ASSESSING THE EFFECTIVENESS OF PERSONALIZED COMPUTER-ADMINSTERED FEEDBACK IN AN INTRODUCTORY BIOLOGY COURSE

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

In Partial Fulfilment of the Requirements

For the degree of Master of Education

In the Department of Educational Psychology and Special Education

University of Saskatchewan

Saskatoon

By

MATTHEW SCHMIDT

© Copyright Matthew Schmidt, June 2019. All rights reserved.

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of Department of Educational Psychology and Special Education
28 Campus Drive Room 3104
University of Saskatchewan
Saskatoon, Saskatchewan S7N 0X1
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

ABSTRACT

Though often held in high regard as a pedagogical tool, the role of feedback within the learning process remains poorly understood (Shute, 2008). The prevailing feedback literature reveals a history of inconsistent if not contradictory findings (Kluger & DeNisi, 1996). This already complicated state is made worse by the recent introduction of learning analytic tools capable of providing students with ongoing personalized computer-generated feedback; the effectiveness of which remains unknown. The present study contributed to this new domain of knowledge by evaluating one such circumstance where a learning analytic feedback intervention was implemented in an introductory biology course at the University of Saskatchewan. The system provided personalized feedback to half of the enrolled students differentiated according their individual characteristics. The remaining students received generic feedback that was common to all students within the condition. The effectiveness of personalized feedback was evaluated with respect to academic achievement (i.e., final grade) and feedback satisfaction. Results of the treatment effect analyses showed no significant differences in student academic achievement but a small significant difference in feedback satisfaction. Follow-up analyses revealed that these significant differences in feedback satisfaction were not consistent from one iteration of the course to the next and that mean feedback satisfaction was in steady decline since the system's implementation. It is suspected that the lack of improvement in academic achievement pertained to poor adherence of the system with the theoretical underpinnings of good feedback practice. Limitations of the study and future directions are discussed.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor, Dr. Amin Mousavi, for his encouragement and guidance. I had not imagined that so much work would have ultimately been conducted on this project. I would like to express my appreciation to my thesis examination committee, Dr. Dirk Morrison and Dr. Vicki Squires for their thoughtful questions and suggestions. I would also like to thank Dr. Tim Claypool for taking time out of his summer to chair the examination.

I would like to thank the Learning Analytics Cluster for their welcoming hand in inviting me to the project. Through the conference presentations and speaking events I have learned far more than I thought possible and feel well-prepared to take on new professional challenges because of it.

I want to acknowledge the enormous amount of gratitude I have for my family for their love and support in this pursuit. Finally, I want to thank my partner Amanda for her unrelenting love, patience, and positivity.

TABLE OF CONTENTS

PERMISSION TO USE ... i

ABSTRACT ... ii

ACKNOWLEDGMENTS ... iii

TABLE OF CONTENTS ... iv

LIST OF TABLES ... vii

LIST OF FIGURES ... viii

Chapter One: Introduction ... 1

Purpose and Significance ... 2

Parameters ... 5

Assumptions ... 5

Delimitations ... 6

Limitations ... 7

Definitions ... 8

The Researcher ... 9

Chapter Two: Literature Review ... 10

Theoretical Frameworks for the Operation of Feedback ... 10

Research on Personalized Feedback ... 26

The Advent of Learning Analytics ... 30

Summary ... 43

Chapter Three: Methodology 45

Introduction 45

Data 45

Measures 49

Statistical Analysis 51

Chapter 4 Results 55

Overview 55

Data Cleaning Procedures 55

Research Question 1 56

Research Question 2 60

Summary 65

Chapter 5 Discussion 67

Overview 67

Research Question 1 68

Research Question 2 74

Limitations and Future Research Directions 78

Conclusion 80

References 82

Appendix A: Feedback Satisfaction Survey 95

Appendix B: Estimating Treatment Effects 96

Appendix C: Sample Differentiated Feedback Message 1 101

Appendix D: Sample Differentiated Feedback Message 2 102

LIST OF TABLES

Table 3.1. Age and gender demographics of classes September 2016 through January 2018 46

Table 4.1. Means, standard deviations, and correlations with confidence intervals for final grade 56

Table 4.2. Means and standard deviations for feedback conditions 57

Table 4.3. Regression results for student final grade 58

Table 4.4. Treatment effect estimates for final grade 59

Table 4.5. Means and standard deviations for feedback conditions 60

Table 4.6. Means, standard deviations, and correlations with confidence intervals for feedback satisfaction 61

Table 4.7. Treatment effect estimates for feedback satisfaction 62

Table 4.8. Feedback satisfaction means and standard deviations for feedback conditions 63

Table 4.9. Descriptive statistics for feedback satisfaction as a function of course offering 64

Table 4.10. ANOVA results using feedback satisfaction as the dependent variable 65

Table 4.11. Tukey multiple comparisons following one-way ANOVA of course offering 65

LIST OF FIGURES

Figure 2.1. Hierarchy of Processing for Feedback Intervention Theory. 17

Figure 2.2. Framework for decision-making in computer-based instruction............................... 19

Figure 2.3. Factors contributing to the informative value of feedback....................................... 23

Figure 2.4. Model of conceptual influences on feedback recipience.. 26

Figure 3.1. Personalized feedback message according to lab exam grade. 49

Figure 3.2. Sample questions from the feedback satisfaction survey. 51

Figure 4.1. Plot of satisfaction means by course year.. 64

Chapter One: Introduction

Feedback over the past ten years has changed dramatically for many at institutions of higher education (Kane, Sandretto, & Heath, 2002). Historically, feedback was believed to have operated in terms of strong behaviorism – feedback thereby taking the role of a reinforcer of correct behavior (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy, 1977; Skinner, 1968). Over time this position changed, and theories attempting at explaining feedback's operation gradually incorporated additional facets, including task difficulty (Mason & Bruning, 2001), locus of attention (Kluger & DeNisi, 1996), and retrieval strategies (Bangert-Drowns et al., 1991). Despite these modifications, feedback in its daily operation remained largely unchanged. Minor changes were observed as computer-based instruction increased in prevalence and many of these existing feedbacks became programmable, but these changes were neither exceptional nor extraordinary. In more recent years however, there has been an expansion of online and blended classrooms (Means, Toyama, Murphy, & Baki, 2013). Computer integration of the education system has continued to increase, carrying with it the instructional tool of feedback.

In a traditional classroom, where there exists a reasonable instructor to student ratio, teaching continues to operate as it has in the past. The instructor can interact personally with individual students and cater their feedback to them based upon their personal characteristics. However, in a typical introductory university lecture-based course this is no longer the case. There simply is not the time of day for a handful of instructors to deliver person-to-person feedback when the instructor to student ratio can be 1 to 300 or more. In response to this recent development, learning analytics has burst onto the scene, now utilizing banks of student information to provide personalized feedback in these courses (Bodily & Verbert, 2017; Siemens

& Long, 2011). Whether these new systems are providing feedback as beneficial as that of an in-person instructor has yet to be determined; more importantly though, whether these systems are creating a net gain in learning outcomes is yet unresolved (Gašević, Dawson, Rogers, & Gasevic, 2016; Teasley, 2017; Viberg, Hatakka, Bälter, & Mavroudi, 2018).

Purpose and Significance

There are many studies looking at the effects of feedback on learning achievement given different task and question characteristics (see for example: Bangert-Drowns et al., 1991; Clariana, Wagner, & Murphy, 2000; Morrison, Ross, Gopalakrishnan, & Casey, 1995; Pridemore & Klein, 1991, 1995; Van-Dijk & Kluger, 2001). In recent years there have also been many studies looking at the effects of highly personalized computer-generated feedback in the context of learning analytics (Tempelaar, Rienties, & Giesbers, 2015). Despite this, there remains a great degree of uncertainty regarding the potential effectiveness of personalized computer-generated feedback. Select recent reviews cast a great deal of doubt on the real-world effectiveness of many of these systems in terms of their ability to improve student learning outcomes, and other notable scholars in the field have produced work questioning the direction in which the field is heading (Gašević et al., 2016; Gašević, Dawson, & Siemens, 2015; Macfadyen & Dawson, 2012). With that said, there remains a rationale and certainly a potential that increasingly personalizing feedback for individual learners should improve both their receptiveness to the feedback and improve learning achievement outcomes as well.

The first major purpose for the study is to learn how to respond to the increasing pervasiveness of technology in education at both the classroom and curriculum level. The rate at which the education system is being technologically integrated is unprecedented, and the increased prevalence of fully electronic or blended classrooms will no doubt require a more settled understanding about how to deliver personalized feedback. An increasing number of online courses are being offered (Jones & Blankenship, 2017). One of the key advantages of online classes are their ability to scale with increasing large class sizes. Despite this, the effects of large class sizes in online courses remain poorly understood (Bettinger, Doss, Loeb, Rogers, & Taylor, 2017). What is understood however, is the difficulty that would be faced by a single instructor that hoped to provide personalized feedback to every student within their course.

Second, this study hopes to add to the new and fast-developing field of learning analytics and its role in education. Learning analytics made its appearance in western education systems around 2010 and has since expanded rapidly (Siemens, 2013). In a modern context where decisions are routinely being made with the assistance of big data, it is no surprise that this movement can also be seen within institutions of higher education. Data-driven decisions have been shown to improve organizational productivity and output over traditional forms of decision making (Brynjolfsson, Hitt, & Kim, 2011) and this increasing prevalence of big data can be seen in numerous other institutions ranging from health (Lee & Yoon, 2017), to insurance (Corea, 2017), to business (Banica & Hagiu, 2015). There has long been a call for reform in education with an emphasis towards increasing the quality of instruction (Slavin, 2017), though how best to manage such a reform or what foundational framework to utilize when navigating these “new waters” remains unknown. According to Siemens and Long (2011) the new means by which educational reform will take place has arrived, and that is learning and academic analytics utilizing data to inform and guide educational decision-making at all levels.

In addition to the proliferation of computing technology to society, the internet age has afforded us a near endless wealth of data to draw upon for learning related decision-making that would otherwise vanish (Ferguson, 2012). For example, emails, tweets, and activity data with online courses can be recorded and mined for any number of purposes. Prior to the internet age conversations in the classroom or lectures were immediately lost shortly after their effects had faded from memory. When guided by theory and analyzed, these electronic data can provide valuable insight into the learning process and inform both educators and administrators of how to proceed with students (Siemens, 2012; Siemens & Baker, 2012). This integration of analytics has already changed the way learning and teaching take place in many higher education classrooms. Examples of in-house learning analytic tools that have been developed and implemented range from Purdue University to the University of Michigan, and even to the University of Saskatchewan (Arnold & Pistilli, 2012; Greer et al., 2015; Wright, McKay, Hershock, Miller, & Tritz, 2014). Among many other applications, learning analytics can be used to generate and deliver students' personalized feedback based on a multitude of different student characteristics and in-course activities — many of which have already been accomplished (Kuo, Peng, & Chang, 2014; Ochoa et al., 2018; Tempelaar, Heck, Cuypers, van der Kooij, & van de Vrie, 2013). Outcomes of the current study hope to provide additional clarity in the appropriate selection of student attributes, activities, and demographics for use in future feedback systems.

The third purpose of the present study is its desire to contribute to both the existing understanding of feedback as a construct and to computer-based instruction in higher education. As a construct, feedback remains relatively poorly understood; a cursory review of the literature will reveal decades of inconsistent findings with numerous attempts at developing a model capable of accurately predicting previously unobserved feedback effects (Shute, 2008). Reviews showing that some feedback interventions produce either no measurable increase in learning outcomes or have a detrimental effect can be found (Azevedo & Bernard, 1995; Kluger & DeNisi, 1996), leading some authors to suggest feedback is a "double-edged sword" and should be used with caution (Kluger & DeNisi, 1998). Many of these earlier attempts focussed heavily on the task and type of feedback but acknowledged minimally the role that individual learner characteristics might play in the effectiveness of feedback (Bangert-Drowns et al., 1991; Kulhavy, 1977). Fortunately, more recent proposals have provided frameworks that incorporate these important learner qualities (Narciss & Huth, 2004; Winstone, Nash, Parker, & Rowntree, 2017) offering a basis for their inclusion. It is the hope that the present study contributes in a meaningful way to the existing feedback literature, while simultaneously incorporating the newly developed techniques and approaches present within learning analytics.

In summary, this paper aims to address these two primary research questions:

1) Will students receiving personalized feedback outperform those receiving generalized feedback with respect to their final grade?
2) Will students receiving personalized feedback report higher levels of feedback satisfaction than those receiving generalized feedback?

Parameters

This study is guided by the following parameters.

Assumptions

The research methodology was fundamentally of quantitative design. Matching procedures permitting the estimation of a treatment effect of our experiment were utilized. It was assumed that the selection of analyses chosen would be best positioned to provide the most in-depth understanding of the variables within the study because of their ability to control for potential bias in group assignment. The results of this study rest upon the following assumptions:

- Participation in the study was voluntary; students were adequately informed of the experiment prior to consenting. Students were also given the opportunity to withdraw their participation at any time.
- Participants answered and engaged with the study's survey honestly and gave the questions genuine effort in their completion.
- It was assumed that no errors were made during the data collection, entry, and analysis. Data were largely collected through the university's data warehouse or generated through student activity and recorded on Blackboard. Data was handled by professional university staff before it was given to the current researcher.
- Measures chosen to assess desired variables are believed to have done so completely and accurately. Evidence of the measures' reliability and validity will be provided.

Delimitations

The first delimitation is the setting and context of the experiment. The study takes place exclusively at the University of Saskatchewan, and thus might limit the generalizability to similar western Canadian universities near this time in history. Though the student ages and backgrounds within the study ranged widely, the overwhelming majority of the students were western Canadians in their first year of university. This population was selected for its convenience in multiple areas. First, the personalized feedback system was not the product of the present paper's author, but rather was already in operation at the University of Saskatchewan for the past several years. Secondly, though the sample is confined largely to western Canadian first year students, the results of the study are still expected to have some applicability in other traditionally western educational contexts that are highly similar.

The second delimitation pertains to the work on intelligent tutoring systems. Though these systems have been in existence for a long time and have demonstrated well their ability to deliver personalized computer-generated feedback, they will not be extensively reviewed within the project for the following reasons. First, feedback exists on a continuum ranging from the most minimal unit of information that an instructor can relay to a student to simply re-teaching a lesson. In large part, intelligent tutoring systems exist on the far end of that continuum and are more akin to teachers in their application than simply feedback systems. Second, one of the fundamental short comings in the research on learning analytic systems is the apparent lack of adherence to established educational theory (Gašević et al., 2015). An expanded discussion of intelligent tutoring systems would be an aside to this more central theme within the paper.

The third delimitation is the decision to utilize quantitative methods exclusively within the data analysis. The analyses chosen for the proposed study have been selected on the basis of their ability to control for selection bias in group assignment. Further, the primary outcome of interest in the present study is academic achievement, a variable that is best operationalized numerically. Lastly, the proposed study intends on using substantial sample sizes. With these challenges in mind, it has been decided to exclusively pursue a quantitative means of analysis.

Limitations

A key limitation is the choice of methods of analysis. Various matching protocols have been chosen for data analysis. These procedures attempt to emulate the circumstances of a randomized controlled experiment when those conditions could not be fully realized. In the case of the present study, stipulations by the ethical review board meant that these ideal experimental conditions could not be achieved. Specifically, random assignment to treatment and control groups could not be ensured. These procedures, however, are not a perfect substitute for a randomized experiment. They merely allow one to estimate the effect of a treatment. Therefore, establishing direct causal inference is not possible within this study.

Within the family of matching statistics is the limitation of a possible lack of data. Matching requires data from known pre-treatment covariates of the outcome variable to ensure that students can be matched appropriately. The proposed study intended on matching with known covariates for which data are available, it is not known whether these data will be contained within the provided datasets.

In addition to the limits on causal inference are the limitations of quantitative analysis more broadly. Teaching and learning can be a highly personal affair, one full of nuance and subtlety (Stronge, Ward, & Grant, 2011). The quantitative methods employed by this study depend upon standardized surveys with closed-ended questions. There is therefore a chance that some information will be unavailable by conducting the study from a purely quantitative perspective with concisely operationalized definitions.

Another limitation is the potential lack of relevant data. Entrance and exit surveys, and other data collection forms might have varied from one term or year to the next with respect to the present study. Though each year will have substantial overlap in terms of the questions asked, some variations will exist and may limit both the generalizability of the results and the analyses that may be performed.

Definitions

Feedback: the process where information is communicated to a learner with the intention of modifying the learner’s behavior to improve learning (Shute, 2008).

Personalized Feedback: the process where individualized information (for example: student demographic characteristics, past performance indices, and learning traits/characteristics), according to the person's own characteristics or actions, is communicated to them with the intention of modifying their behavior for improving learning.

Intelligent Tutoring Systems: computer programs that model learners' psychological states to provide individualized instruction (Ma, Adesope, Nesbit, & Liu, 2014).

Learning Analytics: the measurement, collection, analysis and reporting of data about learners and their contexts, for the purpose of understanding and optimising learning and the environments in which it occurs (First International Conference on Learning Analytics and Knowledge, 2011).

Note: throughout the paper the terms personalized feedback, individualized feedback, and differentiated feedback may be used interchangeably.

The Researcher

The principal researcher for the proposed study is a former high school teacher who completed a Bachelor of Arts degree in Psychology and Bachelor of Education degree from the University of Saskatchewan. The present thesis was conducted for the purposes of fulfilling the requirements of completing a master's program in Educational Psychology, specializing in Measurement and Evaluation.

