class and the course ratings given by responding students in that class. Most EDM/LA research above the student level has focused on small groups of students, either identifying groups in classrooms or assessing how well a group is working (e.g., Gweon et al., 2011; Kay et al.,

2010, 2006; Martinez et al., 2011; Martinez-Maldonado et al., 2013). Class-level work is less common and typically looks at all students in a class as individuals, identifying those who are struggling so an instructor can offer extra help (e.g., Aslan et al., 2019). Some research has also compared student- or observation-level predictions to predictions aggregated up to the class level in order to combat overfitting on classes with more observations (e.g., Kelly et al., 2018).

Summative and Formative Assessment in Learning Analytics

Unlike the conventional, periodic assessment process, learning analytics researchers often build systems that give more immediate feedback (Pardo, 2014). This allows researchers to measure student learning processes as they occur (e.g., Gowda et al., 2013). These systems are able to identify potential problems much earlier than is feasible with traditional assessment, giving teachers areas to focus their attention.

Many current efforts in learning analytics are still summative, providing dashboards or other lists of students who need interventions, but not suggesting specific actions (Gašević et al., 2015). The difficult nature of providing formative assessment is recognized in learning analytics research, with Macfadyen, Dawson, Pardo, and Gašević (2014) terming the problem of assessment a “wicked problem”. Macfadyen et al. (2014) discuss the importance of actually using the data in analytics to improve student learning, a sentiment echoed by Gašević, et al. (2015).

As mentioned earlier, there has been discussion on the increased formative ability of midterm course ratings (Brown, 2008; Hampton & Reiser, 2004). While providing the

same information earlier does not necessarily make it formative, it does give recipients more time to react to it. Many learning analytics systems, due to providing more immediate and specific feedback, are implicitly assumed to be providing formative feedback (see e.g., Greller et al., 2014; Rojas & García, 2012).

EDM, LA, and Student Ratings

Although there are some studies (see citations above; see also Islam, 2018; Makhlouf & Mine, 2020 who discuss automated analysis of students' comments) that discuss the use of educational data mining or learning analytics methods in rating courses, there are relatively few such studies that focus on quantitative course rating data. M.C. Wang, Dziubian, Cook, and Moskal (2009) used classification and regression trees [CART] to predict overall instructor rating from other items on a course rating instrument, finding rules that allow them to predict both good and poor overall ratings with 97.6% accuracy. They conclude that such an analysis may help administrators determine what items students consider important – the items which are predictive of an overall rating.

Jiang, Javaad, and Golab (2016), using a data set with 257,612 rating forms, look at the predictive value of certain items as well, but also discuss the entropy of various items in data from many students – essentially, how much variability there was in students' answers to items and how that variability was affected by other properties of courses. Notably, they find that teaching quality has less entropy than overall course rating – that students agree more on a course's teaching quality than they do on a course's overall quality (Jiang et al., 2016). They also find that there is less entropy in first-year students'

ratings of courses, which they hypothesize is due to new students not knowing what they like (Jiang et al., 2016).

Clow (2012) discusses the importance of getting learning analytics feedback to teachers and students, past the managers who see the feedback now. Similarly, as discussed earlier, there is a great deal of research on the role of course ratings in education, but much of the rating data is used by administrators for personnel decisions. Pardo (2018) points out that while there is a great deal of research on feedback in education, few models of feedback include the possibility of using data as is done in learning analytics. With the wealth of data available from students' interactions with course management systems, now is the right time to ask how learning analytics can generate more formative feedback to help instructors in higher education improve their practice of teaching.

As discussed by Gašević et al. (2016), it is unlikely that a single system would be able to work for all different types of courses, but learning analytics should be able to help instructors better understand the functioning of their courses. Comparing trace data from LMSs with data from course ratings, it may be possible to determine student behaviors that are linked to satisfaction or dissatisfaction with different aspects of a course, giving instructors early feedback that they will be able to act on.

Conclusion

Limitations aside, a preponderance of the research on course ratings suggests that – with proper use – course ratings are generally a reliable and valid measure of teaching quality in a course. However, the quantitative nature of ratings, and the lack of

institutional interest in using ratings to improve courses that are not at the bottom of the barrel means that ratings have little formative utility. Learning analytics researchers, while still arguing for more formative use of analytics, suggest that the real-time feedback presented by many systems can significantly improve teaching and learning. A system that presents predicted course ratings to faculty earlier would be a good first step in improving the formative utility of the ratings.

RESEARCH QUESTIONS

1. Can student LMS data predict course rating scores, and if so, from how far in advance?
2. Can student LMS data predict class-level response rates on course ratings, and if so, from how far in advance?

CHAPTER 3: THEORETICAL CONTEXT

The theoretical grounding of course ratings does not receive a great deal of attention in the literature. Kolitch and Dean (1999) note that the design of most course rating instruments is informed by a teacher-centered, knowledge transmission model of teaching, usually in a lecture or lecture-and-discussion format (see also Centra, 1993; Edström, 2008 for a similar perspective). Most items on rating surveys focus on behaviors in a classroom which are interpreted as evidence of effective teaching (Kolitch & Dean, 1999), although more recent research and surveys are shifting – at least superficially – to learner-centered models, particularly in the Australian and British contexts (Barrie et al., 2008; Talukdar et al., 2013; Darwin, 2021; see also individual studies, e.g., Erikson et al., 2016; Tucker et al., 2003; Shah et al., 2017). In the learner-centered models, students are asked more questions about how the instructor and the course helped them understand the subject and fewer questions about behaviors displayed by the instructor – teaching is judged as good if it works well for the students, rather than if it is done in a certain way. Note that while simple satisfaction surveys appear more similar to learner-centered models than teacher-centered, scholars such as Erikson et al. (2016) argue that these surveys are performed to fulfill administrative requirements and not to help improve classes for learners. Borch et al. (2020), surveying students in Norway, find that students expect that the rating process uses learner-centered models and are discouraged when it focuses on teachers. This finding also affects how students respond to ratings surveys – if they believe that the process centers around their learning,

their answers will be more closely related to the learning they perceived happening in the class.

In particular, the question of why students give the ratings they do is underresearched in the literature. Critics of ratings tend to focus on biases or claim that ratings are motivated by grades (e.g., Stroebe, 2020; Uttl et al., 2017). Proponents typically argue that ratings do reflect something about the effectiveness of the course – although whether that "something" is teaching, learning, or another related concept is rarely made explicit. However, some scholars have offered more specific explanations of why students rate the way they do.

From a psychometric perspective, Valencia Acuña (2017) discusses students' different response styles when filling out ratings surveys – whether students tend to answer all questions similarly, only choose responses at the ends of the scales, or choose midpoint responses, regardless of item content. He found that students tend to answer more questions positively than expected, suggesting that ratings overestimate teaching quality.

Jiang et al. (2016), Park and Dooris (2019), and M.C. Wang et al. (2009) have used decision trees to analyze course ratings by attempting to predict overall scores from other survey items. These analyses help to explain which items are most related to the overall score, and thus which aspects of the course are most important to students. Unfortunately, it is difficult to compare and synthesize results from these three studies – each is based on a different course rating instrument used at a different university.

M.C. Wang et al. (2009) found that Facilitation of Learning and Communication of Ideas and Information were the most important predictors of overall instructor rating, with Organization of the Course, Assessment of Student Progress, and Instructor Interest in Your Learning also being important. Jiang et al. (2016) worked with an instrument that focused more on the instructor's actions and attitudes and less on the student, finding that the instructor's organization and clarity and their response to questions were the most important predictors of teaching quality, with the instructor's encouragement of independent thinking and the professor-class relationship also being quite important. The instrument discussed in Park and Dooris (2019) had items that more directly asked about learning ("The instructor helped me to better understand the course material" and "The instructor made the class intellectually stimulating"); these items were found more predictive of overall teaching quality than items that asked about organization or course grading.

The differences in these findings could be explained by students' different expectations of the rating process (as discussed in Borch et al., 2020), but across the three studies, students' perceptions of the instructor's relationship to the class were predictive of the overall rating. The formation of students' perceptions of courses has been studied from the perspective of attitude formation and change – what causes students to become satisfied or dissatisfied with a course and how those attitudes affect their responses. Gee (2017), studying students' completion strategies when filling out course-of-study rating surveys, found that students' attitudes toward the course were most significantly influenced by personal relationships with staff members and "landmark events," such as a

particularly meaningful lecture or a difficult experience accessing course materials. Landmark events in particular are known to be important factors in attitude formation (Tourangeau et al., 2000).

Researchers have also explored the relationship between students' learning attitudes and their interactions with intelligent tutoring systems or learning management systems. Many scholars have looked at relationships between students' attitudes and emotions and their activity or performance in intelligent tutoring systems or educational games (see, e.g., Novak & Johnson, 2012, for a shorter overview; Calvo et al., 2015 for more background on affect in learning analytics). One analytics study is particularly relevant to this work. Tempelaar, Rienties, and Nguyen (2017) found that attitudes and emotions, measured by various self-report instruments, were predictive of certain indicators of LMS activity. While the work by Tempelaar et al. (2017) lacks post-hoc controls, it is still notable as, unlike most other studies, the authors explored relationships between emotions and attitudes that were measured at the beginning or midpoint of a course and LMS activity collected over the entire semester.

Learning analytics research seldom comes from firm theoretical grounding (Clow, 2013; see also Dawson et al., 2019, who note that although LA researchers do use theory, they rarely refine or revise it). However, some scholars working with data from learning management systems have focused on the meanings in students' interactions (interaction theory; Moore, 1989; Hillman et al., 1994), including their interactions with the system, with their instructors, and with other students (e.g., Conijn et al., 2017; Joksimović et al., 2015). Joksimović et al. (2015) discuss the relative importance of these interactions on

learning, concluding that students who spent more time interacting with the LMS (distinct from interacting with instructors or peers *through* the LMS) had better learning outcomes. Yu et al. (2020) found that students' "academic emotions," such as anxiety and boredom, affected their interactions in an online learning system, although the relationships found were not very strong. This is important to this study as those emotions are likely related to students' opinions of courses.

Although most uses of interaction theory by LA researchers (including both Joksimović et al., 2015 and Yu et al., 2020) focus on pure distance education, where no course activities take place in a physical classroom, the central idea that learning can be described through interactions between students, teachers, learning material, and learning systems is equally applicable to blended learning environments, where some learning takes place in a classroom and some takes place online. This is important as this study's data does not come from online courses.

Finally, it may seem too obvious to state, but teachers' presence and actions do affect students' actions and outcomes in online and blended learning systems (Anthony, 2019; Law et al., 2019). This follows from both interaction theory and common sense. This study follows Babić (2017) and Tempelaar et al. (2017), working with the assumption that a student's attitude, affected by the teacher and course activities, in turn affect their interaction patterns with the course LMS, as discussed by Yu et al. (2020). These attitudes and students' academic motivation should also, as discussed by Gee (2017), affect students' end-of-course ratings. Therefore, features distilled from students' interactions should be predictive of those end-of-course ratings. Predicting ratings would

have significant benefits to the course evaluation process: instructors would receive feedback earlier, predictive models could offer insight into behaviors that influence ratings, and an explainable model might help add credibility to the rating process.

