Human Behavior Modeling and Incentive Design for Online Platforms

by

Mo Zhou

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ken Goldberg, Chair
Assistant Professor Anil Aswani
Assistant Professor Peng Ding

Spring 2018

Human Behavior Modeling and Incentive Design for Online Platforms

Abstract

Human Behavior Modeling and Incentive Design for Online Platforms

by

Mo Zhou

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Ken Goldberg, Chair

This dissertation focuses on human behavior modeling and incentive designs for online platforms, specifically in two application domains: collective intelligence platforms and healthcare intervention platforms. In the last two decades, wide adoption of internet facilitates the emergence of online collective intelligence platforms. In the mean time, the fast developing mobile technology enables real-time healthcare interventions. This dissertation is a step towards implementing human behavior modeling through machine learning and optimization techniques to enhance the efficacy of these online platforms. Case studies show that human behavior modeling combined with smart incentive designs have the potential to improve the performance of such online platforms.

In the first part of this dissertation, I present two collective intelligence platforms: M-CAFE and DebateCAFE. M-CAFE is a mobile-friendly platform that encourages students to check in weekly to numerically assess their course performance, provide textual ideas about how the course might be improved, and rate ideas suggested by other students. For instructors, M-CAFE displays ongoing trends and highlights potentially valuable ideas based on collaborative filtering. M-CAFE is complementary to existing platforms, such as Piazza and stackExchange and allows students to step back and consider their own performance and the performance of their instructors, filling the gap between voluminous transcripts from existing platforms and a one-time-only, end-of-course evaluation. DebateCAFE is an online deliberation platform that introduces a novel incentive mechanism to encourage participants to articulate persuasive arguments on both sides of a complex issue. It uses a combination of uncertainty sampling and collaborative filtering to mitigate bias from selective exposure and highlight/rank the persuasive arguments. Furthermore, DebateCAFE assigns a score to each participant based on the lower of the Wilson scores of the two arguments entered to encourage strong arguments for opposing opinions. Both platforms were built upon the CAFE framework, which is based on Opinion Space.

In the second part of this dissertation, I describe a novel mobile phone application (app) called CalFit and introduce the Discontinuation Prediction Score (DiPS) for non-adherence prediction. CalFit implements important behavior-change features like dynamic goal setting and self-monitoring. Specifically, CalFit uses a reinforcement learning algorithm to generate personalized daily step goals that are challenging but attainable. Two empirical studies with university staff and students indicate that dynamic goal setting is effective in promoting physical activity. Despite the success, I identify that low adherence is a major drawback for mobile-based interventions like ours. Therefore, I further developed the Discontinuation Prediction Score (DiPS), which uses objectively measured past data (e.g., steps and goal achievement) to provide a numerical quantity indicating the likelihood of exercise relapse for the upcoming week for each subject. I present two versions of DiPS using logistic regression and support vector machine methods to demonstrate that DiPS has potential to accurately predict exercise relapse and efficiently allocate resources to improve compliance.

To my parents
Caifeng Xie and Jianlin Zhou.

To my husband
Chen Chen.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **v**

1 Introduction **1**
1.1 Overview 1
1.2 Collective Intelligence Platforms 1
1.3 Healthcare Intervention Platforms 2
1.4 Summary of Contributions 4

I Collective Intelligence Platforms **7**

2 The M-CAFE Platform **8**
2.1 Background and Related Work 8
2.2 M-CAFE 1.0 and 2.0 User Interfaces 11
2.3 Case Studies and Analysis 11
2.4 Conclusion 20
2.5 Future Work 20

3 The DebateCAFE Platform **21**
3.1 Background and Related Work 21
3.2 The DebateCAFE Platform 24
3.3 Case Studies and Analysis 30
3.4 Conclusion 33
3.5 Future Work 34

II Healthcare Intervention Platforms **35**

4 The CalFit Mobile Phone App and the Cal Fitness Studies **36**

4.1 Background and Related Work 36
4.2 The CalFit App 39
4.3 The Cal Fitness Studies 44
4.4 Discussion 58
4.5 Design Implications 60
4.6 Limitations 61
4.7 Conclusion 61
4.8 Future Work 62

5 The Discontinuation Prediction Score (DiPS) 63
5.1 Background and Related Work 63
5.2 Methods 65
5.3 Results 71
5.4 Discussion and Implications 75
5.5 Conclusion 77
5.6 Future Work 78