Chapter Two: Literature Review

This chapter provides a conceptual framework for the study by identifying the relevant literature on feedback's concepts, theories, and usage within the study's context. The review will begin with a brief historical overview of instructional feedback and its interpretation throughout the earlier half of the 20th century. Following the historical context, an overview of formative theoretical frameworks for feedback's operation will be discussed —demonstrating the gradual trend within such proposed frameworks to recognize the importance of individual learner characteristics and their role in the effectiveness of feedback. The chapter will then shift towards a discussion of the disseminated research comparing the effectiveness of "personalized" compared to "generic" feedback. Within this section the effectiveness of personalized feedback will be reviewed both before and after the advent of learning analytics. Finally, the chapter will finish with a summary of the presented work and identify the primary findings dictating the direction of the present paper.

Theoretical Frameworks for the Operation of Feedback

Kulhavy's (1977) review marks the first major critique on the prevailing ideas around instructional feedback theory and practice. Drawing research from multiple lines of evidence, Kulhavy (1977) systematically deconstructed the popular belief that feedback operated as a "reinforcer", acting to strengthen the learner's behavior in a specific direction. Kulhavy's review does not provide a single theoretical framework by which feedback operates; rather, the review serves as a paradigm shift in the understanding of feedback research. Just as the body of feedback literature is full of inconsistencies and contradictions in the present (Shute, 2008), it was so at the time of Kulhavy's (1977) review. It was Kulhavy's position that much of these inconsistencies stemmed from the false belief that feedback operated as the behaviorists believed. At the time of Kulhavy's review conventional belief on the topic of feedback was that it functioned as a form of operant conditioning, where student behavior was governed by either reinforcers or punishers. The basic idea can be illustrated thusly: If a student makes an error on an assignment or test and is immediately informed of its incorrectness, that feedback 'punishes' the student and thus reduces the likelihood that he/she will make the same mistake again. Likewise, a student that has correctly answered a question and is subsequently informed of its correctness will be 'reinforced' in that action and more likely to answer similar questions correctly. Despite being written nearly a decade following Kulhavy's (1977) review, one can observe this attitude towards the reinforcement model of feedback in Shuell's (1986) overview of the cognitive conceptions of learning.

Kulhavy's (1977) review devotes a considerable portion to dispense with the notion that feedback acts as a reinforcing mechanism. The first limitation cited by Kulhavy is the dissimilarity between the motivations of students and animals, that contributed heavily to the development of conditioning theory. Kulhavy (1977) illustrated the difference with reference to a food deprived lab animal. As is the case with many operant conditioning experiments of the time, a pigeon or rat may be deprived of nourishment and then given access to an apparatus that when interacted with correctly results in the animal receiving a food pellet. Drawing on the animal's innate desire to eat, behaviors that result in the animal receiving food will be both reinforced and will be more likely to be emitted under similar conditions. According to Kulhavy (1977) there was no reason to believe that this model will operate in the same way with respect to students engaging in a classroom. Students in these conditions are seldom if ever required to perform under such deprived states and the idea that students' hunger for knowledge in the same way that pigeons' hunger for food is comical if not absurd.

Sorting through the existing literature of the time Kulhavy (1977) also noted that the reinforcement model fails to explain a whole host of discrepancies. Should the reinforcement position be correct, one ought to observe that verifying a student's answer as correct should reinforce that behavior in the student and result in greater learning than simply informing the student of their answer being incorrect. Citing his colleagues' study (Anderson, Kulhavy, & Andre, 1971), Kulhavy (1977) demonstrated that this is not the case. In the study students were sorted into two feedback conditions: feedback for only correctly answered questions and feedback for only incorrectly answered questions. Students were then required to learn a 112-frame computer-controlled lesson. Results showed that both groups made approximately the same number of errors during instruction and had similar post-test scores. The well-documented Delayed-Retention-Effect also challenges the reinforcement theory (Markowitz & Renner, 1966; Sturges, 1972). Reinforcement theory would suggest that retention is best when feedback is delivered immediately following the student's input; however, many studies have shown that delaying the feedback can improve a student's score on a test of retention under some circumstances (More, 1969; Schmidt, Young, Swinnen, & Shapiro, 1989; Schroth, 1992; Sturges, 1972). Lastly, the use of external rewards does not reliably produce increases in student performance (Sullivan, Baker, & Schutz, 1967), a finding that would not be predicted by reinforcement theory. To conclude, Kulhavy (1977) suggested that if one were to objectively assess the state of feedback research they would have to "accept the fact, that whatever feedback does, it rarely acts as a functional reinforcer with text-based materials" (p. 216). This key takeaway put forth by Kulhavy (1977) set the stage for the inevitable decline in the belief of feedback as a reinforcer, making way for new theoretical frameworks.

As it applies to the idea of personalized feedback for individual learners, Kulhavy's (1977) paper is mostly lacking. Near the end of his review Kulhavy touches on the idea that some learner characteristics may be an important subject to study with respect to feedback, although he confines this to the idea of response confidence. The rationale was that varying levels of confidence in a student's answers will produce varying levels of interest in the feedback that may follow. In an earlier study (Kulhavy, Yekovich, & Dyer, 1976) Kulhavy and his colleagues found evidence for their predictions related to response confidence. Specifically, they found that when students had a high degree of confidence in their response to a question, they were far more likely to recall that question if provided feedback. This suggests that at a minimum learner expectations and feedback require further study. Unfortunately, this and a request for future research to increasingly address learner characteristics in the field of instructional psychology is the final statement within the paper pertaining to differentiating feedback by learner characteristics.

One of the more contemporary attempts to synthesize and develop a framework by which to understand the effectiveness of feedback comes from Bangert-Drowns et al. (1991). The authors reviewed 58 effect sizes from 40 reports comparing the effectiveness of feedback. Only studies where a group provided feedback on their coursework was compared in terms of academic achievement to a control group that received no feedback but whose instruction was otherwise identical were included. In keeping with Kulhavy's (1977) finding, despite the commonplace belief that feedback is necessary and absolutely beneficial in instruction, approximately one third of the papers reviewed did not result in improved learning outcomes (Bangert-Drowns, et al., 1991).

In the process of their review the authors (Bangert-Drowns et al., 1991) emphasized a number of features regarding both learning theory and its relationship with feedback. The primary takeaway is that feedback is a process of mutual influence between learners and their environments. Learners cannot learn something new in complete isolation; feedback is the means by which the learner checks their current state of achievement with that of the established standard. In addition to their framework, the authors made the following broader conclusions: First, pre-search availability was negatively related to feedback effectiveness. That is, students that were able to access their corrective feedback before having to generate their own answer were less likely to learn from feedback. Second, feedback type was strongly related to learning outcomes. For example, verification feedback (simply indicating to the student whether their answer was incorrect or correct) produced weaker learning effects than correct response feedback (showing the student the correct response) (Bangert-Drowns et a., 1991). Differing forms of feedback were more appropriate depending on the task. Third, pretests significantly lowered the effect sizes of feedback studies. It is proposed that pretests shape learners' expectations and review content such that feedback effects may be diminished. Lastly, different types of instruction warrant different types of feedback; varying levels of feedback effectiveness can be seen depending on the mode of instruction (e.g. computer-based vs. in-person).

The authors (Bangert-Drowns et al., 1991) proposed the following five-stage framework for properly interpreting feedback in its operation:

1. Initial state. The initial state refers to the learner's individual characteristics (level of motivation, interest, self-efficacy, goals, etc.) and any relevant pre-existing knowledge that they may have.

2. Search and retrieval strategies. These refer to the processes the learner engages in upon encountering a question or problem. These processes may include mindful reflection of the task at hand, exploring one's memory, or evaluating a proposition. This is a crucial stage in ensuring the effectiveness of feedback. The authors found that circumventing this step by permitting feedback before allowing students to thoughtfully respond resulted in minimal learning.
3. Response. Given the situation and the available information, the student responds to the task.
4. Evaluation. The student receives feedback for their response and must evaluate such feedback. Prior to the findings presented within Bangert-Drowns et al.'s (1991) study, response certitude was believed to have played a significant role here. It was proposed that when students were shown to be incorrect but were highly certain that their answers were correct, they would spend greater time reviewing the provided feedback. Though some support exists for this position, the association appears weak and other studies have failed to find such an effect.
5. Adjustments. Having received feedback, students now make any number of meta-cognitive adjustments to ensure improved subsequent performance. Some of these strategies may include increased time spent studying or working on new memorization strategies. Such adjustments may also include more sophisticated overhauls of fallacious beliefs held by the student that led to poor performance. The adjustment stage sets the stage for the student's next initial state.

Referring to the above framework one can observe how a learner cycles through the stages from the initial state to the final adjustment stage and then repeats that process. Each sequence provides the learner with new information by which they can re-evaluate their performance and increase competence. Though the framework acknowledges individual learner characteristics within its initial state, the paper largely ignores the notion that feedback might be made more effective by differentiating it according to these characteristics or activities, opting instead to view feedback as a unitary construct applied equally across all learners.

In another attempt to sort out contradictory findings, Kluger and DeNisi (1996) proposed a new theoretical framework for understanding the role of feedback interventions on performance. As previously mentioned, it was largely presupposed that feedback operated in accordance with the behavioristic law of effect put forth by early behaviorists (Thorndike, 1927; Watson, 1913), though as Kluger and DeNisi (1996) pointed out, this theory is woefully unable to explain the fact that nearly a third of feedback interventions either fail to produce an increase in learning, or worse, result in reduced learning. The way in which feedback correctly operates is almost certainly more complex and nuanced than the behaviorists of the past proposed (Kluger & DeNisis, 1996; Kulhavy, 1977). For their analysis, Kluger and DeNisi defined a feedback intervention as the action of an external agent to provide performance information to a learner. Operating under this definition, the authors systematically reviewed approximately 3000 papers removing those that did not possess a control group, did not measure feedback performance, or used samples smaller than 10 participants. The final review consisted of 131 papers with 607 effect sizes. Their analyses identified a number of feedback moderators such as receiving written vs verbal feedback, praise, task complexity, timing, and task type (physical, knowledge, memory etc.).

In their proposed Feedback Intervention Theory (FIT) (Kluger & DeNisi, 1996), learner attention is examined at three hierarchically arranged levels. The lowest and most basic of these levels are the task-learning processes. These pertain to the essential qualities of any given task at hand and the practical elements of the feedback that may follow. On top of the hierarchy are the meta-task processes. Such processes link the task at hand with higher-order goals pertaining to the evaluation of self. The meta-task processes include processes that have considerable control over effects on performance such as attention to the self, framing effects, and affect. Situated in the middle of the theory are the task-motivation processes. These serve as an intermediary between the task-learning and meta-task processes linking the two through motivation. The basic idea behind the proposed theory is that all things being equal, feedback that focuses the learner's attention towards the task-learning processes is most likely to result in increased learning. Feedback that shifts the attention of the learner up the hierarchy takes attention away from the task at hand and increasingly focuses it on the self, resulting in reduced learning (Kluger & DeNisi, 1996).

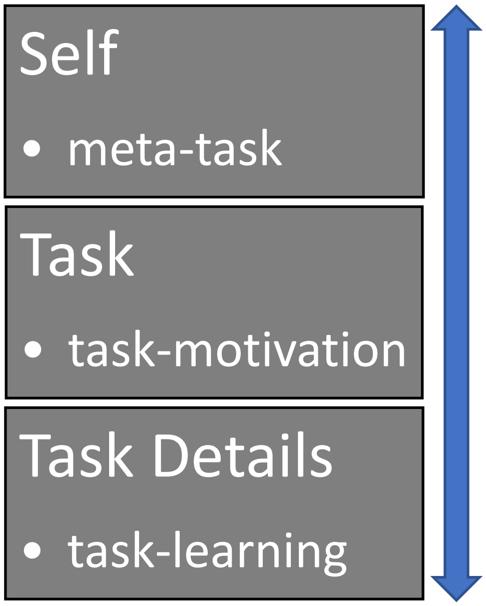

Figure 2.1. Hierarchy of Processing for Feedback Intervention Theory (Kluger & DeNisi, 1996).

FIT is based upon the following five foundational consecutive arguments, each of which is interdependent and is built upon the preceding argument (Kluger & DeNisi, 1996).

1. Behavior is regulated by comparisons of feedback to goals or standards.
2. Goals or standards are organized hierarchically.
3. Attention is limited and therefore only feedback-standard gaps (achievement gaps between a learner's actual performance and that of the feedback) that receive attention actively participate in behavior regulation.
4. Attention is normally directed to a moderate level of the hierarchy.
5. FIs change the locus of attention and therefore affect behavior.

Unlike previous theories that looked at feedback with reference to the law of effect, FIT explicitly examines the learner's locus of attention as they are provided with feedback. In contrast to Bangert-Drowns et al. (1991), Kluger and DeNisi's (1996) FIT deals exclusively with explaining the effectiveness of feedback and is less focused on the broader project of relating learning to feedback. Thus, FIT may be thought of as a supplemental theory to other theories that more fully incorporate other aspects of the learning process.

Kluger and DeNisi's (1996) FIT makes no mention of differentiating feedback according to learner characteristics. However, the theory serves as a powerful tool in explaining why certain forms of feedback may inadvertently move the locus of attention up the hierarchy and threaten the learner's self concept. Among other feedback theories, FIT stands out in its ability to make sense of negative findings when evaluating feedback research, and will no doubt be useful in examining some of the research later discussed in this review.

Citing a dramatic increase in the use of computers for educational purposes and the potential for computer-based instruction to one day provide personalized instruction to all individuals within a classroom, Mason and Bruning (2001) noted that an appropriate theoretical framework for use in such scenarios was much needed. Surveying past feedback literature pertaining to student achievement, task level, feedback timing, prior student knowledge, and feedback type, Mason and Bruning developed a decision-making tree for the administration of feedback in computer-based instruction.

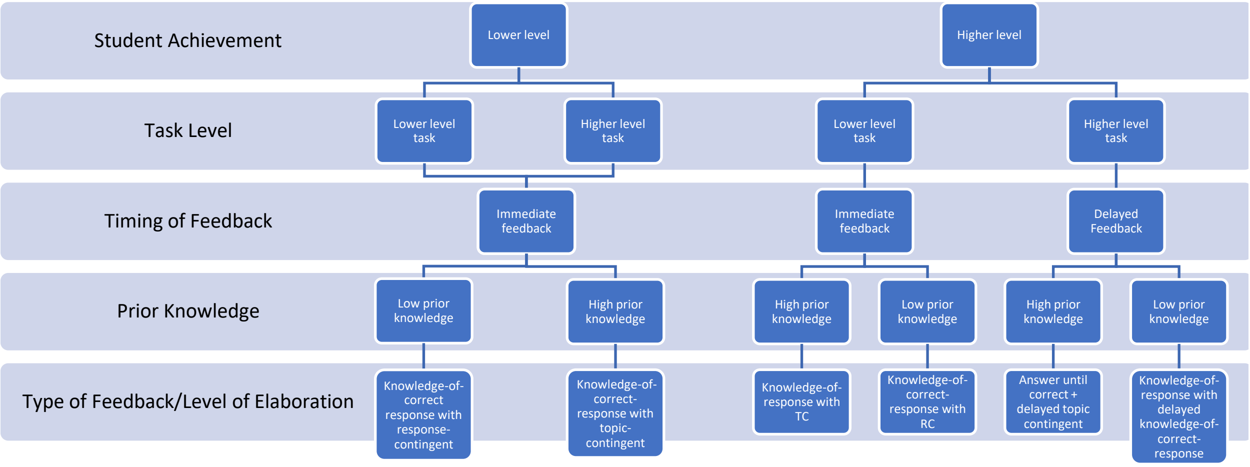

Figure 2.2. Framework for decision-making in computer-based instruction (Mason & Bruning, 2001).

To understand the decision tree and correctly prepare the appropriate feedback for a given learner one must first refer to the top of the tree and determine whether the student in question is high or low achieving in the predetermined content area. High achieving students are those with a greater level of expertise in the chosen area while low achieving students likely possess a less accurate understanding of the given material. Following the classification of student achievement, one must determine the nature of the task and estimate its level of difficulty to time feedback accordingly. Unsurprisingly, task level is contingent upon the goal of instruction. Mason and Bruning (2001) suggested that simple concept acquisition or memorization of new words would constitute a lower level task, and most would likely benefit from immediate feedback to correct simple errors for both high and low achieving students. On the other hand, an assignment requiring abstract reasoning, or the application of theory would constitute a higher-level task and would warrant delayed feedback for high achieving students, but immediate feedback for low achievers. Finally, after determining the timing, one assesses the student's prior level of knowledge on the subject in question. The level of knowledge is then used to identify an appropriate feedback type for use.

This framework (Mason & Bruning, 2001) marks one of the earlier guides by which feedback in instruction is differentiated by student qualities. Though the process of differentiation is confined to the characteristics of student prior ability and level of knowledge in the subject, this is a considerable departure from the theories preceding it.

As has been made plain, the proposed theories for the use and operation of feedback have largely neglected the individual characteristics of the student. Mason and Bruning (2001) incorporated these in a softer sense (for example, considering student achievement and prior knowledge as important variables) but could have been more explicit in mentioning the individual differences between learners that might further influence how feedback takes effect. Narciss and Huth (2004) proposed a model for the creation of effective tutoring feedback. This view of tutoring feedback emphasizes the importance of students becoming self regulated and utilizing feedback to navigate the learning process successfully. Compared to previous formulations of feedback frameworks, their multidimensional view of tutoring feedback explicitly recognizes the importance of catering feedback to individual students based on their characteristics. Unlike Mason and Bruning's (2001) decision tree, Naricss and Huth (2004) proposed a relationship model that relates instructional and individual factors towards the more central tenets of feedback.