CHAPTER 4: METHOD

Data

In order to investigate the research questions, a dataset is needed that contains both course management system log data and end-of-course rating data. For this study, that data is drawn from a large Australian university. The data set includes anonymized course rating data as well as anonymized Moodle interaction records – linked at the course level – from several hundred classes in 2016. The course rating instrument used at this university contains eight Likert-type items on all surveys, with some open-ended questions and additional items for online courses.

Several criteria were used to clean the dataset in preparation for analysis. First, courses were removed which could not be definitively matched to rating data, leaving 341 courses in the dataset. In order for the features to reflect a reasonable amount of variation and improve the likelihood of distinguishing good courses from bad, the dataset was also filtered to exclude courses with insufficient interaction data. Courses with fewer than 10,000 Moodle log events over the term were removed; it is likely that such courses did not have enough students consistently interacting with Moodle for the features to be meaningful. In addition, courses that had fewer than five complete ratings surveys were removed as ratings from courses with fewer responses are less reliable. After filtering the dataset on these criteria, 207 courses remained. The course sizes varied, with the smallest course having only 14 students accessing Moodle over the term and the largest having 949. 50% of the courses had between 76 and 229 students. In total, the dataset contained

data from 18,117 students. Note that these student counts are not from course rosters, instead coming from the Moodle IDs that accessed the site.

Moodle use varied over the sample, but most courses saw about sixteen weeks where at least half of the class logged in.[4] Almost half of the courses had eight or more weeks where at least 75% of the class accessed the site. However, the overwhelming majority of Moodle use was passive, with students only referring to information that already existed on the course site. Students averaged at least one active interaction (such as submitting an assignment, making a forum post, or writing a message) in less than 1% of course-week events. As most of the features focused on students' activity, algorithms were fit and tested on an active subset of the sample – courses where students averaged at least one active interaction per week – in addition to the full sample. The active subset contained 96 courses and data from 12,989 students. It should be noted that Moodle course IDs are not strictly unique to course sections; in some instances, one Moodle course ID corresponded with multiple meeting locations in the same term, suggesting that multiple sections used the same course ID. 106 courses, about half the sample, had no more than one course location per term, although that does not necessarily limit the course to one session; it may have had multiple sessions meeting on the same campus.

[4] Full-term courses at this university last for twenty weeks, including a two-week break midway through the term and a two-week exam period at the end. Half-term courses last for twelve weeks, including a one-week exam period at the end. Twenty courses in the sample were half-term courses.

Factor Analysis

As discussed in the literature review, results from course ratings surveys should be analyzed as a factor, rather than using individual rating items. For this study, the rating instrument used always contained at least eight items, with some additional items for courses that used certain online features. The course rating data here have a complicated, multiply nested structure – each course contains multiple students, each of whom may have responded to multiple rating surveys. As it is not feasible to perform an exploratory factor analysis (EFA) with such a nested structure, multiple EFAs were conducted. For the first, one student's responses were randomly selected from each course, allowing the data to be treated as though there was no course-level nesting and making it less likely that there were multiple responses from any individual student. This EFA contained 440 responses. For the second and third EFAs, a two-level structure was used, with nesting at the student level and at the course level, respectively. Half of the available data (26,229 responses) was included, leaving the remaining half for a confirmatory factor analysis (CFA).

All EFAs and CFAs were conducted using Mplus (Muthén & Muthén, 2017) and geomin rotation, using only the eight items that are always present in the rating instrument. All found a two-factor structure, with four items strongly loading (>0.65, with all but one item >0.85) on each factor and the non-loading items all having values below 0.21. One factor contains items which relate to the course itself and the other factor related to the instructor, although there was very substantial correlation between the two factors (between 0.686 and 0.850, depending on the response nesting). The two-

factor structure was verified using a confirmatory factor analysis on the held-out portion of the data; the model fit well, with $\chi 2 = 30.19$ ($p = 0.0494$), comparative fit index and Tucker Lewis index of 1.00 and a root mean-square error of approximation (RMSEA) of 0.037, with an 80.7% probability of RMSEA ≤ 0.05.

To combine the ratings to factor scores, a method described in Comrey and Lee (1992) was used. Each rating was standardized to a z-score, then the four ratings that loaded onto each factor were summed. These summed factor scores were used to create the course-level outcome variables, as discussed in the next section.

Outcome Variables

Models in this study used seven different outcome variables, six related to rating data and one related to response rate. As rating data is not identifiable down to individual students, the individual ratings must be aggregated to the class level. The instructor and course factors are each aggregated by computing a mean and a median value, leading to four outcome variables, two per factor. To allow for the use of binary classification methods, rating values were also split into two groups based on mean rating, with the cutoff at 0. 129 sections received a mean course rating of at least 0; 132 sections received a mean instructor rating of at least 0.

The last outcome variable is the percentage of students who fill out a course rating form for each course. These percentages were top-coded at 100% (1.0) to account for two courses which saw more students fill out rating forms than access the Moodle site. After top-coding, the mean response rate was 21.2%, the median was 18.5%, and the standard deviation was 14.9%. These response rates are slightly lower than other course rating

data in the literature (Hoel & Dahl, 2019). Table 1 shows descriptive statistics for all outcome variables. Notably, most of the outcome variables were highly correlated with each other, with Pearson correlations at or above 0.6, although response rate was uncorrelated from the others, with correlations below 0.16.

Table 1. Summary statistics for outcome variables.

Outcome variable	Mean	Median	Standard deviation	Skew	Kurtosis	Min	Max
Course mean	0.32	0.64	1.68	-0.79	0.85	-4.97	4.17
Course median	0.69	0.99	1.77	-0.28	1.04	-5.26	4.17
Instructor mean	0.22	0.54	1.94	-1.35	3.16	-7.71	3.55
Instructor median	0.86	0.42	2.29	-0.88	1.53	-8.68	3.55
Response rate	0.21	0.19	0.15	1.98	6.91	0.02	1.00
Course mean binary	0.62	1	0.49	-0.51	-1.75	0	1
Instructor mean binary	0.64	1	0.48	-0.58	-1.68	0	1

Analysis

In educational data mining (EDM), prediction analysis generally starts by establishing a ground truth, some measurement of a construct that is assumed to be accurate (Baker, 2010), which the researcher is trying to predict. In this study, the outcome variables serve as that ground truth. Next, with the exception of approaches like deep learning and autoencoding, features are typically created. Features, in EDM, are essentially distillations of raw data that highlight behaviors of interest – for example, the number of consecutive problems students answered correctly in a learning system, or the average length of time students spend using an LMS each session. Features are usually created with the help of domain experts, who can identify behaviors which they believe are related to the ground truth (Baker, 2019). Once the features have been produced, various algorithms are used to build models that predict the ground truth from the features.

In this study, features were primarily patterned after other features which were used successfully in studies of Moodle dropout prediction (Buschetto Macarini et al., 2019; Cerezo et al., 2016; Monllaó Olivé et al., 2019; Motz et al., 2019). Some of them were constructed weekly at the individual level (e.g., the average length of a student's Moodle sessions over the week), some were constructed weekly at the class level (e.g., the number of assignments handed in late over the week), and some were constructed at the class level but did not change over the entire time of the course (e.g., the number of students accessing the Moodle site). The full list of the 33 features included is presented

in Appendix A. In order to control for the varying size of courses, class-level features were divided by the number of students who accessed the course over the semester. Finally, K-means clustering (Lloyd, 1982) was applied to group all classes into four clusters based on feature values. The resulting clusters were added as a final feature. All feature values were standardized using an interquartile range-based scaler (SciPy 1.0 Contributors et al., 2020).

As the Moodle course activity data for this study was available at the individual level, but course ratings could only be identified down to the class level, individual-level features had to be aggregated to the class level before they could be used in models. The approach used in this dissertation is similar to one that is sometimes used in EDM/LA when log data is recorded at a finer-grained level than label data (e.g., Sao Pedro et al., 2013). To aggregate features, all individual values for a given week were collected, then means, medians, and standard deviations were computed for each feature, for each course. After the course-level aggregation was performed, additional week-level aggregation was performed to include students' past behaviors in the models. The process was similar: each value for each class-level feature was collected for all prior weeks. Sums, means, and standard deviations were computed for each feature, for each course, although not all of these values were used, as described below. These week-level aggregation features were included in the models in addition to the class-level aggregated features for that week.

The aggregation processes described here causes a dramatic increase in the number of features available to the machine learning algorithms: one initial feature may

be aggregated to three (mean, median, standard deviation) at the class level, then each of those three may be aggregated to another seven (the three past calculations, the minimum, maximum, and sum of the past weeks, and its current-week value) after including past behavior, giving a total of 21 post-aggregation features. Given this degree of inflation, even a small number of initial features can lead to hundreds of features post-aggregation, making overfitting a serious threat. This is the "curse of dimensionality" first discussed by Bellman (1961).[5] Although the precise point at which dimensionality becomes a problem varies based on data and algorithm used, a common rule of thumb is that there should not be more features than cases. To reduce dimensionality, there exist various automated methods (see Cunningham, 2008). However, these methods are not themselves entirely safe from threats due to dimensionality (Altman & Krzywinski, 2018). Wujek, Hall, and Güneș (2016), echoing the conventional wisdom, argue that features should first be pruned manually. For these reasons, no features had all of their aggregated values included. Instead, aggregation calculations were individually chosen on a theoretical basis – whether it made sense to include each calculation for each feature, given that feature's meaning (cf. Sao Pedro et al., 2012). Appendix A lists the features and the aggregation calculations that were included.

Even after limiting aggregations, 117 features or aggregations of features were available for the algorithms used, a number which was likely too large and caused overfitting. To address that issue, a nested, cross-validated forward selection process was

[5] This causes more problems than just overfitting; for a detailed discussion, see Zimek et al., 2012.

conducted. In this process, data was subset into two unequal groups. The first, larger group was then used for the forward selection process. To perform forward selection, single-feature models were trained on each feature in turn; the best-performing feature (according to AUC for the classification models and according to Spearman correlation for the regression models) was added to the model, then the process iterated, keeping the first feature and training two-feature models using the remaining features. The forward selection process continued until five iterations were completed with no improvement in model performance; the best-performing features were then used to train a model which was tested on the smaller held-out group. These features were used in the cross-validated training and testing process on the entire dataset. As most of the outcome variables were highly correlated, the forward selection process was only conducted twice for the regression models, using the response rate and the mean of the course rating scores, and once for the classification models, using the binary of the mean course rating scores.