6 Conclusion 79

A Appendix 80

Bibliography 82

List of Figures

2.1 Screenshots of M-CAFE 1.0 interface. 12
2.2 Screenshots of M-CAFE 2.0 interface. 13
2.3 The space interface of course improvement suggestions with topic tags in M-CAFE 2.0. 14
2.4 Course difficulty ratings for IEOR 170, S15 over 10 weeks 15
2.5 Course difficulty ratings for IEOR 170, S15 and IEOR 170, S16 over 15 weeks 16
2.6 Histogram of the number of peer-to-peer CF ratings for ideas in IEOR 170, S15. 17
2.7 Pie plots of IEOR 115, F15 for topic and module breakdowns 19

3.1 DebateCAFE mobile interface screenshots 25
3.2 Information flow of DebateCAFE 26
3.3 IBA summary plots 29
3.4 Histogram of scaled scores of all participants. 33

4.1 The CalFit app information flow 39
4.2 Screenshots of the CalFit app interface 40
4.3 Screening, randomization, and assessments of study participants for Cal Fitness study 1. 49
4.4 Weekly average and moving average steps for intention-to-treat in Cal Fitness study 1 . 52
4.5 Weekly average and fraction of achieved goals for intention-to-treat in Cal Fitness study 1 52
4.6 Weekly average and moving average steps for per-protocol in Cal Fitness study 1 . . . 54
4.7 Weekly average and fraction of achieved goals for per-protocol in Cal Fitness study 1 . 55
4.8 Weekly average steps for intention-to-treat in Cal Fitness study 2 56

5.1 Simplified example of training data augmentation for DiPS 68
5.2 ROC curve of the predictions for week 20 71
5.3 Simulation outcome in terms of number of adhere participants after 3 months 76

List of Tables

2.1 Participation statistics in different stages of M-CAFE. 15

4.1 Baseline characteristics between the control and intervention groups in Cal Fitness study 1. 51
4.2 Run-in adjusted objectively recorded (using iPhone) physical activity in Cal Fitness study 1. 53
4.3 Fraction of achieved daily step goals in 10 weeks in Cal Fitness study 1. 53
4.4 Baseline characteristics between the control and intervention groups in Cal Fitness study 2 57

5.1 Test AUC for predicting weeks 16-30 on all subjects using the fitted DiPS model. . . . 71
5.2 Confusion matrix of the observed and predicted class for week 20 using the augmented Logistic Regression approach with a threshold of 0.5. 72
5.3 Test AUC for predicting weeks 16-30 on each intervention group using the fitted DiPS model. 73
5.4 Feature importance for the fitted augmented Logistic Regression model for week 20. . 74

A.1 Self-reported physical activity scores and barriers to being active of the two groups pre and post the Cal Fitness study 1. 80

Acknowledgments

I am immensely fortunate to spend the past five years here at Berkeley and work with many great researchers. First and foremost, I would like to thank my two advisors, Professor Ken Goldberg and Professor Anil Aswani. Ken has always been inspiring, encouraging and enthusiastic and taught me how to truly do research: from choosing the appropriate topic, reviewing the related literature to designing platforms and conducting empirical studies. From him, I not only learned to think critically in research but also upskilled other professional skills such as communication. I am very grateful to Anil, who has been an advisor, a mentor and a friend. His broad knowledge and open mind encouraged me to challenge the unknown and work out problems in an unfamiliar yet exciting new field with fruitful results.

Next, I would like to thank my committee members: Professor Phil Kaminsky, Professor William Satariano, and Professor Peng Ding for offering insightful and genuine advice throughout the past five years. More importantly, they have all demonstrated to me the standard of a good academician that I will maintain in the future.

I would like to show my gratitude to my exceptional collaborators at Berkeley and UCSF. I deeply appreciate the help Professor Yoshimi Fukuoka offered on my first ever randomized controlled experiment. I am also grateful for all the discussions and learnings I had from the CAFE research group.

My graduate life at Berkeley would not have been so enjoyable without my awesome colleagues: Jiaying Shi, Renyuan Xu, Shiman Ding, Haoyang Cao, Anran Hu, Xu Rao, Tianyi Lin, Nan Yang, Kevin Li, Quico Spaen, Eric Bertelli, Yonatan Mintz, Auyon Siddiq, Cheng Lu, Steward Liu, Junyu Cao, Ying Cao, Meng Qi, and many more. I am very thankful for the laughters we had over the years.