Narciss and Huth's (2004) framework for tutoring feedback is centered around three primary factors:

1) The nature and quality of feedback. Within the nature and quality of feedback are three separate facets.
    a. Functions of Feedback. These pertain to the functional aspects of the feedback related to instructional goals. For example, an instructor may intend that the provided feedback improve a student's ability to recall definitions, or on the other hand, the feedback may be intended to improve a student's ability to stay motivated and on-task.
    b. Presentation of Feedback. As discussed earlier by other theorists, these pertain to the technical qualities of the feedback (timing, type, level of elaboration, frequency, etc.)
    c. Contents of Feedback. This facet contains two subcomponents. The first is the evaluative component that shows the learner the gap between their performance and the ideal performance. The second is the information

component. This contains information relating the task-at-hand and should allow improved performance for the student in subsequent evaluations.

2) The characteristics of the instructional context. Contained within this facet are the learning objectives, tasks, and obstacles inherent in the instructional context. Feedback can only be of value to the student if its qualities are carefully matched to the context for which it is to be used.

3) The individual learner characteristics. A great many learner qualities could be included within this section. Narciss and Huth (2004) drew attention to learning objectives and strategies, student prior knowledge, as well as motivation and meta-cognitive skills. It is further proposed that many of the inconsistent findings in feedback research may have been caused in part by the individual characteristics of learners in those studies. Thus, the authors stressed the importance of taking these qualities into account and considered them of equal importance to the other facets when fitting an appropriate feedback strategy for an individual.

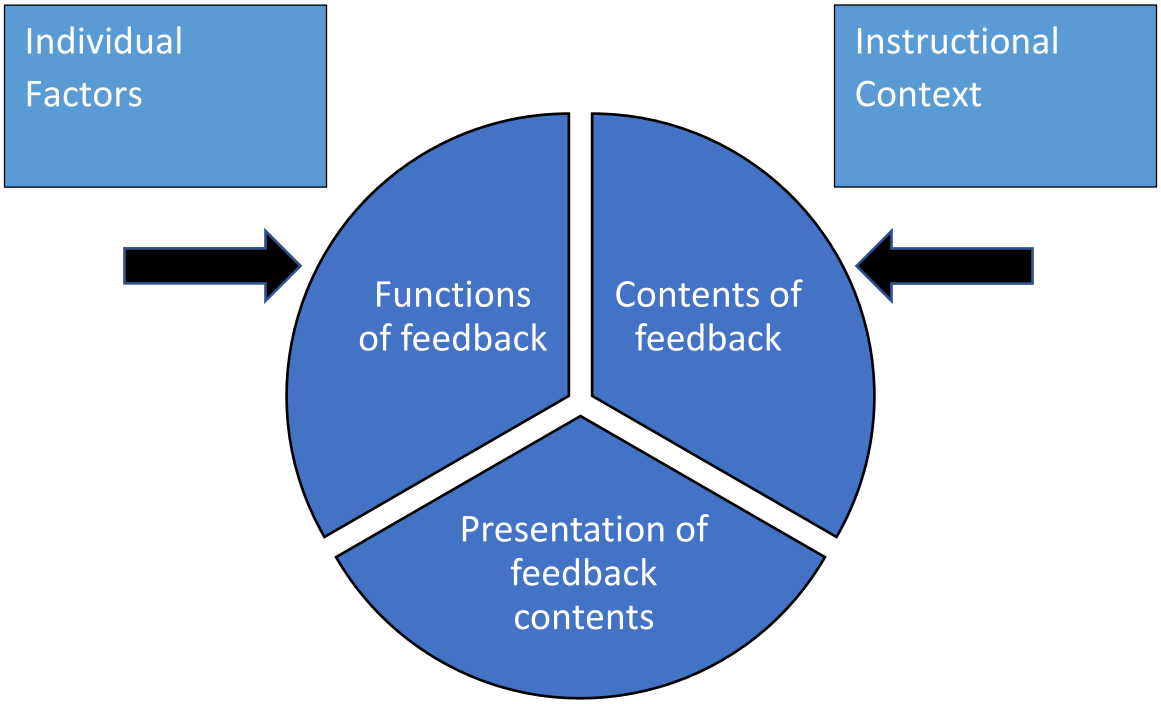

Figure 2.3. Factors contributing to the informative value of feedback (Narciss & Huth, 2004).

Narciss and Huth's (2004) model and guide for developing tutoring feedback is a major step in the direction of incorporating learner characteristics into the application of differing feedback strategies. Many past analyses had concluded that individual learner characteristics were significant factors in the inconsistent findings on feedback studies (Azevedo & Bernard, 1995; Bangert-Drowns et al., 1991; Kulhavy, 1977; Mory, 1992). The paper also calls for the careful adaptation of function, content, and form of feedback for individual learners, in effect personalizing feedback to best maximize learning outcomes. Narciss and Huth (2004) claimed that matching feedback to these learner characteristics is of equal importance to matching the situational factors to the instructional context.

Even though the focus of feedback literature has progressed in moving away from the behaviorist interpretation and the law of effect (Thorndike, 1933), it has still largely remained focussed on the specific features of feedback itself rather than other outside factors (Shute, 2008). Recently a movement towards understanding feedback from the perspective of the receivers and what factors affect successful reception and implementation of feedback has taken place. In a systematic review by Winstone, et al. (2017), the authors cited the long-observed phenomenon of inconsistent effect of feedback that others have observed. Contrary to other reviews however, they suggest that an understudied determinant of this trend is the effectiveness and degree of engagement learners have with their feedback. Asserting that for feedback to be effective the recipient must be properly engaged with the feedback is not a new idea (Hattie & Timperley, 2007; Nicol & MacFarlane-Dick, 2006). However, this branch of feedback literature has largely been disconnected and too far removed from the more mainstream lines of feedback research. Thus, the authors performed a meta analysis of available relevant works and proposed the following framework for the effective utilization of feedback.

The first term that needs to be addressed is that of “proactive recipience.” Proactive recipience is the active engagement with the feedback process and is believed to be associated with a host of favourable outcomes pertaining to feedback utilization (Winstone et al., 2017). Passive recipience, by contrast, is the state where learners engage with feedback in a disinterested and compliant manner. The authors then propose the SAGE (Self appraisal, Assessment literacy, Goal-setting and self regulation, and Engagement and motivation) taxonomy, a set of distinct recipience processes that aim to increase proactive recipience in learners (Winstone et al., 2017). Self appraisal refers to the learner’s ability to make accurate and informative judgements about their own strengths, weaknesses, and reduce their reliance on outside assessment. Assessment literacy refers to the student’s ability to understand the scheme by which they will be graded. Understanding this system affords them a metric to evaluate their own work and performance. Goal setting and self regulation refer to the learner's ability to articulate well-defined goals (for example, achieve greater than 85% on next math assignment) and maintain focus in pursuit of these goals. Lastly, engagement and motivation refer to the learner having an orientation of motivation. To make use of and actively engage with feedback the learner must express a desire to understand and act upon their feedback. The other key aspect of the proposed framework is the role of interpersonal communication with regards to feedback. This component refers to the characteristics of the receiver, the sender, the message, and the learner context. The final formulation of the framework can be seen below:

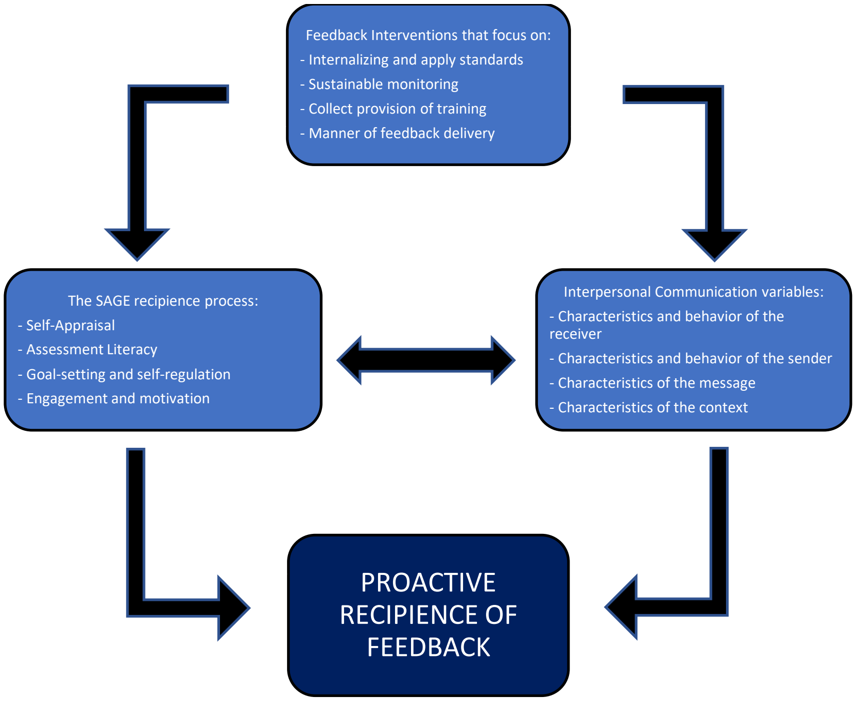
Feedback Interventions that focus on:
- Internalizing and apply standards
- Sustainable monitoring
- Collect provision of training
- Manner of feedback delivery
The SAGE recipience process:
- Self-Appraisal
- Assessment Literacy
- Goal-setting and self-regulation
- Engagement and motivation
Interpersonal Communication variables:
- Characteristics and behavior of the receiver
- Characteristics and behavior of the sender
- Characteristics of the message
- Characteristics of the context
PROACTIVE RECIPIENCE OF FEEDBACK

Figure 2.4. Model of conceptual influences on feedback recipience (Winstone et al., 2017).

As one can see from Figure 4, the relationship between the proactive recipience of the learner and the feedback intervention is moderated by both the SAGE recipience process and by interpersonal communication variables. Thus, according to the theory, for a feedback intervention to enjoy a greater degree of success, conditions within both the SAGE taxonomy and communication factors must be properly addressed (Winstone et al., 2017). In terms of addressing personalized feedback, this model provides an added layer of analysis for consideration, offering a deeper look at what underlying variables may need to be addressed when attempting to personalize instructive feedback.

Research on Personalized Feedback

This section of the review focuses on personalized feedback research prior to the implementation of learning analytics. From the perspective of many within the field of education, feedback is often seen as one of the most powerful instructional tools available to them and its efficacy appears to have no limit (Cohen, 1985; Planar & Moya, 2016). Personalized instructor feedback to students is one such instructional strategy that is not only touted as the highest form of feedback but is also generally preferred by students (Burr, Brodier, & Wilkinson, 2013; Cramp, 2011; Laryea, 2013; Lipnevich & Smith, 2009). Unfortunately, providing personalized instructor feedback is often impossible with many modern university courses where the professor to student ratio becomes increasingly small (i.e. 1/10 compared to 1/300). Given this limitation it seems inevitable that if instructors in higher education want to administer personalized feedback to their students they will need to adapt technologically. In considering ways to deliver personalized feedback more easily and to a greater volume of students, a number of different methods have been proposed or investigated. For example, audio recordings have been shown to provide more complex and detailed feedback than text alone and afford the student the opportunity to infer greater meaning from the speaker's tone and intonation (Gould & Day, 2013; Merry & Orsmond, 2008). Others have proposed the use of video feedback, where students are delivered a video of either their instructor or a teaching assistant marking their tests and assignments in real time and providing ongoing commentary of their marking (West & Turner, 2016). Despite the creativity with the aforementioned modes of delivery, prior success, and ability to reach large numbers of students electronically, they still suffer from a lack of scalability. In a course with over 300 students, no instructor can reasonably be expected to deliver individual feedback to their students whatever the medium. While these examples are no doubt interesting and likely to be of value for other more distantly related projects on personalized feedback, they are less relevant to the present study as personalized written feedback is most pertinent.

Another major contributor to the personalized feedback literature comes from the broader work related to intelligent tutoring systems (ITS). ITS are computer programs that model numerous psychological states within a learner and use these to provide individualized instruction and feedback (Ma et al., 2014). According to Ma et al. (2014), ITS are those that are capable of performing tutoring functions (presenting new information for students, asking questions from students, providing feedback or hints to activities, answering student questions, or offering prompts to elicit a change in student cognition or meta cognition), can develop a model to represent the student's psychological states (subject knowledge, learning strategies, motivation levels, emotions), and can adapt both these qualities should the need arise. The effectiveness of these systems in producing increases in learning outcomes can be readily demonstrated. Ma et al.'s (2014) meta analysis found that ITS instruction produced significantly greater achievement compared to large group teacher-led instruction, non-ITS computer-based tutoring, and workbook instruction. However, it was also found that in terms of academic achievement ITS were not significantly different from those of individualized human tutoring. Another meta analysis performed by Steenbergen-Hu and Cooper (2014) found similar results, the only difference being that individualized human tutors were slightly more effective than the ITS. Possessing their own well-defined place within the learning sciences, ITS are an extensive subject in and of themselves. Further, because their functions are not narrowly confined to the administration of personalized feedback, but rather instruction more broadly, they will only be mentioned briefly within this review.

The following studies are the most applicable works that feature personalized written feedback and compare its effectiveness to generic feedback with reference to well defined learning outcomes.

The study conducted by Gallien and Oomen-Early (2008) correctly identified the trend of higher education increasingly moving towards online instruction and the challenges that this move may impose on teaching and learning. The authors drew attention to the importance of feedback in instruction and identified a number of challenges pertaining to the administration of feedback in exclusively online classrooms. To address this question, they had 84 students from two American universities across multiple classes randomly assigned to either a personalized or generic feedback condition. Those in the personalized condition were provided with one-on-one feedback delivered in written form from the instructor. Within the personalized messages the instructors provided three forms of feedback. 1) Corrective feedback – feedback that corrects the contents of the student's answer to assignments 2) Informative feedback – feedback that addresses the content of a student's answer and links it to course materials, and 3) Socratic feedback – feedback that aims to ask reflective questions about the student's answers to an assignment. Students in the collective feedback condition were given access to a report that discussed different aspects of the assignment and provided examples of well-written answers Gallien and Oomen-Early (2008). The report was generated by the instructor after they had reviewed the performance of all the students throughout the course.

To conclude their study, Gallien and Oomen-Early (2008) found that not only did those students receiving written personalized feedback outperform those in the collective feedback condition, they also reported higher levels of feedback satisfaction. It is important to note however that the personalized feedback generated in this study was done by human instructors, and unsurprisingly required a lot of time commitment. Compared with the collective condition, the personalized feedback took twice as long to generate (approximately 3.5 hours for 39 students). The total number of students within each feedback condition was also relatively small (compared to typical first year university courses) and providing personalized instructor feedback is likely impossible with larger courses, and thus presents the opportunity for learner analytic based feedback systems to address this challenge.

In a study by Parvez and Blank, (2008) a small group of 33 computer science students were required to partake in a multimedia lesson. The ITS classified students according to Felder-Silverman Learning Style Model (Felder & Silverman, 1988), where each learner was classified along the four continuums of sensing-intuitive, visual-verbal, active, reflective, and sequential-global. Students were given a problem description of a movie ticket vending machine and were required to generate a solution to the problem via a diagram. As students worked through the problem, they were provided real time personalized feedback from the ITS according to the learning style they had been classified as. For example, those classified as visual learners were provided visual feedback while those who were classified as verbal learners received written feedback. Overall, students regarded the ITS as being valuable, with 23 out of the 33 students stating they liked the pedagogical advice. In their second study Parvez and Bank (2008) had students placed in either the individualized feedback group using their ITS (personalizing feedback by learning style) or in the control group where they received generic feedback. Between the pre-test before the assignment and the post-test afterwards, it was found that those in the individualized group performed significantly greater, correctly answering 4.6 greater questions (on average) out of a possible 13 questions than those not using the system.

The Advent of Learning Analytics

As touched on briefly in the introduction, learning analytics is a “big data” approach to making decisions about learning, where large volumes of institutional data are analyzed with respect to any number of specified outcomes. A more exact definition was provided at the first

International Conference on Learning Analytics and Knowledge, and reads “learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs (Siemens & Long, 2011, p. 34). This definition is rather broad in its scope, and for good reason. The acquisition of data from various sources and their utilization in the learning sciences could take many forms. Among many other potential applications, Siemens and Long (2011) discussed the ability of learning analytics to help administrative decision-making and resource allocation, identify at-risk learners, and identify institution wide successes and short comings. They further discussed the potential of these systems to be both integrated at the course level to aid students in learning and at the institutional level to reveal statistical trends and habits over time. It must be noted however that learning analytics is a very young field, having only risen to prominence in the past decade (Khalil, Taraghi, & Ebner, 2016). With this youth come some risks and limitations. A number of more recent reviews of the field have found that learning analytics has talked of the potential of the technology far more than they have demonstrated its efficacy (Gašević et al., 2015; Jivet, Scheffel, Specht, & Drachsler, 2018; Viberg et al., 2018). This gap in research has led to a number of papers that have called for a correction within the field, hoping that future developments can be made in tandem with established learning science (Siemens, 2012). These issues as well as an overview of relevant studies to the current paper will be discussed.