Apart from the cross-validation used for forward selection, the main model fitting and testing process also made use of cross-validation. Cross-validation is a process which involves splitting the data and using some of it to train the model and the rest to evaluate it, possibly with a third portion further held out as test data (Richard A. Berk, 2016). In EDM, cross-validation is a common method to address overfitting (Baker, 2019); cross-validation repeats the splitting process multiple times, holding out a different portion of the data each time as an evaluation set, then training on the remainder of the data (Efron & Gong, 1983). In this study, cross-validation was used at the class level, leaving out several classes each time, training the model on the remaining classes, then evaluating on

the held-out classes. Regression models are typically evaluated in educational data mining by looking at the root mean square error [RMSE] and simple Pearson or Spearman correlations (r) or squared correlations (r^2; Baker, 2019). Root mean square error is the square root of the average of the difference between the predicted values and the actual values. Spearman correlations and RMSE values were primarily used in this study. Classification models are evaluated using the area under the receiver operating characteristic curve (AUC; Bradley, 1997), which is the probability that a model will identify a positive case as more likely to be positive than a negative case, averaged over all pairs of positive and negative cases.

Although this study aimed to predict a course rating score for each week over the term, as discussed earlier, not all classes used Moodle extensively every week. As the goal of the study was to use all students' behaviors to predict rating scores which were only available at the course level, it did not make sense to generate predictions in weeks where only a few students accessed the course site – nearly all of the constructed features focus on students' activity, which would not have been representative of the class in these weeks. Therefore, no predictions were generated for low-usage weeks and they are not included in the results, but data from those weeks was still used in the week aggregation process. After removing weeks where fewer than 10% of students accessed the course site, 4016 course-week cases remained, with 1857 course-week cases in the active subsample.

Finally, four different algorithms were used to train and test models. XGBoost (T. Chen & Guestrin, 2016), a popular gradient boosting algorithm, classification and

regression trees (Breiman, 1998), random forest classifiers and regressors (Tin Kam Ho, 1995), and linear and logistic regression. Except for XGBoost, all algorithms were used through the SciPy software package (SciPy 1.0 Contributors et al., 2020). Each algorithm was used in turn to select its own features through forward selection as described earlier, with a separate selection process for the active subset; the algorithms were then trained and tested on the entire dataset and on the active subset, using cross-validation. RMSE values and Spearman correlations were recorded for regression models; RMSE values and AUC values were recorded for classification models.

CHAPTER 5: RESULTS

Results from the final models are show in tables 2-5. Tables 2 and 3 show AUC values, RMSE values, and Spearman correlation coefficients for the final models using XGBoost, random forest, classification and regression trees, and linear and logistic regression on the entire dataset, with tables 4 and 5 showing the same statistics on the active subset. As discussed in the methods section, each of these models was fit after a cross-validated forward selection procedure chose features to retain. Table 1 is also reproduced; its standard deviations make the RMSE values in Tables 2 and 4 easier to compare.

The best performing models on the entire sample only achieved Spearman correlations slightly above 0.2, with one linear regression model seeing a correlation of 0.31 on median course rating. RMSE values depended on the individual outcome variables' ranges, but approximately equaled one standard deviation for the respective outcome variable for better models. Linear regression models tended to have the best RMSE values, followed by random forest, XGBoost, and finally regression tree models, which had the highest RMSEs. The model performance did not vary greatly across the different outcome variables.

The binary classification models attained similarly lackluster results. AUCs ranged from 0.51 to 0.61, with the highest value achieved by logistic regression on instructor classification. RMSEs followed the same pattern as the regression models, with logistic regression best, then random forest, XGBoost, and finally classification tree models. Again, RMSEs for the better models approximately equaled one standard

deviation of the outcome variable. As a zero-feature model that simply predicted the mean of the outcome variable in every case would have an RMSE of one standard deviation, this suggests that in terms of absolute error, none of the regression or classification models are much better than simply taking the mean of the outcome variable.

Results for regression models on the active sample – courses where students averaged at least one active interaction with Moodle per week – were a little bit better for XGBoost, but a little bit worse for the other algorithms. XGBoost achieved Spearman correlations above 0.25 for all variables except response rate, although its RMSE values – and the RMSE values of all other algorithms – were still consistently above one standard deviation. Spearman correlations for algorithms besides XGBoost were close to 0.20. On the binary classification tasks, most models performed slightly better on the active portion of the dataset: although random forest did not achieve any AUCs that were better than random chance, logistic regression and XGBoost both saw AUCs above 0.6 on classifying above-average instructors.

Table 2. Summary statistics for outcome variables.

Outcome variable	Mean	Median	Standard deviation	Skew	Kurtosis	Min	Max
Course mean	0.32	0.64	1.68	-0.79	0.85	-4.97	4.17
Course median	0.69	0.99	1.77	-0.28	1.04	-5.26	4.17
Instructor mean	0.22	0.54	1.94	-1.35	3.16	-7.71	3.55
Instructor median	0.86	0.42	2.29	-0.88	1.53	-8.68	3.55
Response rate	0.21	0.19	0.15	1.98	6.91	0.02	1.00
Course mean binary	0.62	1	0.49	-0.51	-1.75	0	1
Instructor mean binary	0.64	1	0.48	-0.58	-1.68	0	1

Table 3. Spearman correlations and RMSEs for regression models fit and tested on the entire dataset. Best values are in bold.

Outcome variable	Metric	XGBoost	Random forest	CART	Linear regression
Mean course rating	Spearman	0.24	0.16	0.22	**0.26**
	RMSE	1.95	1.76	2.36	**1.63**
Median course rating	Spearman	0.22	0.18	0.14	**0.31**
	RMSE	2.09	1.86	2.50	**1.72**
Mean instructor rating	Spearman	0.23	0.23	**0.24**	0.20
	RMSE	2.25	1.99	2.72	**1.85**
Median instructor rating	Spearman	0.21	0.18	**0.27**	0.24
	RMSE	2.63	2.34	3.20	**2.20**
Response rate	Spearman	0.14	0.13	**0.21**	0.10
	RMSE	0.20	0.17	0.20	**0.15**

Table 4. AUC and RMSE values for binary classification models fit and tested on the entire dataset. Best values are in bold.

Outcome variable	Metric	XGBoost	Random forest	CART	Linear regression
Above-average course rating	AUC	**0.56**	0.52	0.51	0.52
	RMSE	0.60	0.52	0.69	**0.49**
Above-average instructor rating	AUC	0.54	0.55	0.51	**0.61**
	RMSE	0.59	0.51	0.68	**0.47**

Table 5. Spearman correlations and RMSEs for regression models fit and tested on the active subset. Best values are in bold.

Outcome variable	Metric	XGBoost	Random forest	CART	Linear regression
Mean course rating	Spearman	**0.29**	0.22	0.20	0.12
	RMSE	2.01	1.82	2.27	**1.66**
Median course rating	Spearman	**0.27**	0.12	0.20	0.13
	RMSE	2.09	1.88	2.39	**1.72**
Mean instructor rating	Spearman	**0.30**	0.08	0.24	0.20
	RMSE	2.28	2.08	2.44	**1.87**
Median instructor rating	Spearman	**0.30**	0.23	0.19	0.20
	RMSE	2.69	2.37	2.98	**2.20**
Response rate	Spearman	0.16	0.20	**0.23**	0.19
	RMSE	0.16	0.14	0.20	**0.13**

Table 6. AUC and RMSE values for binary classification models fit and tested on the active subset. Best values are in bold.

Outcome variable	Metric	XGBoost	Random forest	CART	Linear regression
Above-average course rating	AUC	0.52	0.47	0.51	**0.58**
	RMSE	**0.50**	0.55	0.70	**0.50**
Above-average instructor rating	AUC	0.61	0.47	0.54	**0.64**
	RMSE	**0.48**	0.54	0.68	0.49

The process of feature building, model fitting, and model testing was iterative and took place over several months. Figures 1-3 show graphical representations of this iterative process: they depict the Spearman correlations and AUCs, respectively, of models changing over time as new features were added and new approaches were tried. These correlations come from XGBoost regression models predicting mean course rating (Figure 1) and rating response rate (Figure 2), with Figure 3 showing AUCs of an XGBoost classification model predicting higher-performing instructors. These graphs include results from the entire dataset as well as results from the active subset; results from the active subset are marked with squares. The best-performing regression models only attained a Spearman correlation of 0.25 on cross-validated data, with one model from each outcome variable attaining that value and most models performing far worse. However, the trajectory does not show constant improvement as more features were added.

The first model represented in the figure only had 10 pre-aggregation features available, mostly related to session times and the number of active sessions. The next four data points reflect the addition of features capturing the number of students in the course, the percentage of students who visited over the week, and the percentage of students who were active over the week. The dramatic improvement following that was driven by the introduction of the forward selection procedure. The three next time points, marked with squares, depict results from the active sample: the first without forward selection, the second with, and the third with both forward selection and K-means clusters. After fitting

on the active sample, more class-level features were added and tested, again with forward selection; these features were also tested on the active sample near the end of the process. The final mean course rating model, after features relating to late assignment hand-ins were added, did approach the best model again, with a Spearman correlation of 0.24.