I also want to thank my friends outside Berkeley IEOR who has always supported me on this PhD journey: Mengyi Wo, Mengdan Ma, Cheng Lin, Mingyang Li, Ruoyun Li, Xiexiao Luo, Shiqi Zhang, and many others. Thank you for listening to my complains, patiently calming me down and always be there for me.

Finally, my deeply grateful acknowledgement goes to my family. My parents Caifeng Xie and Jianlin Zhou has always been my haven. They are always with me through the ups and downs. Their patience, encouragement, and wisdom have left an indelible mark on my life. Thanks to my husband Chen, who has made my life complete. Thank you for loving me for who I am. And thank you very much - the past Mo Zhou in the last five years, for believing in yourself, for working hard, for being strong, for never giving up, for chasing your dream, and for living everyday to the fullest.

Chapter 1

Introduction

1.1 Overview

This dissertation presents human behavior modeling and incentive designs for online platforms based on machine learning, data analytics and optimization techniques. Human behavior modeling is coupled with frontend interface designs and evaluated via empirical studies. Specifically, this dissertation describes new analytical approaches and incentive designs in online collective intelligence platforms and mobile-based healthcare intervention platforms.

1.2 Collective Intelligence Platforms

Collective intelligence are shared intelligence that emerges from collaborations and competitions of many individuals. Though the concept of collective intelligence was introduced decades ago, new technologies - in particular the internet, offer new means of communication that connects individuals from all over the world. Nowadays, many such systems, such as Google and Wikipedia, have become a daily essential. Researchers are also developing new platforms and algorithms to incorporate the idea of collective intelligence in new domains: online deliberation, social media, civic engagement and many more.

In this dissertation, I narrow the scope of collective intelligence platforms to focus on ongoing course evaluation (M-CAFE) and online deliberation (DebateCAFE). Both platforms were built upon the Collaborative Filtering and Feedback Engine (CAFE) with design variations that were tweaked for domain needs.

M-CAFE During MOOCs and large on-campus courses with limited face-to-face interaction between students and instructors, assessing and improving teaching effectiveness is challenging. In a 2014 study on course-monitoring methods for MOOCs [188], qualitative (textual) input was found to be the most useful. Two challenges in collecting such input for ongoing course evaluation are insuring student confidentiality and developing a platform that incentivizes and manages input from many students. To collect and manage ongoing ("just-in-time") student feedback while maintaining student confidentiality, we designed the MOOC Collaborative Assessment and Feedback Engine (M-CAFE). This mobile-friendly platform encourages students to check in weekly to numerically assess their own performance, provide textual ideas about how the course might be improved, and rate ideas suggested by other students. For instructors, M-CAFE displays ongoing trends and highlights potentially valuable ideas based on collaborative filtering. We describe case studies with two EdX MOOCs and three on-campus undergraduate course. Results suggest that comparative plots of past ratings, topic tags and peer-to-peer anonymous suggestion evaluations are valuable in promoting credible and diverse course evaluation.

DebateCAFE Existing Online deliberation platforms such as Consider.it and the Deliberatorium explore a variety of approaches to address polarization resulting from internet-based self-selection that can amplify inherent confirmation bias and result in an "echo chamber" where initial biases are reinforced rather than explored and resolved. We developed DebateCAFE, a prototype platform that introduces an incentive mechanism to encourage participants to articulate persuasive arguments on both sides of an issue. It uses a combination of uncertainty sampling and collaborative filtering to mitigate the bias from selective exposure and highlight/rank the most persuasive arguments. In addition, it assigns scores to participants based on the "weaker" of the Wilson scores of the two arguments entered. To evaluate performance, we used the topic of "personal privacy vs. national security" which was widely discussed in the U.S. in Spring 2016 when Apple refused FBI's request to unlock a terrorist's cellphone. Initial results suggest that DebateCAFE can effectively encourage participants to articulate arguments and identify persuasive arguments for both sides of an issue. Participant's score adequately reflects participant's capability of articulating adversarial arguments. Most participants are more capable of providing persuasive arguments for their own position and they tend to give higher ratings to arguments that align with their position. We summarize data and system performance from a preliminary study with 94 participants who entered 170 arguments on both sides and 1754 peer-to-peer ratings of the arguments.