Though the core message of the reviewed studies below should emphasize the importance and successes of delivering personalized feedback, other studies have been included to emphasize the importance of other qualities. Some of these will include studies that highlight the potential methodological advantages in research using analytics and large volumes of trace data, while others will be examples of work that failed to achieve significant differences in learning outcomes regarding their personalized feedback interventions.

Perhaps one of the most notable and successful implementations of learning analytics for the administration of feedback is that of Course Signals at Purdue University (Arnold & Pistilli, 2012; Gašević et al., 2015). Course Signals was implemented at Purdue University in 2007 and has been actively providing personalized feedback to students ever since. Utilizing trace data from the university's learning management system (Blackboard) and known student information, Course Signals generates risk levels for students with a patented algorithm. Students are provided personalized feedback throughout a course via a highly streamlined 'traffic light system' that presents students one of either a red, yellow, or green light to a student. If a student is provided a red light by the system, it has been determined by the student's existing characteristics (high school GPA, standardized test scores, years of university, age, etc.) and course activities to date that there is a high likelihood of them failing the course and risking dropout. If a student is provided a yellow light the risk of them failing is still present, however much less than that of a red light. Alternatively, if a student is provided with a green light the system has predicted that their chances of success are relatively high. Though considered its key feature, Course Signals provides functionality beyond its traffic light system. Additional functionality permits instructors the ability to interact with students providing individual support for those who need it through messaging systems (Tanes, Arnold, King, & Remnet, 2011).

According to Arnold and Pistilli (2012), the benefits of Course Signals have been numerous. The implementation of Course Signals has seen marked improvements in student academic achievement with a reduction in the number of Cs, Ds, and Fs awarded to students as well as an increase in As and Bs across courses utilizing the system. Similar results have also been observed when comparing courses using Signals to past years before implementation. The data also suggest that Course Signals is effective with respect to student retention. The earlier students encounter Course Signals in their university career the greater the likelihood they will be retained by the institution (Arnold & Pistilli, 2012). Years of surveys assessing student opinion of Course Signals has also revealed its positive reception. Arnold and Pistilli (2012) reported that 89% of students polled said that Course Signals provided a positive experience for them. Additionally, 58% said that would like to have Course Signals operating in all their other courses. Lastly, Course Signals has received considerable support from university faculty and instructors for the benefits that it affords them. The benefit of providing instructors with useful analytic data and the means to identify and refer struggling students to support sources was held in high regard by some teaching staff (Arnold & Pistilli, 2012). However, there were a few commonly cited concerns by faculty using Course Signals. Prior to the system's implementation there were concerns over the potential flood of students seeking instructor help if provided with disheartening feedback from the system. As well, there were concerns about the system becoming a dependency for newly arrived students, instead of focussing on the desired learning traits. Despite Course Signals' successes (Arnold & Pistilli, 2012), the system has received outside criticism within the learning analytic sphere pertaining to Course Signals' adherence to pedagogical theory and educational practice (Gašević et al., 2015), which will be discussed in greater detail later in this chapter.

At the University of Maastricht in the Netherlands, Tempelaar et al. (2013) implemented a learning analytic system that offered students in a statistics class the opportunity to practice their skills with an online testing platform. The system was designed to effectively combine data on learner dispositions with that of formative assessment data. Formative assessment data came largely through course activities courtesy of an online system that allowed students throughout the course access to practice tests (Tempelaar et al., 2013). Results from these tests provided the students personalized feedback regarding their levels of mastery with the course concepts. Learner dispositions were broken down into differences in learning styles and (mal)adaptive thoughts and behaviors. Though different learning styles were found to have benefitted more from having access to the practice tests and the system overall, the authors noted that students in general found benefit from being informed of their learning style and being given access to test-directed learning (Tempelaar et al., 2013).

Perhaps most pertinent to the present study is the case of E^{2}Coach at the University of Michigan. Much like Course Signals, E^{2}Coach is a learning analytic support system that provides students who opt into its services personalized feedback throughout a course (Wright et al., 2014). Unlike Course Signal's traffic light, E^{2}Coach delivers written feedback to students. E^{2}Coach has been operating at the University of Michigan's introductory physics class since January 2012. To determine whether the system was successful Wright et al. (2014) evaluated students' actual final grades relative to their predicted grades. In such a case the learning system predicts an individual student's grade based on their individual characteristics and past performances and should the student's final performance in the course exceed that of their predicted grade it would be considered a success. For example, should a student with a predicted grade of 60 achieve a final grade of 80 in such a course they would have achieved better than expected. Likewise, if a student was predicted to achieve a 90 in a course but only managed a final grade of 75 they would be classified as worse than expected.

According to Wright et al. (2014) students are delivered their first tailored message the first week after they have opted into the system. Their first message is personalized to address their approach to the course based upon their previous academic history and entrance survey responses. Subsequent messages are generated every few weeks and can include personalized advice about how to prepare for exams, how to best utilize the system (testimonials from past successful students), motivational aids, and provide detailed feedback about the student's current course performance. The system also provides normative feedback — allowing the student to see what students who achieved their desired grade in the past did and how well they will need to perform on future test and assignments to achieve theirs. It should be acknowledged that the messages are originally written by a person before they are selected by student trait and delivered to the learner. Huberth, Chen, Tritz, and McKay, (2015) reported that message authors for E^2Coach have historically been recent physics graduates from both undergraduate and graduate programs working in consultation with course instructors.

The effects of this personalized feedback in terms of academic achievement have been well documented (Wright et al., 2014). Though moderated by usage, E^2Coach appears to be increasing student final grades compared to their predicted grade (Wright et al., 2014). Moderate and high usage users were observed to have an average of 0.11 and 0.18 grade points higher final grade while non-users showed no difference in their final grade compared to their predicted grade. Authors also noted that an achievement gap between genders had long been observed in introductory physics at the University of Michigan, and it was initially hoped that E^2Coach might be able to attenuate or fully close this gap. Unfortunately for all, the gap persisted; however, there was no differential improvement in academic achievement between the genders when provided with E^2Coach, suggesting that the system equally benefitted both males and females (Wright et al., 2014). An unfortunate though major drawback to the work done on E^2Coach is that opting in to use the system has been voluntary (Huberth et al., 2015). This limitation calls into question some of the results of the work conducted on E^{2}Coach as bias in group assignment may be present.

In a study conducted by Kim, Jo, and Park, (2016) the effects of a learning analytics dashboard providing personalized feedback for a management statistics course at a private university in Korea was examined. The authors were primarily interested in how student satisfaction with the dashboard's feedback related to their frequency of use, and how the presence of the dashboard would influence student learning achievement. All graphs presented on the dashboard relayed normative information to the student – showing them their performance in a number of domains (for example: total online activity, writing and reading frequency, as well as activity trends over time) with reference to the class average. Students in the course were going to be graded on a relative scale, and so the authors believed that a dashboard providing norm-referenced personalized feedback would be most helpful in improving student participation and activity with the learning materials. Regarding the success of the feedback system, it was observed that students offered access to the dashboard outperformed those not permitted access (Kim et al., 2016). However, the effects of the dashboard on feedback satisfaction were not so straightforward. A negative correlation was observed between student satisfaction and dashboard usage, suggesting that as students increasingly used the dashboard, they became less satisfied with the feedback that it provided. It was also observed that amongst the students who used the dashboard most frequently, high academic achievers reported lower levels of satisfaction with the feedback. The authors suggested that this relationship may be explained in terms of the high achieving learner's motivational state. They proposed that high achieving learners are already highly motivated and thus do not benefit from exposure to a norm-referenced dashboard. They conceded that future work should include additional student information such as motivational state, and achievement when providing dashboard feedback (Kim et al., 2016).

Khan and Pardo (2016) documented the effects of a learning analytics system providing feedback to students in a first-year engineering course using a flipped classroom strategy. Students in the study were provided access to a feedback dashboard that would feature several different displays comparing the student’s course performance with that of the class average. For example, a student may observe a collection of dials (analogous to a speedometer) showing their level of completion of the course’s practice questions in a specific chapter to those of the class. The rationale was that if students can observe their progress and compare it with their classmates, they would be more likely to utilize the information to improve upon their situation. Unlike the previously discussed examples, the analytic system documented by Khan and Pardo (2016) did not utilize any student demographic data in its predictive models. Only data generated by student activities with the course were used; these included interactions with video clips, answers to online formative and summative assessment questions, access to course resources, and access to the dashboard.

A key component of the study was to assess the frequency of use of the dashboard and observe changes with frequency patterns over time (Khan & Pardo, 2016). Additionally, the authors looked at whether access and frequency of use with the dashboard were at all predictive of student academic achievement for the midterm exam. With regards to frequency of use, it was found that students utilized the dashboard with great frequency at the outset of the course and through the first several weeks. Shortly thereafter, use tapered off dramatically, and the dashboard remained largely unused by most participants by the end of the course. It was also discovered that there was no relationship to use of the dashboard and grade achieved on the midterm. This result suggests that the personalized feedback provided by the dashboard was not providing any benefit in terms of academic achievement (Khan & Pardo, 2016).

Of particular interest to the present paper is the work provided by Guarcello et al. (2017) aimed at assessing supplemental instruction with coarsened exact matching. The intervention they provided was not an example of feedback per se, but rather a workshop designed to provide students additional instruction above and beyond the course lectures. The means by which the data were analyzed is key with reference to the current study. Because the intervention had to be voluntary, covariates related to student achievement were tracked and controlled for in subsequent analyses with Coarsened Exact Matching (CEM). This process seeks to overcome the potential selection bias in non-random group assignment by matching students on covariates related to the outcome variable of interest. Using this method one can be more confident that the differences observed between the groups after matching are due to the intervention and not to selection bias, assuming no significant covariates have been overlooked. In the case of Guarcello et al. (2017) controlling for these covariates revealed a net positive effect of their supplemental instruction intervention. Students who received supplemental instruction had a significantly higher final grade in the course and the odds of them passing the class were 2.2 times higher than those who did not receive any supplemental instruction (Guarcello et al., 2017).

Recently Ochoa et al. (2018) presented a study featuring a multimodal learning analytic system called RAP (Spanish acronym for “Automatic Presentation Feedback”) capable of providing highly personalized feedback to students regarding presentation skills. Contrary to many of the previously discussed works, this system incorporated data from a variety of sources pertaining to presentation ability. RAP recorded, analyzed, and evaluated student posture, gaze, volume, filled pauses, and student presentation slides. With regards to feedback, the goals of the developers were to provide objective and readily understood feedback to students delivering presentations. To ensure unambiguous feedback, student reports were also paired with recorded examples of each student’s presentation to show both desirable and undesirable behavior (Ochoa et al., 2018). Within the study, students were required to present to an empty room filled with the requisite cameras, microphones and other equipment needed by RAP. Upon completion of their presentation, students were supplied with an email link to a website that featured a computer-generated personalized feedback report. The report relayed back to students their score in each of the previously identified domains on a 5-point scale ranging from “very good” to “very bad”. Students were also given a total score averaged across all domains. To assess the system’s feedback accuracy, feedback was compared with those of expert human reviewers. Human reviewers and RAP agreed 65% of time with regards to posture, 78% for gaze, 75% for filled pauses, and 71% for volume. Agreement was much lower (ranging from 44% to 59%) for domains related to evaluations of the presentation slides, but the authors believe this was due to the system being unable to properly evaluate pictures contained within a student’s presentation slides. Regarding student perceptions, a post study questionnaire found substantial support for RAP, with 65% and 58% of students reporting that the “usefulness” of the system and “learning from feedback” as excellent (Ochoa et al., 2018).

Another multimodal learning analytic system providing feedback to student learners can be seen in Mangaroska, Sharma, Giannakos, Trætteberg, and Dillenbourg's (2018) study. A group of computer science students were given access to a mirroring tool designed to report back to them a summary of their performance and task activities in a test of debugging. The mirroring tool collected and relayed data on number of lines of code, errors, code warnings, debugging events, execution of commands, and unit test results. Feedback was personalized to individual students in terms of their inputs and was aimed at providing the student support in working towards the exercise's goals. Gaze monitoring technology – where eye-tracking software was used to monitor student gaze – was also utilized within the study to assess how expert and novice debuggers made use of the mirroring tool. Results showed students who processed the information within the mirroring tool improved their performance in the debugging task, though the relationship between the tool and student success was not straightforward (Mangaroska et al., 2018). Expert debuggers benefitted from the tool to a greater extent than novice debuggers and there was a negative correlation between time spent viewing the mirroring tool's interface and performance. The authors claimed this result indicates that the time spent viewing the tool is not important for success, rather it is the time spent processing the information and acting upon it productively (Mangaroska et al., 2018).

While the reviews earlier suggest that learning analytics and data driven education are delivering precious and irrefutably valuable tools to the study of education, the bigger picture may not be so clear. Despite the increasing prevalence of learning analytic systems and software implementations across many institutions of higher education, there appears to be a shortfall of evidence suggesting that such implementations are making a positive difference (Viberg et al., 2018).

With regards to feedback Gašević et al. (2015) pointed out that many learning analytic tools are generally not developed from established educational theories. For example, many learning analytics dashboards provide personalized feedback that is norm-referenced with respect to many student activities and grades within a course. While the thought of, and rationale for, a norm-referenced dashboard might make sense to a developer or an outside observer, the existing research on norm-referenced feedback is quite clear: as a part of Kluger and DeNisi's (1996) review they systematically surveyed the existing literature on norm-referenced feedback and found it to be more detrimental than beneficial. They even provided a theoretical framework that can account for why norm-referenced feedback is unlikely to result in increased learning and why other forms of feedback may be superior. Concerning FIT (Kluger & DeNisi, 1996), one can see that norm-referenced feedback moves the student's locus of attention up the hierarchy of processing towards evaluations of the self. This movement distracts the student from what is most pertinent to closing the gap between their current performance and desire performance – the specifics of the task. Despite this finding, numerous learning analytic systems continue to develop feedback systems providing students norm-referenced feedback. Though one can find some successes of norm-referenced feedback (Delaval, Michinov, Le Bohec, & Le Hénaff, 2017), one will likely find more examples of ineffective systems or students expressing their disinterest in norm-referenced feedback (Khan & Pardo, 2016; Tan, Yang, Koh, & Jonathan, 2016; Wise, Zhao, & Hausknecht, 2014). It may seem then to someone familiar with the established feedback science that these developers may in fact be designing these systems in near isolation of established feedback theory.

In Gasevic, et al.'s (2015) critique of learning analytic practice more broadly, they also went on to criticize the content of Course Signals' messages (Tanes et al., 2011) and its implementation at Purdue University specifically (Arnold & Pistilli, 2012). With reference to the messages, it was found that concise instructional messages provided at the level of the task were rarely observed, despite the large volume of research showing that this is where feedback is most effective in increasing learning (Kluger & DeNisi, 1996; Hattie & Timperley, 2007; Shute, 2008). Criticism levelled at Course Signals at Purdue (Arnold & Pistilli, 2012) was centered around the lack of alignment between Course Signals and established research on instructional practices. Course Signals was originally designed as an early alert system for identifying students at risk, and thus lacks a design more conducive to identifying gaps in student knowledge and aiding instructors (Gasevic, et al., 2015).

A more recent review by Jivet et al. (2018) suggest that little has changed in terms of aligning learning analytic tools and educational theory. Looking more specifically at learning analytic dashboards, the authors wanted to see how individual papers were evaluating their creations. Perhaps unsurprisingly, less than a third of the papers they reviewed explicitly targeted the cognitive competence of students using the dashboard. Additionally, dashboards were often evaluated in terms of their usability and whether students were satisfied with their operation, instead of whether or not they provide a measurable benefit to the learners. In line with Gasevic, et al. (2015) it was also observed that these dashboards frequently utilized norm-referenced comparisons with other students in the course. Lastly, a mere seven papers within their review drew upon educational concepts for use in dashboard evaluation, perhaps indicating that many of these systems are in need of a return to the learning sciences.

Though that criticism might seem harsh or unwarranted, an equally recent review by Viberg et al. (2018) suggests otherwise. Their paper points to the lack of evidence that learning analytic systems are both increasing student achievement indices and supporting teaching. In their meta-analysis of 252 publications pertaining to learning analytics in higher education from 2012 to 2017, a mere 23 papers (9%) demonstrated evidence of learning analytics improving student learning outcomes, prompting the authors to say that the potential for learning analytics still far outweighs the actual evidence. An additional 16% of papers within their review explicitly spoke of this potential for learning analytics to improve learning outcomes but did not present any evidence of this (Viberg et al., 2018).

Viberg et al., (2018) cited earlier in their paper that as a new and emerging field, publications within learning analytics have increased since 2012. As a young academic field, a review of this nature is important in surveying the landscape and assessing the larger directions and intentions of the field. Learning analytics has already succeeded in becoming an integral part of higher education and as both Gaesevic, et al. (2015) and Viberg et al. (2018) have shown, the field in its current state requires both greater adherence to the established learning theory it purports to contribute to, as well as a focus on providing greater evidence of its efficacy instead of touting its potential.

Summary

In this chapter the existing relevant studies on personalized feedback have been reviewed, as well as a number of formative theories regarding the operation of feedback from a psychological and educational perspective historically to the present day. Although there are many studies that utilize learning analytics to deliver personalized feedback for individual students, few have used feedback in a similar manner to that of the present study. Even fewer have demonstrated measurable increases in learning outcomes as a result of that feedback. Research prior to the advent of learning analytic systems has shown that students by and large prefer personalized feedback over generic feedback. These results often reveal higher levels of satisfaction with feedback as well. In the case of ITSs the benefits to academic achievement appear to be robust, though these systems are not necessarily feedback specific, and routinely provide much more to a student by ways of instruction.