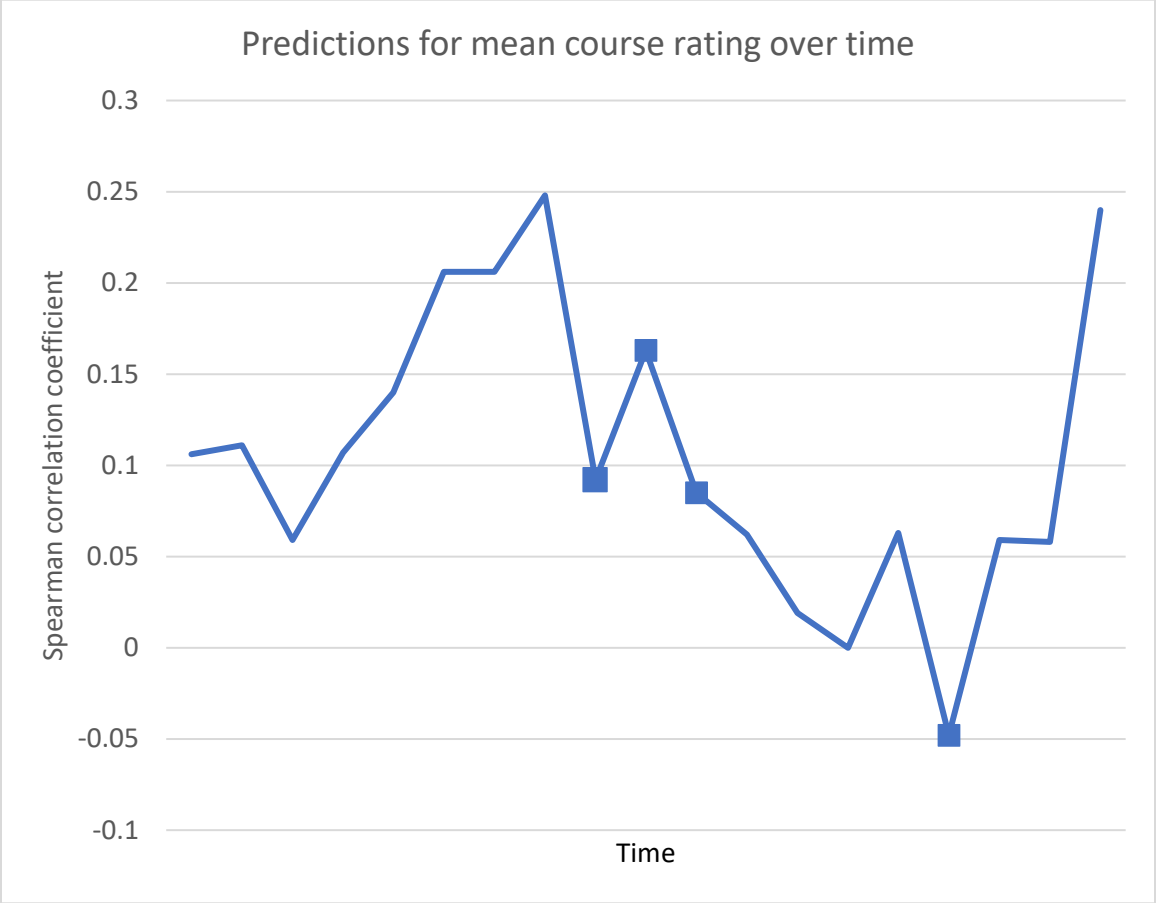


Figure 1. Spearman correlations of XGBoost predictions for mean course rating over time. Squares denote models fit and tested on the active subset.

Figure 2 shows the same process for models predicting the response rate on rating surveys. This graph has fewer data points as response rate models were not trained until midway through the model building process. As before, points representing the active

subset are marked with squares. The same features and approaches were tried for this outcome variable as for mean course rating; the initial time point on this graph is after forward selection had been added. For this outcome variable, none of the models saw correlations below 0.14 and the correlations changed much less as features were added. This model was not helped by the final addition of late hand-ins; it never beat the first model with 14 pre-aggregation features and forward selection.

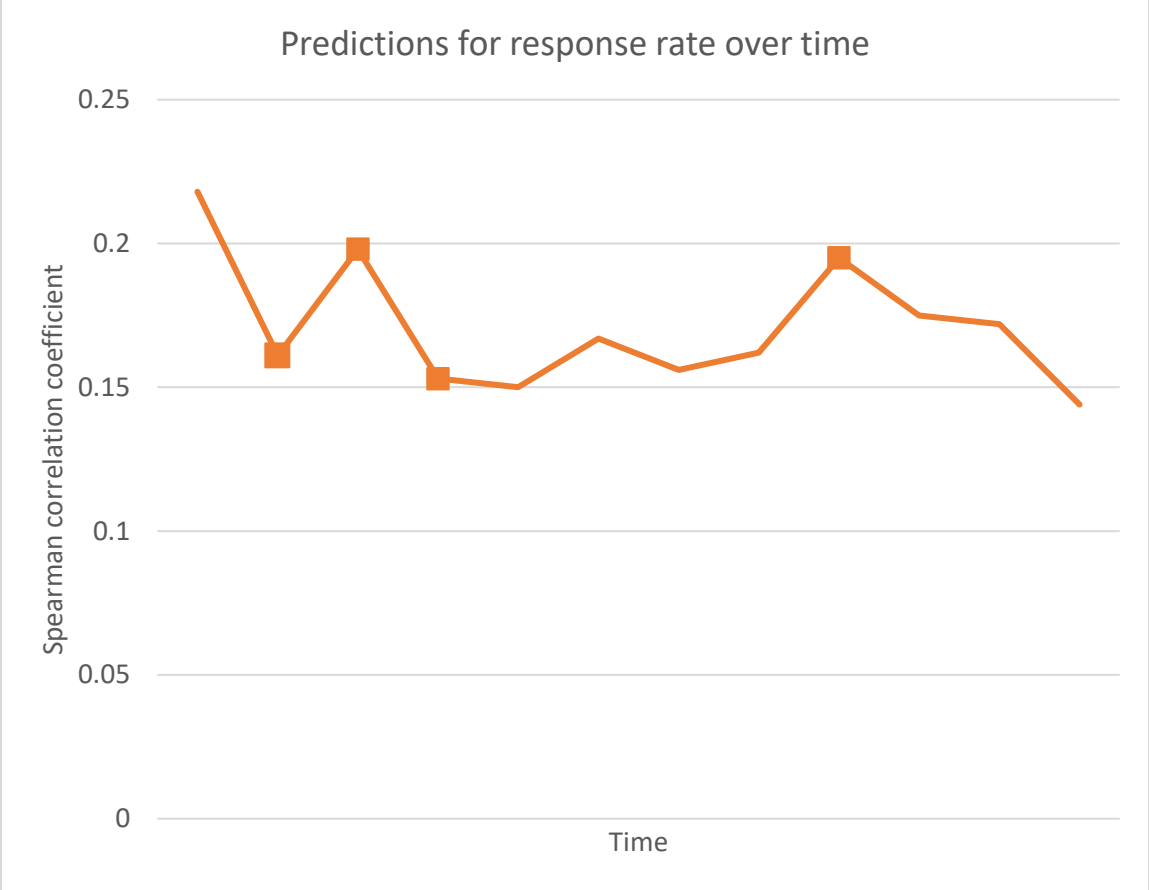


Figure 2. Spearman correlations of XGBoost predictions for response rate on rating surveys over time. Squares denote models fit and tested on the active subset.

Figure 3 shows the AUC values of XGBoost binary classification models which attempted to predict whether a course would receive an above-average or below-average

rating. Binary classification models were added at the same time that models for response rate were first trained. Results from the active subset are again intermingled and marked with squares. Similarly to Spearman correlations, AUC peaked early in the process – at 0.61, on the active subset, with forward selection – and failed to improve despite the addition of more features. Although one model did attain an AUC above 0.6, many saw AUCs at or slightly below 0.5, meaning they performed worse than chance.

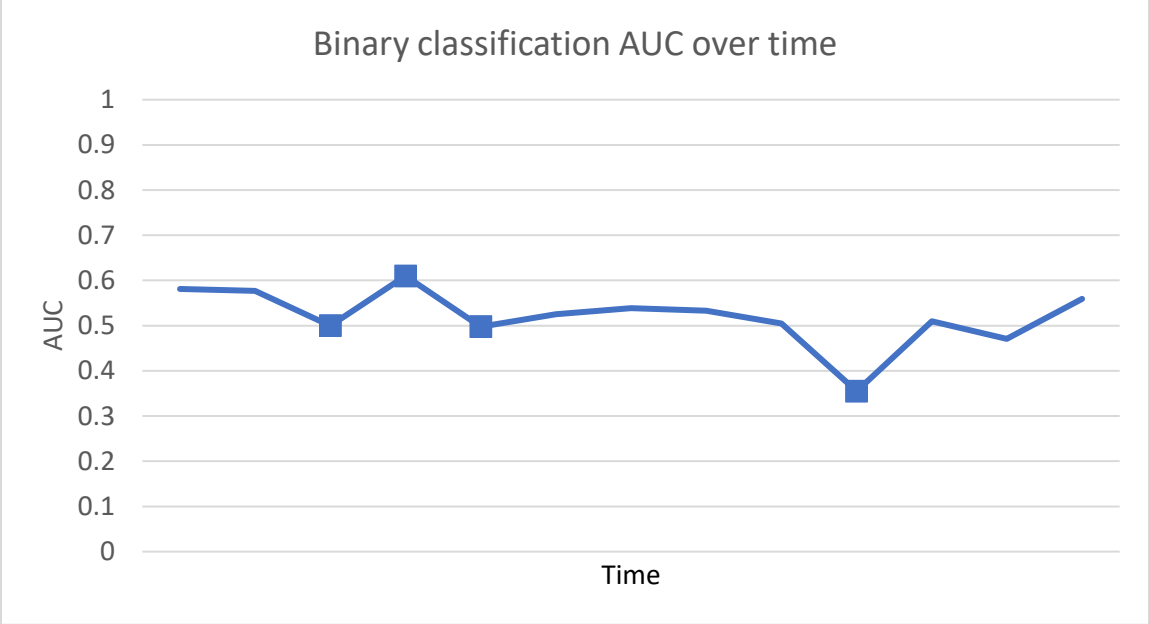


Figure 3. AUC of XGBoost course binary classification models over time. Squares denote models fit and tested on the active subset.

Although the models did not fit well, it may still be instructive to look at the features that were chosen by each algorithm's forward selection processes. As discussed in the methods section, each algorithm ran the forward selection three times for each dataset, twice for the regression models and once for the classification models. In total across all algorithms, 71 distinct features (counting separate aggregations as separate

features) were selected by the forward selection process, 48 features in the entire sample and 53 in the active subset. Although 35 features were only selected once by one algorithm, a few features were favored across algorithms. The median number of active sessions over the week was most popular, being selected 13 times out of a total of 24 models. Other commonly selected features were the number of assessments due that week, the number of active sessions computed at the course level (as though all students in the course were only one student), the median number of actions, the average session time, and the K-means groupings. Looking only at the most successful algorithms across both the entire sample and the active subset, the number of active sessions, average Moodle session time, and time spent on the course forum were frequently chosen.

CHAPTER 6: DISCUSSION

This dissertation was not successful at finding a way to achieve high model goodness when predicting either course ratings or the response rates on rating surveys from students' Moodle activity in seated classes. This study tested a variety of algorithms and a variety of features, many of which had been shown to be productive of students' grades in earlier studies, but none of the models built were reasonably predictive of any of the outcome variables. Many factors made this study more challenging than prior, student-level models with LMS data.

First and foremost, working at the class level dramatically decreases sample size while increasing the feature space. Any feature that is calculated at the individual level must be aggregated to the class level, which makes it more difficult to reliably capture behaviors of interest. In this study, this problem was further exacerbated by the necessity of performing aggregation twice to capture past feature values. For example, suppose that students are more likely to give a course a low rating if their interest in a course gradually wanes – they visit the LMS less often, for less time as the course goes on. While this behavior can be detected in one student by comparing their visits to those of other students in the same class, it's more difficult to compare class-to-class because there are so many more differences between classes: perhaps an instructor chose to emphasize in-class participation over the LMS or perhaps students are busy reviewing for an exam. Further, depending on how individual activity is aggregated together, one student's waning interest may be masked by another student's growing interest.

Class-level analyses also greatly affect sample size. This study included data from more than 18,000 students, which is more students than were included in many other Moodle prediction studies (e.g., Buschetto Macarini et al., 2019; Cerezo et al., 2016; Motz et al., 2019; Quick et al., 2020; cf. Monllaó Olivé et al., 2019, who worked with data from multiple universities), but only 208 classes. This made it difficult to use methods like neural networks, which excel at finding more complicated relationships in larger datasets.