The review of the relevant research on learning analytics-based feedback has shown that feedback can come in many different forms while still falling under the definition of personalized feedback. Learning analytic systems have been used to identify students at risk of dropping out, aid in developing basic task-related skills, deliver advice feedback specifically, and many other applications. In research where learning analytics are used to deliver personalized feedback, it is most often done according to student activities within the course. These can range from task specific actions such as the outcome of a practice assignment or commentary on a student's presentation skills to feedback based upon a student's summative course performance over months of work. Student feedback is also often personalized according to demographic information routinely sourced from individual university's data warehouses or entrance surveys. This information can be used to provide meaningful feedback to students based upon their status within different demographics, for example: students coming from a rural background may receive additional support compared to those who grew up in urban environment.

This chapter also revealed some of the limitations or short-comings of the learning analytics movement. It was shown that learning analytics sometimes seem to develop in isolation of the existing learning sciences, creating products that sometimes ignore existing pedagogical principles. This shortcoming is especially the case with feedback, as many learning analytic dashboards relay norm-referenced feedback to students, a practice that has historically been shown to be largely ineffective or detrimental. Bearing that in mind, it is no surprise that many learning analytic systems over the past decade have not been able to demonstrate their effectiveness in improving student learning outcomes. This observation does not suggest that the greater project is in jeopardy, only that there needs to be a more committed effort to designing systems that harmonize well with the existing learning sciences. This study aims to evaluate the effectiveness of a learning analytic system that delivers computer generated personalized feedback to students based upon their course performance and activities, and demographics with reference to the established literature on feedback.

Chapter Three: Methodology

Introduction

This chapter sets out the analytic methods used in examining the effectiveness of the personalized feedback intervention on indices of academic achievement and feedback satisfaction. The first section describes the sample and provides a collection of demographic information. The second section discusses data collection and provides an overview of the Student Advice Recommender Agent (SARA)(Greer et al., 2015). How SARA differentiates feedback, how this feedback is delivered to students, and a number of supportive examples will be presented. The third section will discuss the measures used and provide an overview of their psychometric properties. Lastly, a conceptual introduction to the statistical technique of matching will be provided with an overview of the matching methods and protocols.

Data

Participants were undergraduate students from the University of Saskatchewan enrolled in multiple sections of an introductory 100-level biology course over the span of two years. The sample was composed of courses held in September 2016 (S16), January 2017 (J17), September 2017 (S17), and January 2018 (J18). A summary of demographic information can be seen below in Table 1.

Table 3.1.
Age and Gender Demographics of Classes September 2016 through January 2018

	S16		**J17**		**S17**		**J18**		**Total**	
N	974		558		906		632		3070	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Gender										
Males	327	33.57	223	40	266	29.36	235	37.18	1051	34.23
Females	647	66.43	335	60	640	70.64	397	62.82	2019	65.77
High School										
Saskatchewan	769	78.95	405	72.58	722	79.70	458	72.47	2354	76.68
Student Type										
New Students	762	78.23	280	50.18	734	81.01	339	53.64	2115	68.89
Aboriginal Ancestry	64	6.57	62	11.11	73	8.10	71	11.2	270	8.79
	m	**SD**	**m**	**SD**	**m**	**SD**	**m**	**SD**	**m**	**SD**
Age (years)	21.66	2.26	21.66	4.34	18.89	2.02	21.26	3.82	19.93	3.28

m: mean
SD: standard deviation

Data used for the differentiation/personalization of feedback was collected from several different sources. Demographic data was attained from the university's data warehouse — which included student status, year of study, high school graduation, age etc. Student activity data within Blackboard (the University of Saskatchewan's learning management system) was also recorded and logged for use in personalizing feedback, some of which include hovers, clicks, and number of links followed. Finally, two surveys were administered to the students during the experiment. At the start of the course, students were required to fill out an entrance survey that included a set of items pertaining to their thoughts about the course, their anticipated performance, and their threshold of disappointment. Students were also requested to complete an exit survey following the completion of the course. Items in the exit survey pertained to students' level of engagement with the feedback system, feedback satisfaction, and a number of other course-related measures (e.g., participation with structured study sessions, mock exams, and weekly quizzes). Following the end of each course, data was collected and compiled at the Gwenna Moss Centre for Teaching and Learning and made available for the proposed study. Datasets from each year were then combined for analysis. All information potentially linking data to identifiable students was removed during this process. Once data from multiple years was merged, it was screened for outliers, missing entries, and data errors.

The learning analytic system responsible for differentiating and distributing feedback is known as SARA (see Greer et al., (2015) for additional information). SARA takes into account approximately 40 distinct student attributes from the above-mentioned data sources and uses those to deliver personalized feedback messages to students. As with analogous learning analytic systems like E^2Coach at the University of Michigan (Wright et al., 2014), feedback messages were originally written by human authors in conjunction with instructor support.

At the outset of each iteration of the course, students were randomly assigned into either the personalized or the generic (control) condition. Students in the generic feedback condition were administered a weekly feedback message through Blackboard that was common across all members within the group. Hence, there was no differentiation by student characteristics with respect to any of the specific components of the feedback message. By contrast, the personalized feedback group received a weekly feedback report that catered to their individual characteristics and course performance indices to date. For example, following the midterm exam a student may receive feedback based on both their performance on the midterm exam and their status as a student coming from a rural high school. Such a feedback message may let them know how their performance on the midterm compares with the class average and may also reassure them that adjusting to life at a university in a city may be difficult. SARA may then refer them to any number of on-campus supports to help with the adjustment. Though the feedback types may differ drastically, it must be stated that both generic and personalized feedback reports contained all essential course information and reminders of course responsibilities. Hence, no group should have been privileged in terms of access to vital course information. See below (Figure 5) as an example of a student feedback message being personalized three ways according to performance on the first lab exam.

BIO 120 Lab Exam 1 Mark <50%	Unfortunately, it seems you did not do so well on the lab exam. Failing the first lab exam is going to hurt your mark, but it isn't the end of the world. The labs change (a lot) in the next few weeks. There will be much less microscope work and almost no organisms to memorize. Instead, you will be focussing on the foundations of genetics: terminology, theory, and math. Please remember that you have a quiz at the beginning of Lab 5. The contents of this quiz include questions on lab 4 and pre-lab questions on lab 5. Be familiar with the manual! After this there will be a genetics quiz, and the last lab exam. Make sure that you put the time in to do well!
BIO 120 Lab Exam 1 Mark 50% > 70%	You passed the lab exam but couldn't quite beat the class average. You are going to have to put some effort in if you want to pull your mark up. You may be happy to hear that the labs change for the second half of the course- no more microscope work, more worksheets and genetics problems. If you like math, this will be your chance to shine! Please remember that you have a quiz at the beginning of lab 5. The contents of this quiz include questions on lab 4 and pre-lab questions on lab 5. Be familiar with the manual! After this there will be a genetics quiz, and the last lab exam. You're all done with microscopes now- the labs from here on out will be focussing on the foundations of genetics: terminology, theory, and math. Keep working hard to maintain a good average, every mark counts!
BIO 120 Lab Exam 1 Mark 70% > 90%	**Congratulations!** You did quite well on the exam, in fact, you beat the class average! It doesn't hurt to celebrate your accomplishments, but don't let this early success make you complacent. Please remember that you have a quiz at the beginning of lab 5. The contents of this quiz include questions on lab 4 and pre-lab questions on lab 5. Be familiar with the manual! After this there will be a genetics quiz, and the last lab exam. You're all done with microscopes now- the labs from here on out will be focussing on the foundations of genetics: terminology, theory, and math. Keep working hard to maintain a good average, every mark counts!

Figure 3.1. Personalized feedback message according to lab exam grade.

Measures

Student final grade was composed of a series of assignments throughout the course in addition to a midterm and final exam. Though the determination of final grade varied slightly from year to year (dependent upon the addition or removal of assignments), grading criterion remained highly consistent. Final grades were absolute measures of course performance and were not graded on a curve.

Feedback satisfaction was assessed via a short single factor five-item survey included within the larger course exit survey and was featured in a related study (Schmidt, Mousavi, Squires, & Wilson, 2018). For the first three installments of the survey, the first two items used dichotomous response scales (yes/no) while the next three items used a five-point Likert-style polytomous response scale with options ranging from “Strongly Disagree” to “Strongly Agree”. For the most recent iteration of the course in January 2018, the scale had been updated. The first two questions were reworded slightly and featured the same Likert-type response scales as the other items.

To ensure that data could be aggregated across years of the course, the first two questions of the satisfaction scale were dichotomized for the January 2018 installment. Responses to those first two items were then considered to either endorse the item or not endorse the item. More specifically, the responses “Agree” and “Strongly Agree” were coded as endorsed, while the responses “Disagree” and “Strongly Disagree” were coded as not endorsed. Hanisch's (1992) work on the Job Descriptive Index discovered that neutral responses to positively keyed items behaved much like negative responses. Thus, neutral responses within the feedback satisfaction survey were also treated as not endorsed. Despite the potential loss of information in the dichotomization of the response scales, the satisfaction scale reported adequate reliability with a Cronbach’s alpha estimate of α = .85 (DeVellis, 2016).

All five items on the feedback satisfaction survey are positively keyed, and calculating an overall score requires the researcher to sum a participant’s responses. Higher sums imply greater levels of feedback satisfaction than lower sums. The maximum achievable value is seventeen while the lowest possible value is 5. See below (Figure 6) an example of both a dichotomous and polytomous item from the feedback satisfaction scale (For full scale see Appendix A):

	Yes	No	N/A
Did you appreciate receiving your weekly note from SARA?	○	○	○

To what extent do you agree or disagree with the following statements:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Agree	N/A
My weekly note was a good reminder of my performance in Biology 120.	○	○	○	○	○	○

Figure 3.2. Sample questions from the feedback satisfaction survey.

Statistical Analysis

Statistical analyses for the study were conducted using both the program R (R Core Team, 2018) and Stata version 15 (StataCorp, 2017). Basic descriptive statistics (e.g., frequencies, means, standard deviations, correlations) were used to describe the characteristics (e.g., age, gender, nationality, background, aboriginal ancestry) of the sample as well as search for relevant covariates for use in matching using tests of correlation. Identified covariates were then used to build progressively complex linear models seeking to explain as much of the observed variance as possible. Successive models were compared using an ANOVA to determine whether a model is of superior fit. The final selection of covariates for matching were extracted from the most parsimonious linear model, or more specifically, the model that can account for as much variance as possible using the fewest number of covariates.

The study was of post-test only quasi-experimental design. Students were randomly assigned to either feedback condition at the outset of the course. However, because students in the study were working for real grades, ethical approval required the ability of students to self select a different feedback condition if they preferred. Because of the lack of true random assignment, matching is being proposed as a means of better approximating an ideal data set and allowing greater statistical inference from findings. Because the existing dataset already approximates a fully randomized experiment (i.e., neither group demonstrates any significant differences in known covariates of final grade), Mahalanobis distance matching using nearest neighbor without replacement protocol was chosen as the main method of analysis to address the primary research question.

Among the many matching distances available, Mahalanobis matching has been chosen because it possesses both a number of theoretical advantages over other alternatives and for its suitability to the specific data of the proposed study (Gu & Rosenbaum, 1993). Compared to the most popular form of matching, propensity score matching (PSM), Mahalanobis matching results in the approximation of a fully-blocked design as opposed to a randomized design. Further, as an observational dataset approaches randomization through matching, it has been demonstrated that PSM begins to remove observations at random, resulting in an eventual increase in model dependence and bias (King & Nielsen, 2016). Because the present dataset already approximates a fully randomized design it was suspected that PSM would only result in increasing bias. Regarding the specific suitability of the present dataset, Mahalanobis matching appeared more appropriate for analysis than Coarsened-Exact-Matching (CEM). Within CEM, one or more continuous variables are temporarily coarsened such that exact matching can be more easily accomplished (Blackwell, Iacus, King, & Porro, 2009). Following matching, both treated and controlled observations without matches are removed and the remaining data are assigned weights to account for differences in frequency (Iacus, King, & Porro, 2012). For example, total years of education may be coarsened to two-year diploma, four-year degree, master's degree etc. to make exact matching on education easier. This approach was deemed inappropriate for the present study as there was no shortage of data or difficulties in finding matches.

Conceptually speaking, the goal of matching is to improve a given dataset by balancing both known and unknown covariates of the outcome variable of interest within both the treatment and control groups. Doing so should result in a new balanced dataset that approximates either a randomized (where known and unknown covariates are balanced on average) or a fully blocked design (where known covariates are balanced exactly, and unknown covariates are balanced on average) (King & Nielsen, 2016).

$$TE_i = Y_i(1) - Y_i(0)$$

The above equation describes how matching allows one to calculate the treatment effect (TE_i) of the ith subject. This task is accomplished by taking the value of Y as if it were treated $Y_i(1)$ and subtracting the value of Y as if it weren't treated $Y_i(0)$. The inherent difficulty here is that the first term is observed, while the second term is unobserved and must be estimated. Matching attempts to solve this problem by finding an analogous unit from the control group that is highly similar to the treated unit in known pre-treatment control covariates. Given the quasi-experimental design of the proposed study, this conceptual approach appears most appropriate for determining whether students within the personalized feedback condition are in fact out performing those within the control condition. Referring to the below equation, an average treatment effect on the treated (ATET) is calculated by summing the treatment effects of all subjects within the treatment group. The significance of the ATET will serve as the test of whether the personalized feedback condition is improving final grades over the control group.

$$Sample\ Average\ Treament\ Effect\ on\ the\ Treated = \frac{Mean}{i \in \{T_i\ = 1\}}(TE_i)$$

The below equation describes the Mahalanobis distance between a control unit X_C and a treated unit X_T. (X_C – X_T) is a matrix of distances between controlled and treated units, with the term $(X_C - X_T)^T$ representing the transpose of that same matrix (Mahalanobis, 1936). If attempting to calculate the ATET, S^{-1} refers to the inverse variance covariance matrix of X in the entire control group. When using nearest neighbor matching, each treated participant is matched with the closest control unit (in terms of the Mahalanobis distance) that possesses the same specified covariates.

$$Mahalanobis\ Distance(X_C, X_T) = \sqrt{(X_C\ - X_T)^{\mathrm{T}} S^{-1} (X_C\ - X_T)}$$

The same procedure was used for evaluating the success of the personalized feedback system as it relates to feedback satisfaction. An ATET was calculated following the creation of a matched dataset using Mahalanobis distance matching, with the intention of balancing all known and unknown covariates relating to feedback satisfaction.

In summary, the proposed methodology entails searching the dataset for pre-treatment covariates of both final grade and student feedback satisfaction. Linear models using the collection of covariates for both final grade and feedback satisfaction will then be created. Models will be compared to determine which is the most parsimonious collection of covariates. The covariates present within the best-fit models will then be used for creating the matched datasets from which treatment estimates (both the ATET and the ATE) will be generated.

Chapter 4 Results

Overview

This chapter will present the results of the study with respect to the research questions outlined in the previous chapters. First, will students receiving personalized feedback exhibit higher mean final grades compared to those receiving generalized feedback? Second, will students receiving personalized feedback exhibit higher levels of feedback satisfaction compared to those receiving generalized feedback?

Data Cleaning Procedures

At varying levels of analysis, data loss was substantial. Regarding the first research question, far more data could be retained. A very small number of cases with incomplete data were removed. Analyses relating to gender also warranted the removal of students that self-identified as neither “male” or “female” on the entrance survey, as they represented fewer than .01% of the total sample. In total, this left a sample of 3053 students for the evaluation of the first research question. With respect to the second research question assessing feedback satisfaction, there was substantial data loss, with only 920 students completing the entire feedback satisfaction survey.

SARA’s system collects data on approximately 110 variables (the total number varies by year according to research interest). As such there exist many variables outside of those presented or discussed in the preceding chapters that were explored with reference to both research questions. The variables presented within this chapter represent those found to have relevance to evaluating the proposed research questions.

Research Question 1

Will students receiving personalized feedback outperform those receiving generalized feedback with respect to their final grade?

Based on the previous research, it was hypothesized that students receiving personalized feedback would outperform those only receiving generalized feedback. To begin evaluating this question, SARA's dataset was explored for pre-treatment variables with significant correlations with student final grade appropriate for matching. Table 2 displays the means, standard deviations, and correlations with final grade. Two such appropriate variables were found and can be seen in Table 2: (1) Student predicted grade (according to SARA's algorithm) *r* = .66, p <.001, and (2) Student term GPA *r* = .91, p < .001.

Table 4.1.
Means, standard deviations, and correlations with confidence intervals for final grade

Variable	*M*	*SD*	1	2
1. Final Grade	65.66	15.10		
2. Predicted Grade	66.63	11.19	0.66** [0.63, 0.68]	
3. Term GPA	68.43	13.38	0.91** [0.90, 0.91]	0.64** [0.62, 0.66]

Note: M and *SD* are used to represent mean and standard deviation. Values within the square brackets indicate the 95% confidence interval for each correlation. **indicates *p* < .01

Table 3 displays basic descriptive statistics across feedback conditions (means and standard deviations). It should be noted that term GPA includes the final grade of biology within its calculation, and thus is artificially inflated. Despite this potential limitation, no greater alternative approximating student ability within a postsecondary context could be found.

With respect to SARA's algorithm for student predicted grade, the exact nature of the algorithm is not known. However, the variables used for predicting final grade by the algorithm are: student self efficacy (determined on the entrance survey), postal district, whether the student attended an urban or rural high school, student high school GPA, and their grade in Biology 30. Though these variables possess a relationship to final grade, they were not used by themselves for matching, as they have already been accounted for with SARA's prediction.

Table 4.2.
Means and standard deviations for feedback conditions

	Personalized		Generalized	
Variable	*M*	*SD*	*M*	*SD*
Final Grade	65.18	15.81	65.49	15.68
Term GPA	68.31	13.39	68.53	13.38
Predicted Grade	66.32	11.42	66.75	11.17

Note: M and *SD* are used to represent mean and standard deviation

To determine the most appropriate selection of variables for matching, two regression models were created to determine which combination of variables resulted in the most parsimonious prediction of student final grade. The results of the regression models can be seen below in Table 4. The first model utilized predicted grade as the single predictor, while the second model utilized both predicted grade and term GPA. Results showed that the combination of both predicted grade and term GPA provided the best fit to the data, and thus both variables were selected for matching.

Table 4.3.
Regression results for student final grade

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	*sr^2*	*sr^2* 95% CI [LL, UL]	*r*	Fit	Difference
(Intercept)	6.76**	[4.31, 9.20]							
Predicted Grade	0.88**	[0.85, 0.92]	0.66	[0.63, 0.69]	.43	[NA, NA]	.66**		
								R^2 = .439**	
(Intercept)	-9.38**	[-10.76, -8.00]							
Predicted Grade	0.17**	[0.14, 0.19]	0.12	[0.11, 0.14]	.01	[.01, .01]	.66**		
Term GPA	.093**	[0.91, 0.95]	0.83	[0.81, 0.85]	.40	[.38, .43]	.91**		
								R^2 = .832**	ΔR^2 = .403** 95% CI [.38, .43]

Note: A significant b-weight indicates the beta-weight and semi-partial correlation are also significant.
b represents unstandardized regression weights. beta indicates the standardized regression weights.
sr^2 represents the semi-partial correlation squared. r represents the zero-order correlation.
Square brackets are used to enclose the lower and upper limits of a confidence interval.
* indicates p < .05. ** indicates p < .01.

With matching variables identified, treatment effect estimates were made for all years combined as well as for each course year separately (See Table 5). It was found that treatment effect estimates for both ATEs (average treatment effect) and ATETs (average treatment effect on the treated) hovered around zero, and no estimates reached statistical significance, suggesting that the personalized feedback treatment was not providing a benefit to students in terms of academic achievement.

Table 4.4.
Treatment effect estimates for final grade

Final Grade		*Coef.*	*SE*	*z*	*p*	95% CI
All years combined	ATE	.03	.25	0.12	0.90	[-.467, .529]
	ATET	.22	.28	0.80	0.42	[-.321, .764]
September 2016	ATE	.50	.42	1.2	0.23	[-.321, 1.329]
	ATET	.28	.47	0.60	0.55	[-.642, 1.213]
January 2017	ATE	-.71	.67	-1.06	0.29	[-2.017, .605]
	ATET	-.86	.76	-1.14	0.26	[-2.339, .622]
September 2017	ATE	.19	.41	.47	0.64	[-.609, .992]
	ATET	.37	.45	0.82	0.41	[-.513, 1.249]
January 2018	ATE	-.70	.59	-1.19	0.24	[-1.852, .455]
	ATET	-.96	.62	-1.55	0.12	[-2.164, .251]

Note: ATE refers to the average treatment effect in the population. ATET refers to the average treatment effect on the treated. Coef. is the specified treatment effect and SE is the standard error. * indicates p < .05. ** indicates p < .01.

Table 6 shows the mean final grades for the total sample, as well as by individual year. Looking at the total sample one can see that the final grades are essentially identical, differing by less than half a percentage point. Differences are slightly more pronounced looking at each course iteration individually, but again are essentially the same. The only notable difference across years is that January course offerings generally report lower grades than September offerings. However, this seems to affect each feedback condition equally.

Table 4.5.
Means and standard deviations for feedback conditions

	Personalized		Generalized	
Final Grade	*M*	*SD*	*M*	*SD*
Total	65.18	15.81	65.49	15.68
September 2016	67.97	14.08	68.05	13.64
January 2017	63.53	17.56	64.75	18.14
September 2017	67.29	14.25	66.11	14.72
January 2018	59.64	17.07	61.08	16.76

Note: M and *SD* are used to represent mean and standard deviation.

Research Question 2

Will students receiving personalized feedback report higher levels of feedback satisfaction than those receiving generalized feedback?

Based on the previous research, it was hypothesized that students receiving personalized feedback would report higher levels of feedback satisfaction than those only receiving generalized feedback. To address this question, SARA's dataset was explored for pre-treatment variables with significant correlations with feedback satisfaction for matching. Table 7 (below) shows the pre-treatment variables found within SARA's dataset used to test for correlations with feedback satisfaction. Gender was found to be the only pre-treatment variable with a significant correlation to feedback satisfaction. Though the relationship was weak, it was found that being female was significantly correlated with feedback satisfaction r = .09, p < .01.

Table 4.6.
Means, standard deviations, and correlations with confidence intervals for feedback satisfaction

Variable	*M*	*SD*	1	2	3	4	5
1. Satisfaction Total	11.38	3.11					
2. Female	0.66	0.47	.09** [.03, .15]				
3. Aboriginal Ancestry	0.09	0.28	-.01 [-.07, .06]	.03 [-.00, .07]			
4. Age	19.92	3.28	.01 [-.05, .08]	-.09** [-.13, -.06]	.12** [.09, .16]		
5. Urban High School	.71	0.45	-.05 [-.13, .02]	.01 [-.03, .05]	.01 [-.03, .05]	.02 [-.02, .06]	
6. Term GPA	68.43	13.38	-.03 [-.09, .04]	-.01 [-.04, .03]	-.15** [-.18, -.11]	-.11** [-.14, -.07]	.07** [.03, .11]

Note: M and *SD* are used to represent mean and standard deviation. Values within the square brackets indicate the 95% confidence interval for each correlation. **indicates *p* < .01

Given that gender was the only pre-treatment variable found to be significantly predictive of feedback satisfaction, it was deemed not necessary to build a succession of regression models to determine which collection of variables would be most appropriate for matching. Subjects were matched by gender alone before treatment effects were calculated.

Table 4.7.
Treatment effect estimates for feedback satisfaction

Satisfaction		*Coef.*	*SE*	*z*	*p*	95% CI
All years combined	ATE	.56**	.20	2.78	0.005	[.166, .960]
	ATET	.56**	.20	2.78	0.005	[.166, .960]
September 2016	ATE	.69*	.34	2.02	0.043	[.022, 1.360]
	ATET	.67*	.34	1.96	0.050	[.000, 1.345]
January 2017	ATE	1.56*	.76	2.06	0.039	[.079, 3.040]
	ATET	1.52*	.75	2.05	0.041	[.064, 2.985]
September 2017	ATE	.35	.34	1.02	0.308	[-.323, 1.022]
	ATET	.38	.35	1.06	0.289	[-.312, 1.046]
January 2018	ATE	.25	.41	0.60	0.546	[-.551, 1.042]
	ATET	.22	.41	0.53	0.595	[-.579, 1.011]

Note: ATE refers to the average treatment effect in the population. ATET refers to the average treatment effect on the treated. Coef. is the specified treatment effect and SE is the standard error. * indicates p < .05. ** indicates p < .01.

Treatment effect estimates were made for all years combined as well as for each course year separately and can be seen in Table 8 (above). Significant treatment effects were found for all years combined, as well as the first two implementations of SARA (September 2016 & January 2017). With respect to the entire dataset both treatment effects showed a benefit of .56 points increase on the feedback satisfaction scale when participants were in the personalized feedback condition. However, this is only half of the story. For the September 2016 installation, a significant ATET was estimated at .67, and an even greater estimate was found for the January 2017 installation of 1.52. However, from that point forward, there appears to be no significant differences in feedback satisfaction between the personalized and generalized groups, with notably smaller and non-significant treatment effect estimates for September 2017 and January

2018. Table 9 shows the means and standard deviations in feedback satisfaction between treatment groups across years.

Table 4.8.
Feedback satisfaction means and standard deviations for feedback conditions

	Personalized		Generalized	
Satisfaction	*M*	*SD*	*M*	*SD*
Total	11.67	2.87	11.11	3.30
September 2016	12.07	2.71	11.39	3.13
January 2017	12.23	3.36	10.74	3.92
September 2017	11.51	3.25	11.22	3.25
January 2018	11.13	2.70	10.75	3.32

Note: M and *SD* are used to represent mean and standard deviation.

Figure 7 plots the means of feedback satisfaction overall across years revealing an overall steady decrease in satisfaction with SARA.

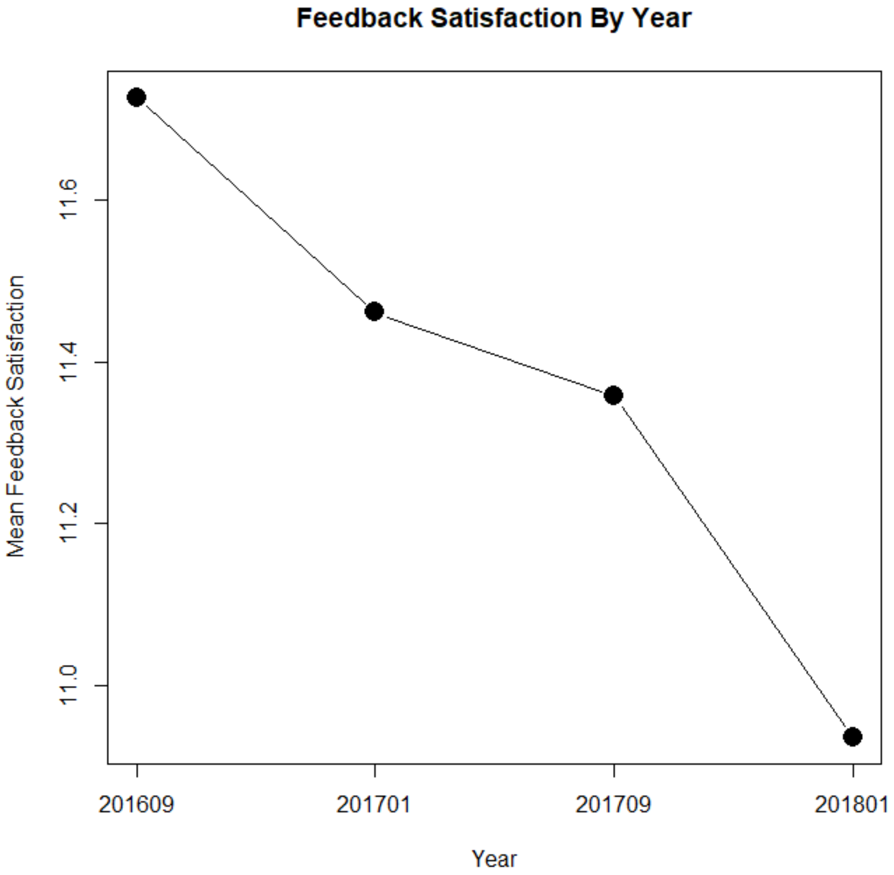

Figure 4.1. Plot of satisfaction means by course year.

To investigate this finding further and determine if overall mean satisfaction was significantly different from earlier years onward, a one-way ANOVA was conducted – results can be seen in Table 10 and 11, and reveal a significant difference in means across years.

Table 4.9.
Descriptive statistics for feedback satisfaction as a function of course offering

Year	*M*	*SD*
September 2016	11.73	2.94
January 2017	11.46	3.71
September 2017	11.36	3.10
January 2018	10.94	3.04

Note: M and *SD* are used to represent mean and standard deviation.

Table 4.10.
ANOVA results using feedback satisfaction as the dependent variable

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2	90% CI partial η^2 [LL, UL]
(Intercept)	39741.60	1	39741.60	4131.48	.000		
Year	78.34	3	26.11	2.71	.044	.01	[.00, .02]
Error	8811.21	916	9.62				

Note: LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval.

Table 12 shows the follow-up results of Tukey's HSD; the only significant difference in mean feedback satisfaction was observed between September 2016 and January 2018, at which point the difference is approximately one quarter of a standard deviation of the pooled variance between course offerings (Cohen's *d = 0.*27).

Table 4.11.
Tukey multiple comparisons following one-way ANOVA of course offering

Comparison	*Diff*	95% CI	*p*
September 2016 – January 2017	-0.27	[-1.234, 0.702]	0.89
September 2016 – September 2017	-0.37	[-1.015, 0.276]	0.45
September 2016 – January 2018	-0.79	[-1.510, -0.074]	0.02*
January 2017 – September 2017	-0.10	[-1.059, 0.851]	0.99
January 2017 – January 2018	-0.53	[-1.530, 0.480]	0.53
September 2017 – January 2018	-0.42	[-1.121, 0.278]	0.41

Note:Diff refers to the difference in means. * indicates p < .05.

Summary

With regards to the first research question, it was determined that students receiving personalized feedback did not achieve beyond those receiving generalized feedback in terms of final grade when matched on both their predicted grade and their term GPA. It was observed that all calculated treatment effects were near zero, and none reached statistical significance. This finding suggests that the personalized computer-generated feedback provided by SARA was not providing a benefit to these students in terms of their achievement in biology.

In regard to the second research question, it was found that students receiving personalized feedback reported higher levels of feedback satisfaction than those receiving generalized feedback. Students were matched by gender alone, as it was discovered that gender was the only significant pre-treatment variable associated with feedback satisfaction. Looking at all years in total, the average treatment effect on the treated was calculated at .56 points on the feedback satisfaction scale. Unfortunately, this is only half of the story. The first two installments of the present feedback intervention resulted in significant positive treatment effects (ATETs), calculated at .67 points increase for September 2016, and 1.52 points increase for January 2017. Subsequent years saw far reduced treatment effect estimates that did not achieve statistical significance. It was also observed that mean feedback satisfaction had been in steady decline since the first implementation of SARA.

Chapter 5 Discussion

Overview

This chapter will begin with a discussion of the results of each research question with reference to the literature discussed in Chapter Two. The implications of the findings of each of the research questions will then be discussed. Following that, the chapter will conclude with an overview of the limitations of the research and future directions. It is hoped that future research regarding the implementation and evaluation of analogous learning analytic (LA) systems will be informed by the present paper.

Learning analytics systems are an inevitable part of the future of postsecondary education (Dawson, Gašević, Siemens, & Joksimovic, 2014; Siemens, 2012). Therefore, there is a need to establish the parameters under which these systems operate to their greatest utility. The present paper identified a situation where a LA system (SARA) administered computer-generated feedback to students in an introductory biology course at the University of Saskatchewan. One group received personalized feedback, where feedback was tailored to their individual characteristics, while the other received generalized feedback that was common to all members of the group. Data collection for this study begin in September 2016 and continued until January 2018 for a total of four course offerings. The current study was designed to assess the effectiveness of personalizing computer-generated feedback — in terms of both academic achievement and satisfaction with feedback — compared to generalized feedback. Despite the precedent set by the literature, it was found that personalized feedback did not improve students' academic achievement. However, it was observed that those receiving personalized feedback reported significantly higher levels of feedback satisfaction.

Research Question 1

Will students receiving personalized feedback outperform those receiving generalized feedback with respect to their final grade?

It was discovered that the personalized feedback condition failed to improve the academic achievement over those receiving generalized feedback. One of the potential reasons for this lack of improvement in academic performance is the breadth of the definition regarding personalized feedback. The current study considered personalized feedback as individualized information (differentiated according to the learner's own characteristics or actions) communicated to a learner for the intended purpose of improving learning. This definition is obviously wide-ranging and without careful consideration opens one to inappropriate comparisons within the greater body of research. As instances of LA-based feedback systems are numerous, it was the intention of this paper to compare SARA with highly analogous LA-based systems that administered their feedback in a similar manner.

One of the most commonly cited examples within the LA field is Course Signals at Purdue (Arnold & Pistilli, 2012). Course Signals has demonstrated its ability improve academic achievement through the delivery of highly streamlined personalized feedback in the form of a 'traffic light' dashboard. This form of feedback, while similar in principle, may operate in a dissimilar way to that of the present study, calling into question its suitability as a direct comparison. Additionally, Arnold and Pistilli (2012) determined the success of Course Signals by comparing their current year to a cohort of students from the previous year. The present study opted to approximate an experimental design with treatment and control groups running simultaneously. When compared to Arnold and Pistilli (2012), the difference in methodologies lessens the confidence one can have in both their results and their applicability as a comparison.

Unfortunately, highly analogous systems to SARA are in short supply. Only Wright et al.'s (2014) investigation of the effectiveness of E^{2}Coach at the University of Michigan delivered personalized written feedback in a similar way. Unique to their study, however, was they way in which they defined success. Students were considered a success if their achieved grade exceeded their predicted grade (as determined by the E^{2}Coach algorithm). Within the present paper, success of the personalized feedback was determined by whether students were able to achieve, on average, higher final grades compared to the generic group in an absolute sense. This subtle distinction may have partially contributed to the negative finding in the present paper. Although not an example of an LA system, Gallien and Oomen-Early (2008) also assessed personalized text-based feedback. Their results showed a substantial benefit for those students who received personalized feedback compared to generalized feedback. However, a few key distinctions suggest why the situation in the present paper is different. First, within Gallien and Oomen-Early's (2008) study, messages were personalized exclusively with reference to student performance on tests and assignments. The instructor did not differentiate feedback according to any survey, percentile rank, or demographic membership data. Further, all feedback was task-focused, the importance of which will be discussed at length later. Second, though the feedback medium was the same between studies, feedback in Gallien and Oomen-Early's (2008) study was confined to corrective, informative, and Socratic forms. By contrast, SARA delivers a much greater range of feedback types, such as advice on time management, how to use course resources, and study tips. SARA also uses some feedback types that have less support within the literature, such as norm-referenced feedback and praise. Lastly, in their study the observed difference for the personalized feedback group was over three standard deviations above those in the generalized condition. The benefits of computer administered feedback interventions are generally much more modest (Azevedo & Bernard, 1995), suggesting that Gallian and Oomen-Early's (2008) study may not be an appropriate case with which to compare.