Although class-level differences have not received as much attention by educational data mining researchers (cf. Conijn et al., 2017, who, although working at the student level, discuss class differences in their data), course ratings have been the subject of multilevel analysis in the broader literature (Cho, 2013; Toland & De Ayala, 2005; M.C. Wang et al., 2017), with most scholars finding that there is a notable amount of variation in students' ratings between classes – that is, the ratings of students in different classes show different levels of agreement with the ratings of other students in the same class. This makes intuitive sense, as some classes are effective for nearly all students, while other classes only work for some. It is reasonable to expect the same variation in students' LMS interactions: that in some classes, most students will behave quite similarly to each other, while in others, behavior will cover a wider range (see again Conijn et al., 2017). While these differences would be extremely difficult to quantify, their existence makes the task of building class-level models very complicated as the meaning of feature values would likely shift from class to class. Learning analytics researchers working at the individual level have also noticed features which change their

meaning over time (Baker et al., 2015); this may be the case at the class level as well. It would almost certainly be a problem when working with data over several years as norms around LMS use evolve over time.

Ultimately, successfully predicting course ratings may necessitate models that work at the student level, which would require data that breaches the confidentiality of the ratings. Barring this, it may be interesting to attempt to predict ratings only in very similar courses – perhaps courses in the same department, or even one popular introductory course that is offered in several sections, taught by different instructors. Of course, any attempt to limit the available courses would also shrink the sample size, making it more difficult to determine the most predictive features and increasing the likelihood of overfitting.

Another important challenge of this work relates to the use of LMS data as the only way to measure the activities of seated classes. While LMS logs in an online course capture all or nearly all of the activity in a course, in traditional seated classes much of the lectures, discussions, and work take place offline and are not captured by any data system. In this study, that is exemplified by the rarity of active Moodle sessions as compared to passive sessions: students did turn in work through Moodle and used features like course forums, but the overwhelming majority of their Moodle use was just looking at course content such as readings. Although one of the main aims of this study was predicting course ratings from data that has been collected *in situ*, constructing predictions from data that only represents a fraction of course activity severely limited the accuracy of the predictions.

It is difficult to firmly establish the magnitude of this limitation – more research tends to focus on online courses (e.g., Joksimović et al., 2015; You, 2016), and few if any studies compare the performance of models trained on data from online courses to those trained on data from blended or seated courses. Such comparisons would be questionable in most cases – there are many differences between LMS interaction patterns even within the same delivery method at the same university (Conijn et al., 2017), so it would be exceedingly difficult to determine how much of the difference between online-trained models and seated-trained models was attributable to the delivery method. Nevertheless, it is almost certain that predictions made about seated courses with only LMS data would be significantly inferior to predictions that had access to more information about classroom events or instructor behavior. While researchers have accurately predicted student performance and student dropout in seated courses using only LMS data (e.g., Romero et al., 2013; Zacharis, 2015), those predictions are at the individual level.

In this study, assessment data were included where available to partly address this limitation, but these data were not linked to individual students' Moodle activity or course ratings and thus could only be used at the class level. For example, one feature counted the total number of late assignment submissions made each week in a course, but features could not capture an individual student's Moodle accesses immediately prior to turning in a late assignment. Nevertheless, as many scholars have argued over the relationship between course ratings and grades (G. Wang & Williamson, 2020), it is noteworthy that including aggregated grades in the models in this study did not lead to more accurate predictions.

The comprehensive nature of LMS interaction logs makes it appealing to develop many features and let the algorithms choose the most predictive, particularly given that most studies predicting students' success, dropout, or behavior from LMS data succeed with features which are simple to calculate – such as "calendar visits" (Motz et al., 2019), "Time spent on forum" (Cerezo et al., 2016), or "Time spent on assignment pages" (Quick et al., 2020). While that approach may be feasible in some cases, it likely led to overfitting in this study.

The forward selection process used to combat overfitting improved both the Spearman correlations – from near-zero to 0.2 – and the AUC values – from about 0.5 to 0.55 – but may not have been ideal for these data. Despite the forward selection being cross-validated at the class level, the metrics seen in the innermost cross-validation loop were much better than those outside the loop, and similarly better than those on the larger dataset. This suggests that there may have still been problems with overfitting to training data, which means that the other results presented here may reflect overfitting as well even with the use of cross-validation. The graphs in Figure 1 support this: as more features were added, correlations stayed the same or decreased.

Due to the exploratory nature of this study, there were other limitations as well. Working with course rating data, there was the ever-present possibility that respondents may have differed from nonrespondents. Since the course rating data used was anonymized to protect student privacy, there was no way of knowing if students are rating multiple classes. This affected the cross-validation process; the same students may have appeared in both training and testing folds of the data despite the class-level cross-

validation. In addition, students who rate courses may have differed from those who do not. Ideally, ratings would be missing at random from students in classes, but in fact, research on student responses to course ratings has found that many student-level factors influence both responses and response rates (Adams, 2010; Adams & Umbach, 2012; Macfadyen et al., 2016). This suggests that some students will be more likely to rate courses, so they were likely overrepresented in the data. Essentially, this study was trying to determine differences between populations using small, nonrepresentative subpopulations which may have been more similar to each other than the larger populations that they were taken from. This became particularly troublesome when aggregating individual behaviors to the class level – it was not clear if the aggregations reflected the behavior of the small group of students who actually responded to the ratings surveys.

This study also has some data-related limitations. First, Moodle access data was only present if students had accessed the system. Although students in an online class would have to access the course site in order to complete the class, it is conceivable that seated classes did not require all students to log in. Some other studies have included features that capture the number of registered students who never accessed the course site at all (Monllaó Olivé et al., 2019), and features like this could better capture the dynamic of a class and show that the Moodle data is or is not representative of all students. This is similar to – and compounds – the problem of ratings not representing all students in a course.

In addition, there was very little available data that described each class, and no data on the individual instructors. Class size was included in the models, but beyond size it was not clear if a class was primarily lectures, discussions, or another format; there was also not enough data to include course subject. These and other class-level factors have been shown to be related to course ratings (Benton & Cashin, 2014; Centra, 2003; Darwin, 2021) and their inclusion may have made the models more accurate. Whether instructor characteristics affect ratings is more debated (Darwin, 2021; Fischer & Hänze, 2019), but their inclusion would have made it possible to investigate this issue.

Time Series

One other method investigated in this dissertation was whether the temporal patterns of events may be indicative of students' attitudes towards courses. Researchers have successfully used time series neural networks to predict students' grades in MOOCs (Yang et al., 2017), which naturally led to the idea of using time series methods to analyze feature data in this study. Unfortunately, there were several challenges in using time series algorithms for this project.

The majority of available time series methods focus on either forecasting – predicting future values in the series – or classification – predicting whether a time series represents case A or case B. Time series regression methods are far less common and often consist of using the time series methods to generate additional features, which are then fed to another regression algorithm (see e.g., the implementations presented in

Tavenard et al., 2020). In addition, because most time series methods are univariate (Faouzi & Janati, 2020), the number of multivariate time series regression algorithms is quite small. Finally, time series methods tend to focus on longer series than were used in this study: Bagnall et al.'s (2017) large comparison of classifiers included 85 datasets, all of which had more data points than the 22 here.

In this study, it was decided to use an existing multivariate time series classifier rather than working with regressors. Word Extraction for Time Series Classification (WEASEL; Schäfer & Leser, 2017) and its extension, Multivariate Unsupervised Symbols and Derivatives (MUSE; Schäfer & Leser, 2018) is a domain agnostic time series classifier that has been shown to be accurate on typical classification tasks. WEASEL+MUSE compares favorably to some deep learning methods while requiring much less data for training (Schäfer & Leser, 2018). In this study, WEASEL+MUSE was used via the pyts package in Python (Faouzi & Janati, 2020).

Only courses with 22 weeks of data were included in time series analyses, leaving 154 courses. This also means that all time series predictions were made "at the end of the course", with all of a course's data available for the algorithm. WEASEL+MUSE was trained and tested in a cross-validation loop similarly to the other algorithms, although course groupings were unnecessary as each course only appeared once in the dataset. The algorithm output feature values for each case, which were then used as input to a logistic regression algorithm as described in Schäfer and Leser (2018). To lessen dimensionality threats, week-level aggregation was not performed when working with time series

methods. Features that displayed no variation over any individual course were also not included, bringing the total number of features available to the time series classification algorithm down to 30. This process was followed twice: once for each binary outcome variable.

Unfortunately, the results from the time series classifier were also poor. The precise number of features generated varied based on hyperparameter values, but the default hyperparameters led to between 2400 and 4400 features, depending on cross-validation fold. AUCs and RMSEs were unimpressive, with AUC values ranging from 0.50 to 0.53 and RMSEs all above 0.50. The number of features suggested that overfitting was all but certain; however, adjusting the hyperparameters so as to generate no more than 15 features did not lead to an improvement in AUC or RMSE.

CHAPTER 7: CONCLUSION

Although the models built in this study were not predictive of end-of-course rating scores, this should not be taken as conclusive evidence that ratings cannot be predicted. There are several other avenues which may be worth further investigation. Working with data from online courses would likely lead to more effective detectors, even if such an approach may not eliminate all of the difficulties encountered in this study. Data from online courses more completely represents students and instructors' educational interactions, allowing models to better capture the class environment.

However, the true challenge of this study was aggregating student-level predictors to the class level. Despite this study using features which had been predictive of student grades and dropout in previous research, the resulting aggregated features were not successful at the prediction task. As course ratings surveys are nearly always administered with students having the expectation of confidentiality, it is likely that future models that aim to predict ratings will be built at the class level, which introduces several difficulties. These difficulties are worth elaboration; although some will only arise when working with unevenly sampled dependent variables, others will have to be addressed by any class-level aggregation approach.

First, there are the problems related to dependent variables that have only been collected from a subset of the population – in this study, fewer than half of the students in most courses responded to the course rating survey. As there may be considerable variation between students' interaction patterns in the same class, aggregations may represent behavior from respondents, nonrespondents, or a mixture of both. As response

rates shrink, this problem becomes more acute; aggregations are less likely to represent respondents. The gap between respondents and other students is also affected by how different the respondents are from the others, which may be quantifiable.

A clustering approach could offer one measurement of variation within classes without having identifiable respondents. Students within one class could be randomly assigned to groups. After features or covariates are aggregated, a dataset could be built with groups from multiple classes. If a clustering algorithm were run on that dataset, the algorithm's clusters can be compared to the actual classes; the algorithm's separation indices could then provide a quantifiable estimate of within-group and between-group variation. The process could be repeated an arbitrary number of times with different group assignments. If the clustering algorithm is able to reproduce the original classes repeatedly, regardless of group assignment, that would suggest that the students in the classes are quite similar to each other – there is less within-group variation – and the data may be more suitable for aggregation. Although the goal is different, this is similar to external validation of clustering (Dalmaijer et al., 2020; Ritter, 2014; Ullmann et al., 2021).

There also may be broader issues related to class-level predictions in educational data mining. Well-defined ground truth labels are rarer at the class level than at the student level, particularly labels that are more than just aggregations of student-level measurements.[6] This is particularly true in higher education, where course ratings are

[6] Other aggregations of ratings were considered as potential outcome variables in this study, but ultimately discarded due to having insufficient variation.

much more popular than all other class-level metrics. Course ratings exist somewhere between class-level and student-level; they are nearly always aggregated for analysis, but students are asked about their individual experiences in a course (Boysen, 2015; Kitto et al., 2019; Miller & Seldin, 2014). Aggregating student-level measurements, as in this study, is a typical approach, but may oversimplify the class – a class that half the students love and half hate is not the same as one that all students are indifferent towards.

Further research is needed to better understand the class-level aggregation issues encountered in this study. They may only apply to course rating data, or be artifacts of either the feature aggregation or the aggregation of the dependent variables. Class- and institution-level effects can account for a good deal of variance in multilevel models but are rarely discussed in learning analytics or educational data mining research. Knowing more about how aggregation affects independent and dependent variables could improve both modeling and prediction of nested data.

Teaching and learning researchers in higher education tend to support the use of course ratings with other data as part of an evaluation portfolio (Paulsen, 2002; Seldin, 1999). Seldin (1999) notes that 40% of liberal arts colleges report using data from classroom observations to evaluate teaching performance, with Ronald Berk (2005, 2013) and Ackerman, Gross, and Vigneron (2009) favoring their use (note that this is somewhat distinct from formative peer observations, see Fletcher, 2018 for more). Considering the success of learning analytics researchers in predicting student affect using ground truth labels from classroom observations (e.g., Baker et al., 2020; Botelho et al., 2018), it may be interesting to attempt to predict ratings from classroom observations. The variability in

rating approaches would make validation difficult, but peer observation practices have been studied in higher education (see citations in Berk, 2013), with Gleason and Cofer (2014) describing the validation of an observation protocol for undergraduate math classes.

Apart from predictive models, there may be a simpler way to offer early feedback to instructors and administrators. The factor analyses for this study found a great deal of correlation between the questions on the rating survey and even between the two separate scales. Given this, it may be worthwhile to ask students to respond to single-question rating surveys earlier in the term.[7] While the results from these surveys would not be as trustworthy as the results from the entire rating survey, they should be sufficient to serve as a warning system. Gathering data from single-question surveys multiple times over a term would also show more about the trajectory of a course, as well. Finally, as single-question surveys are quicker and easier to answer, they might see higher response rates.

Teaching, especially in higher education, is a complicated process, both in how instructors act and in how their actions affect students. There is not one prototype of a good class; what works for one group of students, at one university, in one term may be dramatically different from what works for a different group at a different school at a different time. The wide range of successful courses makes teaching tricky to evaluate, regardless of whether it is evaluated with ratings surveys, peer observations, student learning gains, or other methods. Even with perfectly reliable evaluations, relationships

[7] This is similar to, but simpler than, the midterm rating strategies discussed by e.g., Blash et al., 2018; Taylor et al., 2020.

between students' actions and the evaluations are, as demonstrated here, very difficult to detect.

In this study, it was hoped that connecting course ratings to classroom activity could improve the understanding and increase the credibility of the course rating process. In addition, early identification of courses that are likely to receive lower ratings or fewer responses could help the ratings process become more formative. Although the models built here were ultimately insufficiently predictive, providing formative feedback to college instructors would still be very valuable, as would further understanding of what makes successful classes successful. Teaching is one of the most important functions of higher education institutions, but more research is needed to help instructors do the best jobs they can.

APPENDIX A: FEATURE LIST

Individual features	Class-level aggregation			Week-level aggregation				
	Mean	Median	S.D.	Min	Max	Mean	Median	S.D.
Early sessions	X					X		X
Morning sessions	X					X		X
Afternoon sessions	X					X		X
Evening sessions	X					X		X
Number of active events	X	X	X			X		X
Total time on Moodle	X	X	X	X	X	X		X
Average session length	X	X	X			X		X
Number of sessions with 1+ active events		X	X				X	X
Time spent on course forum		X	X			X		X
Longest time between Moodle sessions		X				X		X
Total sessions		X	X			X		X

Course-level features	Min	Max	Mean	Median	S.D.
Total sessions			X		X
Longest time between Moodle sessions			X		X
Total time on Moodle	X	X	X		X
Time spent on course forum			X		X
Time spent on assignment pages			X		X
Time spent viewing resources			X		X
Total sessions			X		X
Total students					
Students this week					

Percent of students who visited this week					X
Percent of students who took actions this week					
Assignments this week*		X	X		X
Late turnins this week*		X	X		X
On-time turnins this week*		X	X		X
Average grade on assignments			X		X
Cumulative average grade					
Number of active sessions				X	X
Early sessions			X		X
Morning sessions			X		X
Afternoon sessions			X		X
Evening sessions			X		X
Average session length			X		X

Table 7. List of features with their course-level and week-level aggregation.

Course-level features are already at the course level and thus only need week-level aggregation. Course-level features have asterisks if their means and standard deviations were computed only including weeks with at least one assignment.

BIBLIOGRAPHY

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., & Hess, C. W. (1990). Satisfaction with processes of collecting student opinions about instruction: The student perspective. *Journal of Educational Psychology*, *82*(2), 201–206. https://doi.org/10.1037/0022-0663.82.2.201

Ackerman, D., Gross, B. L., & Vigneron, F. (2009). Peer observation reports and student evaluations of teaching: Who are the experts? *The Alberta Journal of Educational Research*, *55*(1), 18–39.

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, *53*(5), 576–591. https://doi.org/10.1007/s11162-011-9240-5

Altman, N., & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, *15*(6), 399–400. https://doi.org/10.1038/s41592-018-0019-x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Andersen, L. P. (2018). How to increase teachers' user of their quantitative evaluation data: Case: The technical faculty of Lund University. Master's thesis: Aalborg Universitetsforlag. https://www.ucviden.dk/en/publications/how-to-increase-teachers-use-of-their-quantitative-evaluation-dat

Anthony, E. (2019). (Blended) learning: How traditional best teaching practices impact blended-learning classrooms. *Journal of Online Learning Research*, *5*(1), 23-48.

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, 267–270. https://doi.org/10.1145/2330601.2330666

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., & Arslan Esme, A. (2019). Investigating the Impact of a Real-time, Multimodal Student Engagement Analytics Technology in Authentic Classrooms. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. https://doi.org/10.1145/3290605.3300534

Babić, I. Đ. (2017). Machine learning methods in predicting the student academic motivation. *Croatian Operational Research Review*, *8*(2), 443–461. https://doi.org/10.17535/crorr.2017.0028

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112–118.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. In *International Educational Data Mining Society*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED560553

Baker, R. S., Ocumpaugh, J., & Andres, J. M. L. (2020). BROMP Quantitative Field Observations: A Review. In R. Feldman (Ed.), *Learning Science: Theory, Research, and Practice* (pp. 127–156). McGraw-Hill.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Bentley, P. J., & Kyvik, S. (2012). Academic work from a comparative perspective: A survey of faculty working time across 13 countries. *Higher Education*, *63*(4), 529–547. https://doi.org/10.1007/s10734-011-9457-4

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In *Higher Education: Handbook of Theory and Research* (pp. 279–326). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-8005-6_7

Benton, S. L., & Ryalls, K. R. (2016). Challenging Misconceptions about Student Ratings of Instruction. IDEA Paper# 58. *IDEA Center, Inc.*

Berk, Richard A. (2016). *Statistical Learning from a Regression Perspective*. Springer International Publishing. https://doi.org/10.1007/978-3-319-44048-4

Berk, Ronald A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48–62.

Berk, Ronald A. (2013). Top five flashpoints in the assessment of teaching effectiveness. *Medical Teacher*, *35*(1), 15–26. https://doi.org/10.3109/0142159X.2012.732247

Berk, Ronald A. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *The Journal of Faculty Development*, *32*(1), 73-81.

Blash, A., Schneller, B., Hunt, J., Michaels, N., & Thorndike, J. (2018). There's got to be a better way! Introducing faculty to mid-course formative reviews as a constructive tool for growth and development. *Currents in Pharmacy Teaching and Learning*, *10*(9), 1228–1236. https://doi.org/10.1016/j.cptl.2018.06.015

Borch, I., Sandvoll, R., & Risør, T. (2020). Discrepancies in purposes of student course evaluations: What does it mean to be "satisfied"? *Educational Assessment, Evaluation and Accountability*, *32*(1), 83–102. https://doi.org/10.1007/s11092-020-09315-x

Botelho, A., Baker, R. S., Ocumpaugh, J., & Heffernan, N. T. (2018). Studying affect dynamics and chronometry using sensor-free detectors. *Proceedings of the 11th International Conference on Educational Data Mining*, 157–166.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology*, *42*(2), 109–118. https://doi.org/10.1177/0098628315569922

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education*, *39*(6), 641–656. https://doi.org/10.1080/02602938.2013.860950

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71–88. https://doi.org/10.1016/j.econedurev.2014.04.002

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance.* ERIC.

Braun, E., & Leidner, B. (2009). Academic course evaluation: Theoretical and empirical distinctions between self-rated gain in competences and satisfaction with teaching behavior. *European Psychologist*, *14*(4), 297–306. https://doi.org/10.1027/1016-9040.14.4.297

Breiman, L. (Ed.). (1998). *Classification and regression trees* (Repr). Chapman & Hall [u.a.].

Brinko, K. T. (1993). The Practice of Giving Feedback to Improve Teaching. *The Journal of Higher Education*, *64*(5), 574–593. https://doi.org/10.1080/00221546.1993.11778449

Brown, M. J. (2008). Student perceptions of teaching evaluations. *Journal of Instructional Psychology*, *35*(2), 177–181.

Buschetto Macarini, L. A., Cechinel, C., Batista Machado, M. F., Faria Culmant Ramos, V., & Munoz, R. (2019). Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Applied Sciences*, *9*(24), 5523. https://doi.org/10.3390/app9245523

Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. Oxford University Press.

Capa-Aydin, Y. (2016). Student evaluation of instruction: Comparison between in-class and online methods. *Assessment & Evaluation in Higher Education*, *41*(1), 112–126. https://doi.org/10.1080/02602938.2014.987106

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. Face-to-face) using a multilevel growth modeling approach. *Computers & Education*, *53*(2), 429–435. https://doi.org/10.1016/j.compedu.2009.03.001

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, *9*(2), 137–151. https://doi.org/10.1016/j.jarmac.2019.12.009

Carrell, S. E., & West, J. E. (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, *118*(3), 409–432. https://doi.org/10.1086/653808

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited. IDEA Paper No. 32*. https://eric.ed.gov/?id=ED402338

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. Jossey-Bass Inc.