Concerns relating to the ethical compromises in SARA's development may also be somewhat responsible for the negative findings. SARA's operation was a real-world phenomenon that existed outside of ideal laboratory conditions. Thus, the students subjected to SARA's treatments were working for real grades. To help ensure that students in the personalized group were not being given an unreasonable advantage, ethical review required that the experimental conditions remain highly similar. It should be readily apparent to the reader that one group could easily be provided with information that could drastically improve their performance within the course over the other group. For example, the personalized group could have been provided with information with direct relevance to assessments within the course not afforded to the generalized group. Consider such a situation where students with a specific set of characteristics in the personalized group are likely to benefit substantially from some feedback intervention to the exclusion of those within the control group. To provide one group with such an advantage in a situation where students are working for real grades would be considered unethical. Hence, this risk of over-privileging one group had to be kept in mind during development. This cautionary approach might have resulted in groups that were too similar to create a meaningful difference, especially as it pertains to key task-related information. This suggestion is of course speculation, but considering the ethical compromises made in the execution of this experiment, must be mentioned.

Another potential reason for the absence of a positive effect of achievement pertains to SARA's feedback practices and its lack of adherence to a strong theoretical background.

Specifically, there are shortcomings that need to be discussed regarding feedback personalization, as well as feedback theory more generally.

SARA's ability to personalize feedback could be argued to be superficial, when it pertains to improving learning. In many of the differentiated reports provided by SARA, the majority of the text within messages is the same across groups. Further, in many cases the students in the generalized feedback group receive one of the notes that was otherwise seen by one of the subgroups within the personalized group. For example, prior to the February break students in the generalized group received a copy of one of the messages that the personalized group received – which was differentiated according to the amount of study time students said they allocated to the course (See Appendix C for example). While other messages are more highly differentiated and not guilty of this practice (See Appendix D for example), it is perhaps unsurprising that no differences in academic achievement were observed given the similarity of the personalized feedback emails.

Perhaps the greatest reason SARA's personalization failed to improve academic achievement outcomes lies in a lack of adherence to good feedback practices. Numerous authors have provided theoretical frameworks for the proper operation and administration of feedback, many of which were discussed in the earlier portion of Chapter Two. The main areas in which SARA deviates from good feedback practice are: lack of task specific feedback, use of praise, and the improper use of norm-referenced feedback, all of which will be addressed in greater length.

As touched on briefly, one of the most salient criteria for ensuring the success of feedback in improving learning is the task specificity of the feedback (Kluger & DeNisi, 1996). A common feature in virtually all theoretical attempts in explaining successful feedback has been the central importance of keeping feedback task-focused. This directive means that the information conveyed to the learner should pertain specifically to closing the gap between their current level of performance, and that of ideal performance. For example, consider a situation where a young student is learning how to add numbers of increasing digit span, and is struggling with the concept of ‘carrying over’ numbers. In this case, task-specific feedback could range from re-teaching the concept of place value, practicing through questions with an instructor, or perhaps having the student talk aloud their thinking as they complete a question and correct misunderstanding when it arises. What would not constitute task-specific feedback, however, would be to deliver the student a message instructing them to allocate more mental energy to the task in the future, a strategy that SARA sometimes employs.

Tying into the argument concerning SARA’s lack of task specificity is Mason and Bruning’s (2001) feedback decision-making model for computer-based instruction. SARA’s approach shares some aspects of Mason and Bruning’s model. For instance, SARA differentiates feedback according to student ability and achievement quite regularly and does so using a variety of important performance indices. However, Mason and Bruning’s model is built around task specificity, and this marks the departure between these two approaches. A task, from the perspective of Mason and Bruning’s model, is a singular exercise of learning that ranges in complexity. In the case of SARA, it is difficult to identify a task in a similar way. In a more abstract sense, one could consider the course, as whole, to be the task in which SARA is aiding students. But this approach breaks down almost immediately when applied to Mason and Bruning’s model, as SARA’s highly regimented application of feedback does not bear any resemblance to the protocols in the model.

Yet another potential reason for the lack of improvement in academic achievement lies in SARA's modest use of social comparison feedback. For example, SARA routinely informs students of their performance within the course relative to the class average. As has been mentioned at length, social comparison feedback is largely scorned within the broader feedback literature (Shute, 2008). Granted, there are situations where social comparison feedback has resulted in improved academic achievement; Kim, et al.'s (2016) study provided students with access to a dashboard that featured a series of norm-referenced graphs reporting course performance. Students given access to the dashboard were found to have performed greater than those not given access. However, students in the Kim, et al.'s (2016) study were informed that they would be receiving norm-referenced final grades, which may have motivated participating students to maximize their use of the dashboard beyond what would otherwise have been observed. There was no such motivating factor within the present study; final marks within the course were absolute measures of performance. In another instance, Delaval et al., (2015) found that students given access to personalized norm-referenced feedback outperformed the control group in terms of academic achievement. However, this was only found to be the case when academic procrastination was taken into account, suggesting that this type of feedback only benefitted highly motivated students with low levels of academic procrastination.

Despite these examples, the totality of the feedback literature recommends against the use of norm-referenced feedback (Shute, 2008). Kluger and DeNisi's (1996) feedback intervention theory states that feedback should be as task-oriented as possible in addressing student misconceptions. This demand for task-oriented feedback is done to ensure that the pupil's attention remains fixed on the exercise. Under this theory, norm-referenced feedback takes student attention away from the task and focuses it up the hierarchy of processing towards the self. This movement pulls valuable cognitive resources away from where learning takes place and is likely to result in reduced performance.

Kluger and DeNisi’s (1996) feedback intervention theory also provides some insight into SARA’s use of praise within personalized feedback messages. Praise, as defined by Baumeister, Hutton, and Cairns (1990), is favourable interpersonal feedback expressing approval. This type of feedback is relatively common within SARA messages. For example, SARA messages not only inform students of their relative standing in the class, they congratulate students on these successes. While praise might seem like an appealing method of feedback, the research is fairly clear regarding its efficacy: as it pertains to feedback intervention theory (Kluger & DeNisi, 1996), praise functions in a similar way as norm-referencing, moving the student’s locus of attention up the hierarchy to deal with evaluations of the self. This movement again, takes cognitive resources away from the task at hand and focuses on evaluating the self. Baumeister, et al. (1990) acknowledged that almost everyone enjoys receiving praise, but their experiments showed conclusively that praise leads to reduced effort, introduces unnecessary pressure for continued good performance, and draws attention away from the task at hand. All things considered, SARA’s use of praise is not extensive in its application. However, from the perspective of the aforementioned theorists, it is likely used too liberally and may account for why the personalized feedback condition failed to improve learning.

Research Question 2

Will students receiving personalized feedback report higher levels of feedback satisfaction than those receiving generalized feedback?

Unlike the case of academic achievement, SARA was able to make a meaningful positive difference with regard to feedback satisfaction. Important to note however, this positive effect was observed to be decreasing over time.

Regarding satisfaction more generally, the opinion within the feedback literature appears to be that personalized feedback is so self-evidently more satisfactory than generic feedback that the question is seldom addressed (Burr et al., 2013; Cramp, 2011; Laryea, 2013; Lipnevich & Smith, 2009). However, this notion comes out of the educational literature that precedes the advent of LA and is likely not representative of situations dealing with computer-generated personalized feedback. It is suspected that what constitutes personalized feedback within the LA sphere may only be ostensibly so in the eyes of the typical in-person instructor.

Unfortunately, little research examining feedback satisfaction with LA-based systems exists. Jivet et al.’s (2018) review cites three studies that featured satisfaction as a primary variable by which an LA tool was evaluated. Loboda, Guerra, Hosseini, and Brusilovsky (2014) examined students’ satisfaction of use with a social progress visualization tool, that delivered personalized feedback to students based upon both their own and their peers’ performance in a learning task. Results showed that students were generally highly satisfied with the system and responded favourably to many of its features permitting comparisons between oneself and past performances and with one’s peers. In Kim, et al.’s (2016) study on a LA dashboard, it was discovered that a slight positive correlation between dashboard usage and learning achievement existed. Further, they found that learners who used the dashboard more often reported lower levels of satisfaction with its use. Ruiz-Calleja, Dennerlein, Ley, and Lex (2016) also found that their participants were generally satisfied with the feedback provided by their highly visual LA dashboard.

In the above mentioned studies, participants were given access to a selection of graphs displaying normative information on course performance indices. By contrast, students in the present study were delivered written messages personalized according to a selection of characteristics. Despite this difference in feedback mediums, students within the present study and those previously mentioned, all appeared to have been satisfied with the personalized feedback provided by their respective systems. Unfortunately for the present study, the most highly analogous LA-based system to SARA — E^2Coach at the University of Michigan — was not evaluated with regards to feedback satisfaction (Wright et al., 2014).

The aforementioned criticism that SARA's feedback was only superficially personalized appears to conflict with the findings on feedback satisfaction. One might expect that if SARA's personalized feedback was only superficial, it likely would not register a difference in feedback satisfaction. However, because the personalized feedback was received more favourably than the generalized feedback, one must conclude that the feedback is indeed making a positive difference, at least with respect to some aspect of its functioning. One possible reason for this positive finding, considering the negative finding on academic achievement, is that the threshold level of personalization required to be considered more favourable in the eyes of the recipient is far lower than that required to produce measurable improvements in learning. This complication is perhaps compounded by the ethical compromises made to ensure the experimental groups remained highly similar. Ultimately, this compromised state might have resulted in a situation where SARA's personalized feedback was better able to communicate its message to its intended audience, but courtesy of the imposed ethical limitations, was unable to deliver the learning specific information needed to improve achievement. The result is a system that has succeeded in its ability to appeal to the listener but does not know what to say in terms of improving learning performance.

Yet another possibility is that the personalized component of SARA's feedback has managed to improve Winstone et al.'s (2017) term of proactive recipience or the desire to actively engage in feedback. By addressing specific student characteristics, SARA may be improving outcomes related to Winestone's (2017) SAGE model. With respect to their model, SARA's personalized feedback may be assisting students in their ability to self appraise, regulate, and goal-set. These effects may be small enough that they failed to produce any improvements in academic achievement, though their appreciation and satisfaction was recorded within the satisfaction survey.

From the perspective of Narciss and Huth's (2004) feedback framework for tutoring feedback, SARA is in fact successfully addressing some important dimensions. SARA feedback is highly aware of the instructional context, actively linking feedback messages to important course activities and outcomes. For example, students are kept well aware and are encouraged to take full advantage of additional learning opportunities directly related to major course assessments, including mock exams and structured study sessions. Further, SARA works very well to measure individual learner characteristics and use this information to inform its feedback. Among these traits include student self-efficacy and prior academic performance, both of which provide great utility in personalizing feedback (Shute, 2008). Where SARA's shortfall lies is primarily in Narciss and Huth's (2004) feedback quality. Specifically, these short-comings lie within the contents of feedback. This aspect of the theory is concerned with both evaluation and information. Evaluation includes that which shows the learner their gap between current and ideal performance, and information includes the task-specific advice that will bridge the gap in performance.

Limitations and Future Research Directions

Though the present study did not suffer a limitation regarding sample size, attention should be drawn to the issue relating to data loss. Participation rates with SARA's entry and exit surveys were very low. As was observed in Chapter 4, less than a third of participants completed the feedback satisfaction survey. It may be the case that these select students were distinct in some way and not fully representative of the greater population. In any case, future research should work to find ways of encouraging higher participation rates.

The present study was done on an existing dataset. As such, there was no opportunity to introduce additional measures to the entrance or exit surveys that would span the length of the experiment. There was a shortage of pre-treatment variables associated with both the outcome variables of interest. Certain commonly measured constructs have been known to be associated with academic achievement. For example, conscientiousness, openness to experience, and agreeableness have all been found to possess a significant relationship to achievement in education (Carthy, Gray, McGuinness, & Owende, 2014). These and perhaps more variables would have been helpful in creating higher quality matches for analysis.

Within the LA sphere there is a great breadth of feedback forms on offer, many of which do not resemble the feedback within the present study. This discrepancy limited the extent to which one was able to predict the effect of the feedback intervention within the present study. For example, though both are the same in principle (i.e. they are offering feedback differentiated by student demographics and activity), how similar can the 'traffic light' function of Course Signals (Arnold & Pistilli, 2012) be to that of the text-based messages provided by E^{2}Coach

(Wright et al., 2014)? In light of the results of the present study, it is suspected that these varying forms of feedback may have little in common. This mismatch between the available literature and the operation of the LA feedback system may have led to the over confidence of the present paper's primary hypothesis. It is the hope that future research endeavours take greater care in selecting appropriate LA systems for comparison during evaluation.

In a similar vein, SARA combined multiple types of feedback interventions ranging from praise to norm referencing. It was difficult to predict how the unique combination of these varying forms of feedback would work together, and especially so in combination with examining the effectiveness of personalizing feedback. Because of this, the current paper functions more as an evaluation of SARA, rather than a serving as a major contribution to feedback theory and practice. That is not to say that this is a closed case, however. SARA was found to be succeeding with regards to some of its functionality, at least with respect to feedback satisfaction. In light of this, the primary takeaways from the present study should be the following: 1) Personalized computer-generated feedback can improve feedback satisfaction over generic feedback 2) Personalized computer-generated feedback will not necessarily result in improvements in learning achievement, and 3) To ensure a proper evaluation of feedback, adherence to the theoretical underpinnings of feedback must be maintained.

Gasevic, et al.'s (2015) article, made the insightful point that the LA world appears to have been working in near isolation of good pedagogical practice. They went on to caution future developers to mesh good learning principles with these data driven enterprises. This same sentiment is expressed here. Though SARA was able to influence participants in terms of their feedback satisfaction, the lack of improvement in learning achievement is believed to be attributed to the mismatch between SARA's practices and the feedback literature. Future research should bear this in mind in the development of other personalized text-based feedback systems.

Regarding the significant effect of feedback satisfaction, it is encouraging to see that SARA is succeeding in providing personalized feedback that is more satisfactory than its generic feedback. However, the direct cause of this is unknown, as SARA combines multiple distinct forms of feedback in its operation. Future research should investigate which specific features of personalization are responsible for the observed differences in feedback satisfaction. Further, one must bear in mind that the gap between personalized and generic feedback was closing over time. Future research with SARA specifically, should investigate this phenomenon and 'course correct' the problem, as properly addressing this issue may shed new light on how future LA-based systems can improve their own user experience and feedback satisfaction.

Conclusion

The history of feedback research is one of mixed results and seemingly contradictory findings (Shute, 2008). Less controversial, is the role of personalized feedback, which is often seen as superior to generic feedback both in terms of theory and in practice. However, it has been made plain in the present paper that LA-based feedback systems appear to be an exception. Results of the present study failed to support the proposed hypotheses based on the current literature; they have however contributed meaningfully to the growing body of research in LA. The findings drawn from this study suggest that personalizing feedback will not by itself produce positive effects on academic achievement. To produce increases in learning, feedback (whether personalized or not) must be well aligned with well-established pedagogical theory. In the present paper, it is suspected that the negative findings on personalized feedback are due to such a misalignment.

It is important to note that despite the lack of improvement in academic achievement, student perceptions of the personalized feedback were positive. Results showed that students consistently reported greater levels of mean feedback satisfaction when receiving personalized feedback. This finding suggests that the system is at least partially aligned with important aspects of feedback theory pertaining to the instructional context and adapting to individual student characteristics.

Recent commentary in LA research suggested that the discipline was too boastful of its potential benefits, relied too heavily on user perceptions for evaluations, and was poorly aligned with proper pedagogical theory (Gasevic, et al., 2015; Viberg et al., 2018). The system within the present paper may be such an example.

References

(2011). *1st International Conference on Learning Analytics and Knowledge.* Banff, Alberta, Canada: ACM.

Abadie, A., Herr, J. L., Imbens, G., & Drukker, D. M. (2004). NNMATCH: Stata module to compute nearest-neighbor bias-corrected estimators. Chesnut Hill, MA, USA: Boston College Department of Economics.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology*, *62*(2), 148–156.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, (April 2012), 267. https://doi.org/10.1145/2330601.2330666

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, *13*(2), 111–127.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238.

Banica, L., & Hagiu, A. (2015). Big data in business environment. *Scientific Bulletin-Economic Sciences*, *14*(1), 79–86.

Baumeister, R. F., Hutton, D. G., & Cairns, K. J. (1990). Negative effects of praise on skilled performance. *Basic and Applied Social Psychology*, *11*(2), 131–148.

Bettinger, E., Doss, C., Loeb, S., Rogers, A., & Taylor, E. (2017). The effects of class size in online college courses: Experimental evidence. *Economics of Education Review*, *58*, 68–85.

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *The Stata Journal*, *9*(4), 524–546.

Bodily, R., & Verbert, K. (2017). Trends and issues in student-facing learning analytics reporting systems research. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 309–318. https://doi.org/10.1145/3027385.3027403

Brynjolfsson, E., Hitt, L., & Kim, H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? In *International Conference on Information Systems 2011, ICIS 2011* (Vol. 1, pp. 541–558).