Centra, J. A. (1998). The development of the student instructional report II. *Princeton, NJ: Educational Testing Service*.

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. https://doi.org/10.1023/A:1025492407752

Centra, J. A. (2015). Student evaluations of instruction: Research evidence and their utility. *Journal of Collective Bargaining in the Academy*, *10*, 49.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42–54. https://doi.org/10.1016/j.compedu.2016.02.006

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, 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. https://doi.org/10.1080/02602930301683

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual likert-type items. *Journal of Agricultural Education*, *35*(4), 31–35.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16–30. https://doi.org/10.1177/0273475308324086

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. *Journal of Education for Business*, *88*(1), 26–35. https://doi.org/10.1080/08832323.2011.633580

Clow, D. (2012). *The learning analytics cycle: Closing the loop effectively*. 134. https://doi.org/10.1145/2330601.2330636

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. https://doi.org/10.3102/00346543051003281

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis, 2nd ed* (pp. xii, 430). Lawrence Erlbaum Associates, Inc.

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Cunningham, P. (2008). Dimension Reduction. In M. Cord (Ed.), *Machine learning techniques for multimedia: Case studies on organization and retrieval* (pp. 91–112). Springer. http://site.ebrary.com/id/10223660Darwin, S. (2017). What contemporary work are student ratings actually doing in higher education? *Studies in Educational Evaluation*, *54*, 13–21. https://doi.org/10.1016/j.stueduc.2016.08.002

Darwin, S. (2021). The changing topography of student evaluation in higher education: Mapping the contemporary terrain. *Higher Education Research & Development*, *40*(2), 220–233. https://doi.org/10.1080/07294360.2020.1740183

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 446–455. https://doi.org/10.1145/3303772.3303784

Denson, N., Loveday, T., & Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction? *Higher Education Research & Development*, *29*(4), 339–356. https://doi.org/10.1080/07294360903394466

Donovan, J., Mader, C. E., & Shinsky, J. (2010). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning*, *9*(3), 283–296.

Drysdale, M. J. (2010). *Psychometric properties of postsecondary students' course evaluations* [Ph.D., Utah State University]. https://search.proquest.com/docview/817554206/abstract/5460C2164DEF44BEPQ/1

Edström, K. (2008). Doing course evaluation as if learning matters most. *Higher Education Research & Development*, *27*(2), 95–106. https://doi.org/10.1080/07294360701805234

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.1080/00031305.1983.10483087

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37–46. https://doi.org/10.1108/09684880310462074

Erikson, M., Erikson, M. G., & Punzi, E. (2016). Student responses to a reflexive course evaluation. *Reflective Practice*, *17*(6), 663–675. https://doi.org/10.1080/14623943.2016.1206877

Feistauer, D., & Richter, T. (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation*, *59*, 168–178. https://doi.org/10.1016/j.stueduc.2018.07.009

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, *30*(2), 137–194. https://doi.org/10.1007/BF00992716

Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., Rienties, B., Ullmann, T., & Vuorikari, R. (2016). *Research evidence on the use of learning analytics: Implications for education policy*.

Fischer, E., & Hänze, M. (2019). Bias hypotheses under scrutiny: Investigating the validity of student assessment of university teaching by means of external observer ratings. *Assessment & Evaluation in Higher Education*, *44*(5), 772–786. https://doi.org/10.1080/02602938.2018.1535647

Fletcher, J. A. (2018). Peer Observation of Teaching: A Practical Tool in Higher Education. *The Journal of Faculty Development*, *32*(1), 51–64.

Flodén, J. (2017). The impact of student feedback on teaching in higher education. *Assessment & Evaluation in Higher Education*, *42*(7), 1054–1068. https://doi.org/10.1080/02602938.2016.1224997

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x

Gee, N. (2017). A study of student completion strategies in a Likert-type course evaluation survey. *Journal of Further and Higher Education*, *41*(3), 340–350. https://doi.org/10.1080/0309877X.2015.1100717

Gleason, J., & Cofer, L. D. (2014). Mathematics classroom observation protocol for practices: Results in undergraduate mathematics classrooms. In T. Fukawa-Connelly, G. Karakok, K. Keene, & M. Zandieh (Eds.), *Proceedings of the 17th Annual Conference on Research in Undergraduate Mathematics Education, Volume 1 (Refereed Articles)* (pp. 93–103).

Golding, C., & Adam, L. (2016). Evaluate to improve: Useful approaches to student evaluation. *Assessment & Evaluation in Higher Education*, *41*(1), 1–14. https://doi.org/10.1080/02602938.2014.976810

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341–364. https://doi.org/10.1007/s11162-016-9429-8

Gowda, S. M., Baker, R. S., Corbett, A. T., & Rossi, L. M. (2013). Towards automatically detecting whether student learning is shallow. *International Journal of Artificial Intelligence in Education*, *23*(1–4), 50–70. https://doi.org/10.1007/s40593-013-0006-4

Granzin, K. L., & Painter, J. J. (1973). A new explanation for students' course evaluation tendencies. *American Educational Research Journal*, *10*(2), 115–124. https://doi.org/10.3102/00028312010002115

Greller, W., Ebner, M., & Schön, M. (2014). Learning analytics: From theory to practice – data support for learning and teaching. *Computer Assisted Assessment. Research into E-Assessment*, 79–87. https://doi.org/10.1007/978-3-319-08657-6_8

Gweon, G., Jun, S., Lee, J., Finger, S., & Rosé, C. P. (2011). A framework for assessment of student project groups on-line and off-line. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing Interactions in CSCL*. Springer US. https://doi.org/10.1007/978-1-4419-7710-6

Hammonds, F., Mariano, G. J., Ammons, G., & Chambers, S. (2017). Student evaluations of teaching: Improving teaching quality in higher education. *Perspectives: Policy and Practice in Higher Education*, *21*(1), 26–33. https://doi.org/10.1080/13603108.2016.1227388

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research*

in Higher Education, *45*(5), 497–527. https://doi.org/10.1023/B:RIHE.0000032326.00426.d5

Harrison, C. J., Könings, K. D., Schuwirth, L., Wass, V., & van der Vleuten, C. (2015). Barriers to the uptake and use of feedback in the context of summative assessment. *Advances in Health Sciences Education*, *20*(1), 229–245. https://doi.org/10.1007/s10459-014-9524-6

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*(3), 311–323. https://doi.org/10.1023/B:RIHE.0000019592.78752.da

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131. https://doi.org/10.3102/00346543071001105

Haskell, R. E. (1998). *Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century.* https://eric.ed.gov/?id=ED426114

Hativa, N. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences. *Research in Higher Education*, *37*(3), 341–365.

Hativa, N. (2000). *Teaching for effective learning in higher education*. Kluwer Academic.

Hillman, D. C. A., Willis, D. J., & Gunawardena, C. N. (1994). Learner-interface interaction in distance education: An extension of contemporary models and strategies for practitioners. *American Journal of Distance Education*, *8*(2), 30–42. https://doi.org/10.1080/08923649409526853

Hirschberg, J., & Lye, J. (2016). The influence of student experiences on post-graduation surveys. *Assessment & Evaluation in Higher Education*, *41*(2), 265–285. https://doi.org/10.1080/02602938.2014.1001318

Hmieleski, K., & Champagne, M. V. (2000). Plugging in to course evaluation. *The Technology Source*.

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. *Assessment & Evaluation in Higher Education*, *44*(3), 361–378. https://doi.org/10.1080/02602938.2018.1511969

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

Howard, G. S., & Schmeck, R. R. (1979). Relationship of changes in student motivation to student evaluations of instruction. *Research in Higher Education*, *10*(4), 305-315.

Ishiyama, J., Balarezo, C., & Miles, T. (2014). Do Graduate Student Teacher Training Courses Affect Placement Rates? *Journal of Political Science Education*, *10*(3), 273–283. https://doi.org/10.1080/15512169.2014.924801

Islam, N. (2018). A novel framework using machine learning to effectively analyze the faculty evaluations. *Journal of Education & Social Sciences*, *6*(2), 40–52. https://doi.org/10.20547/jess0621806204

Jiang, Y. H., Javaad, S. S., & Golab, L. (2016). Data mining of undergraduate course evaluations. *Informatics in Education*, *15*(1), 85–102. https://doi.org/10.15388/infedu.2016.05

Joksimović, S., Gašević, D., Loughin, T. M., Kovanović, V., & Hatala, M. (2015). Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education*, *87*, 204–217. https://doi.org/10.1016/j.compedu.2015.07.002

Jones, J., Gaffney-Rhys, R., & Jones, E. (2014). Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education*, *38*(1), 37–56. https://doi.org/10.1080/0309877X.2012.699514

Kay, J., Koprinska, I., & Yacef, K. (2010). Educational data mining to support group work in software development projects. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. Baker (Eds.), *Handbook of Educational Data Mining*. CRC Press.

Kay, J., Maisonneuve, N., Yacef, K., & Zaïane, O. (2006). Mining patterns of events in students' teamwork data. *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, 45–52.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*, *47*(7), 451–464. https://doi.org/10.3102/0013189X18785613

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

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

Kite, M. E., Subedi, P. C., & Bryant-Lees, K. B. (2015). Students' Perceptions of the Teaching Evaluation Process. *Teaching of Psychology*, *42*(4), 307–314. https://doi.org/10.1177/0098628315603062

Kitto, K., Williams, C., & Alderman, L. (2019). Beyond Average: Contemporary statistical techniques for analysing student evaluations of teaching. *Assessment & Evaluation in Higher Education*, *44*(3), 338–360. https://doi.org/10.1080/02602938.2018.1506909

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. https://doi.org/10.1080/03075079912331380128

Kontogiannis, S., Valsamidis, S., Kazanidis, I., & Karakos, A. (2014). Course Opinion Mining Methodology for Knowledge Discovery, Based on Web Social Media. *Proceedings of the 18th Panhellenic Conference on Informatics*, 50:1-50:6. https://doi.org/10.1145/2645791.2645827

Krull, E., & Leijen, Ä. (2015). Perspectives for Defining Student Teacher Performance-Based Teaching Skills Indicators to Provide Formative Feedback through Learning Analytics. *Creative Education*, *06*(10), 914. https://doi.org/10.4236/ce.2015.610093

Law, K. M. Y., Geng, S., & Li, T. (2019). Student enrollment, motivation and learning performance in a blended learning environment: The mediating effects of social, teaching, and cognitive presence. *Computers & Education*, *136*, 1–12. https://doi.org/10.1016/j.compedu.2019.02.021

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology*, *82*(2), 232–241. https://doi.org/10.1037/0022-0663.82.2.232

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Love, D. A., & Kotchen, M. J. (2010). Grades, Course Evaluations, and Academic Incentives. *Eastern Economic Journal*, *36*(2), 151–163. https://doi.org/10.1057/eej.2009.6

Lutovac, S., Kaasila, R., Komulainen, J., & Maikkola, M. (2017). University lecturers' emotional responses to and coping with student feedback: A Finnish case study. *European Journal of Psychology of Education*, *32*(2), 235–250. https://doi.org/10.1007/s10212-016-0301-1

Macfadyen, L. P., Dawson, S., Pardo, A., & Gaševic, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, *9*, 17–28.

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. https://doi.org/10.1080/02602938.2015.1044421

Makhlouf, J., & Mine, T. (2020, November). Automatic feedback models to students freely written comments. In *28th International Conference on Computers in Education, ICCE 2020* (pp. 336-341). Asia-Pacific Society for Computers in Education.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(1), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319–383). Springer, Dordrecht. https://doi.org/10.1007/1-4020-5742-3_9

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness. *The Journal of Higher Education*, *64*(1), 1–18. https://doi.org/10.1080/00221546.1993.11778406

Marsh, H. W., Ginns, P., Morin, A. J. S., Nagengast, B., & Martin, A. J. (2011). Use of student ratings to benchmark universities: Multilevel modeling of responses to the Australian Course Experience Questionnaire (CEQ). *Journal of Educational Psychology*, *103*(3), 733–748. https://doi.org/10.1037/a0024221

Marsh, H. W., & Hocevar, D. (1991a). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education*, *7*(1), 9–18. https://doi.org/10.1016/0742-051X(91)90054-S

Marsh, H. W., & Hocevar, D. (1991b). 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. https://doi.org/10.1016/0742-051X(91)90001-6

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.

Martinez, R., Wallace, J. R., Kay, J., & Yacef, K. (2011). Modelling and Identifying Collaborative Situations in a Collocated Multi-display Groupware Setting. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (Vol. 6738, pp. 196–204). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_27

Martinez-Maldonado, R., Yacef, K., & Kay, J. (2013). Data mining in the classroom: Discovering groups strategies at a multi-tabletop environment. *Proceedings of the 6th International Conference on Educational Data Mining*.

Matos-Díaz, H. (2012). Student evaluation of teaching, formulation of grade expectations, and instructor choice: Explorations with random-effects ordered probability models. *Eastern Economic Journal*, *38*(3), 296–309. https://doi.org/10.1057/eej.2011.7

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. *New Directions for Teaching and Learning*, *2003*(96), 39–48. https://doi.org/10.1002/tl.121

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, *52*(11), 1218–1225. https://doi.org/10.1037/0003-066X.52.11.1218

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation. *Academe*, *100*(3), 35–38.

Mircioiu, C., & Atkinson, J. (2017). A comparison of parametric and non-parametric methods applied to a Likert scale. *Pharmacy*, *5*(2), 26. https://doi.org/10.3390/pharmacy5020026

Monllaó Olivé, D., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A Quest for a One-Size-Fits-All Neural Network: Early Prediction of Students at Risk in Online Courses. *IEEE Transactions on Learning Technologies*, *12*(2), 171–183. https://doi.org/10.1109/TLT.2019.2911068

Moore, M. G. (1989). Editorial: Three types of interaction. *American Journal of Distance Education*, *3*(2), 1–7. https://doi.org/10.1080/08923648909526659

Morris, L. V., Finnegan, C., & Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221–231. https://doi.org/10.1016/j.iheduc.2005.06.009

Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019). The validity and utility of activity logs as a measure of student engagement. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 300–309. https://doi.org/10.1145/3303772.3303789

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Norton, A., Sonnemann, J., & Cherastidtham, I. (2013). *Taking university teaching seriously*. Grattan Institute Melbourne.

Novak, E., & Johnson, T. E. (2012). Assessment of students' emotions in game-based learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives* (pp. 379–399). Springer. https://doi.org/10.1007/978-1-4614-3546-4_19

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314. https://doi.org/10.1080/02602930701293231

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49–64.

Pardo, A. (2014). Designing learning analytics experiences. In J. A. Larusson & B. White (Eds.), *Learning Analytics* (pp. 15–38). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_2

Pardo, A. (2018). A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education*, *43*(3), 428–438. https://doi.org/10.1080/02602938.2017.1356905

Park, E., & Dooris, J. (2019). Predicting student evaluations of teaching using decision tree analysis. *Assessment & Evaluation in Higher Education*, *45*(5), 1–18. https://doi.org/10.1080/02602938.2019.1697798

Pascarella, E. T., Seifert, T. A., & Whitt, E. J. (2008). Effective instruction and college student persistence: Some new evidence. *New Directions for Teaching and Learning*, *2008*(115), 55–70. https://doi.org/10.1002/tl.325

Paulsen, M. B. (2002). Evaluating teaching performance. *New Directions for Institutional Research*, *2002*(114), 5–18. https://doi.org/10.1002/ir.42

Piccinin, S., Cristi, C., & McCoy, M. (1999). The impact of individual consultation on student ratings of teaching. *International Journal for Academic Development*, *4*(2), 75–88. https://doi.org/10.1080/1360144990040202

Quick, J., Motz, B., Israel, J., & Kaetzel, J. (2020). What college students say, and what they do: Aligning self-regulated learning theory with behavioral logs. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 534–543. https://doi.org/10.1145/3375462.3375516

Robinson, T. E., & Hope, W. C. (2013). Teaching in Higher Education: Is There a Need for Training in Pedagogy in Graduate Degree Programs? *Research in Higher Education Journal*, *21*. https://eric.ed.gov/?id=EJ1064657

Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science*, *28*(5), 439–468. https://doi.org/10.1023/A:1026576404113

Rojas, I. G., & García, R. M. C. (2012). Towards efficient provision of feedback supported by learning analytics. *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 599–603. https://doi.org/10.1109/ICALT.2012.171

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384. https://doi.org/10.1016/j.compedu.2007.05.016

Sao Pedro, M. A., Baker, R. S. J. d., & Gobert, J. D. (2012). Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User*

Modeling, Adaptation, and Personalization (pp. 249–260). Springer. https://doi.org/10.1007/978-3-642-31454-4_21

Sao Pedro, M. A., de Baker, R. S. J., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, *23*(1), 1–39. https://doi.org/10.1007/s11257-011-9101-0

Sauer, T. M. (2012). *Predictors of student course evaluations* [Ph.D., University of Louisville]. https://search.proquest.com/docview/1151735427/abstract/CF7243DEB3594F34PQ/1

SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Seldin, P. (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*. Anker Publishing.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shah, M., Cheng, M., & Fitzgerald, R. (2017). Closing the loop on student feedback: The case of Australian and Scottish universities. *Higher Education*, *74*(1), 115–129. https://doi.org/10.1007/s10734-016-0032-x

Sheehan, E. P., & DuPrey, T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology; Milwaukee, Wis.*, *26*(3), 188–193.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin*, *106*(1), 148.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, *27*(5), 397–409. https://doi.org/10.1080/0260293022000009285

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

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

Talukdar, J., Aspland, T., & Datta, P. (2013). Australian higher education and the course experience questionnaire: Insights, implications and recommendations. *Australian Universities' Review*, *55*(1), 27–35.

Taylor, R. L., Knorr, K., Ogrodnik, M., & Sinclair, P. (2020). Seven principles for good practice in midterm student feedback. *International Journal for Academic Development*, *25*(4), 350–362. https://doi.org/10.1080/1360144X.2020.1762086

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2017). Towards actionable learning analytics using dispositions. *IEEE Transactions on Learning Technologies*, *10*(1), 6–16. https://doi.org/10.1109/TLT.2017.2662679

Tetenbaum, T. (1977). The factor invariance of student ratings of instruction under three sets of directions. *Research in Higher Education*, *6*(1), 11–23. https://doi.org/10.1007/BF00992013

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*(2), 197–208. https://doi.org/10.1177/0013164496056002001

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Tomes, T., Coetzee, S., & Schmulian, A. (2019). Prediction-Based Student Evaluations of Teaching as an Alternative to Traditional Opinion-Based Evaluations. *Assessment & Evaluation in Higher Education*, *0*(0), 1–15. https://doi.org/10.1080/02602938.2019.1594157

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Tucker, B., Jones, S., Straker, L., & Cole, J. (2003). Course evaluation on the web: Facilitating student and teacher reflection to improve learning. *New Directions for Teaching and Learning*, *2003*(96), 81–93. https://doi.org/10.1002/tl.125

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

Valencia Acuña, E. A. (2017). *Response Styles in Student Evaluation of Teaching* [Thesis, University of Toronto]. https://tspace.library.utoronto.ca/handle/1807/80897

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, *52*(4), 1003–1017. https://doi.org/10.1177/0013164492052004025

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A clustering methodology of web log data for learning management systems. *Journal of Educational Technology & Society*, *15*(2), 154–167.van der Lans, R. M. (2018). On the "association between two things": The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(4), 347–366. https://doi.org/10.1007/s11092-018-9285-5

Veeck, A., O'Reilly, K., MacMillan, A., & Yu, H. (2016). The use of collaborative midterm student evaluations to provide actionable results. *Journal of Marketing Education*, *38*(3), 157–169. https://doi.org/10.1177/0273475315619652

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

Wang, G., & Williamson, A. (2020). Course evaluation scores: Valid measures for teaching effectiveness or rewards for lenient grading? *Teaching in Higher Education*, 1–22. https://doi.org/10.1080/13562517.2020.1722992

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 D. M. C. S. II, D. L. D. Yore, & D. B. Hand (Eds.), *Quality Research in Literacy and Science Education* (pp. 383–398). Springer Netherlands. http://link.springer.com/chapter/10.1007/978-1-4020-8427-0_19

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, *40*(3), 227–261. https://doi.org/10.3200/JECE.40.3.227-261

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

Wujek, B., Hall, P., & Güneș, F. (2016). *Best Practices for Machine Learning Applications*. SAS Institute. https://support.sas.com/resources/papers/proceedings16/SAS2360-2016.pdf

Xu, Y. (2012). Developing a comprehensive teaching evaluation system for foundation courses with enhanced validity and reliability. *Educational Technology Research and Development*, *60*(5), 821–837. https://doi.org/10.1007/s11423-012-9240-y

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, *29*, 23–30. https://doi.org/10.1016/j.iheduc.2015.11.003

Young, R. D. (1993). Student evaluation of faculty: A faculty perspective. *Perspectives on Political Science*, *22*(1), 12–16. https://doi.org/10.1080/10457097.1993.9944513

Yu, J., Huang, C., Han, Z., He, T., & Li, M. (2020). Investigating the influence of interaction on learning persistence in online settings: Moderation or mediation of academic emotions? *International Journal of Environmental Research and Public Health*, *17*(7), 2320. https://doi.org/10.3390/ijerph17072320

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