Burr, S. A., Brodier, E., & Wilkinson, S. (2013). Delivery and use of individualised feedback in large class medical teaching. *BMC Medical Education*, *13*(1), 63.

Carthy, A., Gray, G., McGuinness, C., & Owende, P. (2014). A review of psychometric data analysis and applications in modelling of academic achievement in tertiary education.

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, *48*(3), 5–22.

Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology*, *25*(1), 33–37.

Corea, F. (2017). Big data and insurance: Advantageous selection in european markets. *Data Science Journal*, *16*, 33.

Cramp, A. (2011). Developing first-year engagement with written feedback. *Active Learning in Higher Education*, *12*(2), 113–124. Retrieved from http://search.proquest.com/docview/896195330/

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 231–240).

Delaval, M., Michinov, N., Le Bohec, O., & Le Hénaff, B. (2017). How can students' academic performance in statistics be improved? Testing the influence of social and temporal-self comparison feedback in a web-based training environment. *Interactive Learning Environments*, *25*(1), 35–47.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Thousand Oaks, Calif: Sage publications.

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering Education*, *78*(7), 674–681.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304–317.

Gallien, T., & Oomen-Early, J. (2008). Personalized versus collective instructor feedback in the online courseroom: Does type of feedback affect student satisfaction, academic performance and perceived connectedness with the instructor? *International Journal on E-Learning*, *7*(3), 463–476.

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.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, *59*(1), 64–71.

Gould, J., & Day, P. (2013). Hearing you loud and clear: Student perspectives of audio feedback in higher education. *Assessment & Evaluation in Higher Education*, *38*(5), 554–566.

Greer, J. E., Frost, S., Banow, R., Thompson, C., Kuleza, S., Wilson, K., & Koehn, G. (2015). The student advice recommender agent: SARA. In *UMAP Workshops*.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning*, *22*(3), 335–352.

Hanisch, K. A. (1992). The job descriptive index revisited: Questions about the question mark. *Journal of Applied Psychology*, *77*(3), 377–382.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Huberth, M., Chen, P., Tritz, J., & McKay, T. A. (2015). Computer-tailored student support in introductory physics. *PLoS ONE*, *10*(9), 1–18. https://doi.org/10.1371/journal.pone.0137001

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24.

Imai, K., King, G., & Lau, O. (2009). Zelig: Everyone's statistical software. *R Package Version*, *3*(5). 1–591.

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018). License to evaluate: Preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 31–40).

Jones, I. S., & Blankenship, D. (2017). Learning style preferences and the online classroom. *Research in Higher Education Journal*, *33*, 1–8. Retrieved from http://search.proquest.com/docview/2011266681/

Kane, R., Sandretto, S., & Heath, C. (2002). Telling half the story: A critical review of research on the teaching beliefs and practices of university academics. *Review of Educational Research*, *72*(2), 177–228.

Khalil, M., Taraghi, B., & Ebner, M. (2016). Engaging learning analytics in MOOCS: The good, the bad, and the ugly. *ArXiv.Org*. Retrieved from http://search.proquest.com/docview/2079150371/

Khan, I., & Pardo, A. (2016). Data2U: Scalable real time student feedback in active learning environments. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 249–253).

Kim, J., Jo, I.-H., & Park, Y. (2016a). Effects of learning analytics dashboard: analyzing the relations among dashboard utilization, satisfaction, and learning achievement. *Asia Pacific Education Review*, *17*(1), 13–24.

Kim, J., Jo, I. H., & Park, Y. (2016b). Effects of learning analytics dashboard: analyzing the relations among dashboard utilization, satisfaction, and learning achievement. *Asia Pacific Education Review*, *17*(1), 13–24. https://doi.org/10.1007/s12564-015-9403-8

King, G., Ho, D., Stuart, E. A., & Imai, K. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8).

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. *Copy at Http://J. Mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, *378*.

Kluger, A., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword on JSTOR. *Current Dirrections in Psychological Science*, *7*(3), 67–72. Retrieved from https://www-jstor-org.gate2.library.lse.ac.uk/stable/20182507?seq=1#metadata_info_tab_contents

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–282.

Kulhavy, R. (1977). Feedback in written instruction. *Review of Educational Research*, *47*(2). Retrieved from http://search.proquest.com/docview/1290958872/

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology*, *68*(5), 522–528.

Kuo, C.H., Peng, J.W., & Chang, W.C. (2014). Hanzi handwriting acquisition with automatic feedback. In *Proceedings of the Fourth International Conference on learning analytics and knowledge* (pp. 261–262). ACM.

Laryea, S. (2013). Feedback provision and use in teaching and learning: A case study. *Education + Training*, *55*(7), 665–680.

Lee, C. H., & Yoon, H. J. (2017). Medical big data: promise and challenges. *Kidney Research* *and Clinical Practice*, *36*(1), 3.

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *15*(4), 319–333.

Loboda, T. D., Guerra, J., Hosseini, R., & Brusilovsky, P. (2014). Mastery grids: An open source social educational progress visualization. In *European conference on technology enhanced learning* (pp. 235–248).

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901–918.

Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, *15*(3), 149–163.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Mangaroska, K., Sharma, K., Giannakos, M., Trætteberg, H., & Dillenbourg, P. (2018). Gaze insights into debugging behavior using learner-centred analysis. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 350–359).

Markowitz, N., & Renner, K. E. (1966). Feedback and the delay-retention effect. *Journal of Experimental Psychology*, *72*(3), 452.

Mason, B. J., & Bruning, R. (2001). Providing feedback in computer-based instruction: What the research tells us. *Retrieved February*, *15*(August), 2007.

Means, B., Toyama, Y., Murphy, R. F., & Baki, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 115*(3).

Merry, S., & Orsmond, P. (2008). Students' Attitudes to and Usage of Academic Feedback Provided via Audio Files. *Bioscience Education E-Journal, 11*(1), 1–11.

More, A. J. (1969). Delay of feedback and the acquisition and retention of verbal materials in the classroom. *Journal of Educational Psychology*, *60*(5), 339–342.

Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology*, *20*(1), 32–50.

Mory, E. (1992). The use of informational feedback in instruction: Implications for future research. *Educational Technology Research and Development*, *40*(3), 5–20.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. *Instructional Design for Multimedia Learning: Proceedings of the 5th International Workshop of Sig 6 Instructional Design of the European Association for Research on Learning and Instruction (Earli), June 27-29, 2002 in Erfurt*, 181. https://doi.org/10.1042/CS20100475

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and selfregulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Ochoa, X., Dominguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and lowcost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 360–364).

Parvez, S. M., & Blank, G. D. (2008). Individualizing tutoring with learning style based feedback. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5091, pp. 291–301).

Planar, D., & Moya, S. (2016). The effectiveness of instructor personalized and formative feedback provided by instructor in an online setting: Some Unresolved issues. *Electronic Journal of E-Learning*, *14*(3), 196–203.

Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development*, *39*(4), 27–32.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology*, *20*(4), 444–450.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. Version 3.4. 2. Released September 28, 2017.

Ruiz-Calleja, A., Dennerlein, S., Ley, T., & Lex, E. (2016). Visualizing workplace learning data with the SSS dashboard. In *CrossLAK* (pp. 79–86).

Schmidt, M., Mousavi, A., Squires, V., Wilson, K. (2018). Assessing the effectiveness of automated personalized feedback in an undergraduate biology course. In *Proceedings of Hawaii International Conference Science, Technology & Engineering, Arts, Mathematics & Education* (pp. 1–17).

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 352–359.

Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology*, *17*(1), 78–82.

Shuell, T. J. (1986). Cognitive conceptions of learning. *Review of Educational Research*, *56*(4), 411–436.

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

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In *ACM International Conference Proceeding Series* (pp. 4–8).

Siemens, George. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400.

Siemens, George, & d Baker, R. S. J. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254).

Siemens, George, & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30.

Skinner B. F. (Burrhus Frederic), 1904-1990. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students*

Placed at Risk (JESPAR), *22*(3), 178–184.

StataCorp, L. P. (2017). Stata statistical software: Release 15.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331–347.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339–355.

Sturges, P. T. (1972). Information delay and retention: Effect of information in feedback and tests. *Journal of Educational Psychology*, *63*(1), 32–43.

Sullivan, H. J., Baker, R. L., & Schutz, R. E. (1967). Effect of intrinsic and extrinsic reinforcement contigencies on learner performance. *Journal of Educational Psychology*, *58*(3), 165–169.

Tan, J., Yang, S., Koh, E., & Jonathan, C. (2016). Fostering 21st century literacies through a collaborative critical reading and learning analytics environment: user-perceived benefits and problematics. In *Proceedings of the Sixth International Conference on learning analytics & knowledge* (Vol. 25-29-, pp. 430–434). ACM.

Tanes, Z., Arnold, K. E., King, A. S., & Remnet, M. A. (2011). Using Signals for appropriate feedback: Perceptions and practices. *Computers and Education*, *57*(4), 2414–2422. https://doi.org/10.1016/j.compedu.2011.05.016

Teasley, S. (2017). Student facing dashboards: One size fits all? *Technology, Knowledge and*

Learning, *22*(3), 377–384.

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). Formative assessment and learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 205–209).

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, *47*, 157–167.

Thorndike, E L. (1933). A proof of the law of effect. *Science*, *77*(1989), 173–175.

Thorndike, Edward L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Van-Dijk, D., & Kluger, A. N. (2001). Goal orientation versus self-regulation: Different labels or different constructs. In *16th annual convention of the Society for Industrial and Organizational Psychology, San Diego, CA*.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110.

Watson, J. B. (1913). Psychology as the behaviourist views it. *Psychological Review*, *20*(2), 158–177.

West, J., & Turner, W. (2016). Enhancing the assessment experience: improving student perceptions, engagement and understanding using online video feedback. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2014.1003954

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, *52*(1), 17–37.

Wise, A. F., Zhao, Y., & Hausknecht, S. N. (2014). Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics*, *1*(2), 48–71.

Wright, M. C., McKay, T., Hershock, C., Miller, K., & Tritz, J. (2014a). Better than expected: Using learning analytics to promote student success in gateway science. *Change: The Magazine of Higher Learning*, *46*(1), 28–34.

Appendix A: Feedback Satisfaction Survey

Feedback Satisfaction Survey for September 2016, January 2017, and September 2017 course offerings.

Your NSID and survey feedback will remain confidential. Your instructors will never see individual data.

		Yes	No	N/A
1.	Did you appreciate receiving your weekly note from SARA?	○	○	○
2.	Did you find that your weekly note from SARA was personalized to your academic situation?	○	○	○

To what extend do you agree or disagree with the following statements:

		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	N/A
3.	My weekly note increased my sense of belonging in the University community.	○	○	○	○	○	○
4.	My weekly note was encouraging with respect to my academic situation.	○	○	○	○	○	○
5.	My weekly note was a good reminder of my performance in Biology 120.	○	○	○	○	○	○

Note: Survey was administered electronically following completion of the course and was nested within the larger Biology 120 Student Resource Exit Survey.

Feedback Satisfaction Survey for January 2018 course offering.

Your NSID and survey feedback will remain confidential. Your instructors will never see individual data.

To what extend do you agree or disagree with the following statements:

		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	N/A
1.	I appreciated receiving my weekly note from SARA.	○	○	○	○	○	○
2.	My weekly note was personalized to my own academic situation.	○	○	○	○	○	○
3.	My weekly note increased my sense of belonging in the University community.	○	○	○	○	○	○
4.	My weekly note was encouraging with respect to my academic situation.	○	○	○	○	○	○
5.	My weekly note was a good reminder of my performance in Biology 120.	○	○	○	○	○	○

Note: Survey was administered electronically following completion of the course and was nested within the larger Biology 120 Student Resource Exit Survey

Appendix B: Estimating Treatment Effects

To estimate treatment effects in Stata 15.1

#In both cases the Mahalanobis distance is selected as the default distance metric and need not be specified.

#The average treatment effect can be obtained using the following code:

```
. teffects nnmatch (outcome.variable covariate.one covariate.two
covariate.three) (grouping.variable)
```

#The outcome.variable refers to your continuous outcome variable of interest. Covariates covariate.one, covariate.two etc. may be a dichotomous or continuous pre-treatment variables. Lastly, grouping.variable must be a dichotomous variable specifying whether each unit belongs to either the treatment or control group.

#The average treatment effect on the treated can be obtained using the following code:

```
. teffects nnmatch (outcome.variable covariate.one covariate.two
covariate.three) (grouping.variable), (atet)
```

#The only difference between the calculation of the ATE and the ATET is the inclusion of (atet) following the existing code. For additional information see (Abadie, Herr, Imbens, & Drukker, 2004).

Creating the matched dataset in R(R Core Team, 2018) and estimating treatment effects:
#Load R, then install and load the MatchIt package. For additional information see King, Ho, Stuart, and Imai (2011).

```
> library(MatchIt)
```

#Ensure that all variables to be matched on are complete cases with no missing data. MatchIt cannot create matched datasets with incomplete or missing data.
#Use the matchit() function to create a MatchIt object. In the below example it is named "matched.object".
#For treatment.variable fill in your dichotomous treatment variable. var1, var2, and var3 represent the pre-treatment covariates you intend on matching on. Ensure your dataset is selected and specify the distance measure as "mahalanobis".

```
> matched.object <- matchit(treatment.variable ~ var1 + var2 +
var3…, data = your.data, method = "mahalanobis"
```

```
#To check the balance of your newly matched dataset use the
summary() function on your MatchIt object.

> summary(matched.object)

#Balance may also be assessed by looking at the diagnostic plots
produced from the plot() function.

> plot(matched.object)
#To generate the average treatment effect on the treated (ATET)
the MatchIt object must be converted back into data using the
match.data() function.

> matched.data <- match.data(matched.object)

#The package Zelig is then loaded to generate treatment effects
(Imai, King, & Lau, 2009).

> library(Zelig)

#The Zelig object – zelig.object – is created using the zelig()
function. Outcome.variable refers to the outcome variable of
interest. For model, specify “ls” for least squares regression
for continuous dependent variables. Specifying “control” within
```

```
the matched.data function extracts only the matched control
units.

> zelig.object <- zelig(outcome.variable ~ treatment.variable +
var1 + var2 + var3…, model = “ls”, data =
matched.data(matched.object, “control”))

#Now we use the control group’s estimated coefficients and
combine them with the covariate values set to the treated units.
#This is first done using the setx() function. Be sure to
specify “treat” within the data argument.

> explanatory.object <- setx(zelig.object, data =
match.data(matched.object, “treat”), cond = TRUE)

#The sim() function does the imputation. For the x argument, set
x = to the explanatory.object from earlier.

> simulate.object <- sim(zelig.object, x = explanatory.object)

#Use the summary function to observe the results.

> summary(simulate.object)
```

```
#To estimate the average treatment effect overall (ATE)refer
back to the previous example but fit the linear model to the
treatment group instead.

> zelig.object.2 <- zelig(outcome.variable ~ treatment.variable
+ var1 + var2 + var3…, model = “ls”, data =
matched.data(matched.object, “treat”))

#The same procedure is then repeated.
> explanatory.object.2 <- setx(zelig.object.2, data =
match.data(matched.object, “control”), cond = TRUE)

> simulate.object.2 <- sim(zelig.object.2, x =
explanatory.object.2)

> ate.all <- c(simulate.object$qi$att.ev, -
simulate.object.2$qi$att.ev)

#The following codes then provide the point estimate of the ATE,
its standard error, and its 95% confidence interval.

> mean(ate.all)
> sd(ate.all)
> quantile(ate.all, c(0.025, 0.975))
```

Appendix C: Sample Differentiated Feedback Message 1

February Break Message Differentiated by Study Time:

Below Average Study Time:	Enjoy a week without lab! Your TAs are hard at work marking your exams, so you get a week off of labs (not to mention a short week overall). But before you set up for a Netflix binge, think about using that extra free time to catch up on some studying – keeping on top of it now will be a big help in the long run!
Average Study Time:	Enjoy a week without lab! Your TAs are hard at work marking your exams, so you get a week off of labs (not to mention a short week overall). But before you completely check out, remember that it's midterm season right away. A little extra study time now would be perfect for going over your notes, or making sure you're ready for your other classes!
Above Average Study Time:	Enjoy a week without lab! Your TAs are hard at work marking your exams, so you get a week off of labs (not to mention a short week overall). You could use the extra free time to keep up with your studying, but remember that sometimes taking a break can help reduce stress and retain information! Why not splurge on an afternoon nap when your lab would usually be?
Generic Feedback Message:	Enjoy a week without lab! Your TAs are hard at work marking your exams, so you get a week off of labs (not to mention a short week overall). But before you set up for a Netflix binge, think about using that extra free time to catch up on some studying – keeping on top of it now will be a big help in the long run!

Note: Blue highlighted text is common to all messages.

Appendix D: Sample Differentiated Feedback Message 2

Structured Study Sessions Message Differentiated by Predicted GPA:

Predicted GPA > 80%:	Structured Study Sessions are a great opportunity to share your Biology knowledge with your peers, and strengthen your own understanding in the process!
Predicted GPA 80% > 60%:	Structured Study Sessions are a great way to help someone else with the topics you know, and get help for the topics you don’t!
Predicted GPA < 60%:	Structured Study Sessions are a great way to study and learn with and from your peers in a friendly, low-pressure environment.
Generic Feedback Message:	No generic message.

Note: Blue highlighted text is common to all messages.