1.3 Healthcare Intervention Platforms

Regular physical activity (e.g., walking or running) is an important factor in preventing the development of chronic diseases like type 2 diabetes, cardiovascular disease, depression, and certain types of cancer [103, 219, 220]. Because of its importance in maintaining good health, the 2008 Physical Activity Guidelines for Americans recommend that adults engage in at least 150 minutes a week of moderate-intensity physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity [203, 219]. However, about 50% of adults in the U.S. [39] are physically inactive. In fact, over 3 million deaths worldwide are attributed to physical inactivity [218].

Given the high prevalence of physical inactivity, it is necessary to develop new cost-effective, scalable approaches to increase physical activity. One promising direction is the use of smartphones in the delivery and personalization of programs that motivate individuals to increase their physical activity. Over 40% of adults worldwide and 77% of adults in the U.S. own a smartphone [156]. Smartphones have powerful computation and communication capabilities that enable the use of machine learning and other data-driven analytics algorithms for personalizing the physical activity programs to each individual. Furthermore, the past several generations of smartphones integrate reliable activity tracking features [7, 38, 64, 85], which make possible the real-time collection of fine-grained physical activity data from each individual.

The second half of this dissertation addresses the problem of designing such an mobile-based physical activity promotion application (app) to improve its efficacy and adherence. We first describe the CalFit app, which is a novel iOS app that sets personalized, adaptive step goals for each individual based on past physical activity performance. Next we discuss the Discontinuation Prediction Score (DiPS), which can be incorporated to physical activity promotion apps to enhance adherence by accurately predicting the probability of exercise relapse.

CalFit Despite the vast number of mobile fitness apps and their potential advantages in promoting physical activity, many existing apps lack behavior-change features and are not able to maintain behavior change motivation. CalFit, on the other hand, implements important behavior-change features like dynamic goal setting and self-monitoring. CalFit uses a reinforcement learning algorithm to generate personalized daily step goals that are challenging but attainable. We conducted two studies with college staff members and students to evaluate the efficacy of the CalFit app in promoting physical activity. Both studies show that the intervention group (receiving personalized goals) had significantly more daily steps than the control group (receiving step goals of 10,000 per day) over 10 weeks.

DiPS Nonadherence is a big challenge in mobile-based lifestyle modification programs. As a result, many past mHealth interventions involve regular in-person counseling sessions besides the mobile intervention to motivate adherence [54, 68, 205]. However, in-person counseling sessions are costly and put a burden on both the participants and the research staff. Past studies have identified that adherence is correlated with self-efficacy, perceived environment, exercise history, health condition, and many more. In this dissertation, we aim to use Logistic Regression to develop a model to predict exercise relapse. We introduce the Discontinuation Prediction Score (DiPS), which uses objectively measured past data (i.e., steps, MVPA, goal achievement) to fit prediction models and provide a score indicating the likelihood of exercise relapse for a specified time horizon (i.e., a week) in the future. Results show that Logistic Regression with augmented training data gives the highest test accuracy of 80% compared to other classification methods and this approach is robust across different weeks and various thresholds. Coefficients from the augmented Logistic Regression indicate that steps data (including run-in average steps, last week steps, average steps over the entire study) and physical activity intensity data are most predicable of an individual's DiPS for the following week. Implementing DiPS in mobile-based physical activity promotion programs has the potential to perform just-in-time intervention and reduce exercise relapse.

1.4 Summary of Contributions

An outline of the contributions of this dissertation can be found below, along with the associated publications.

M-CAFE The MOOC Collaborative Filtering and Feedback Engine (M-CAFE) is the first to automatically collect ongoing course evaluation and scalably identify the most valuable suggestions.
a) **Design** We present the interface and the backend mechanism design of the M-CAFE platform. Both M-CAFE 1.0 and M-CAFE 2.0 are discussed in this dissertation.
b) **Empirical Studies** M-CAFE was used in 5 courses (including 2 MOOC courses and 3 in-person courses). We present empirical results to demonstrate the efficacy of the M-CAFE platform.

This work was performed in collaboration with the California Report Card research group (The CAFE group) at UC-Berkeley. Details of this piece of work is presented in Chapter 2 of this dissertation and was published in:

- Mo Zhou, Alison Cliff, Allen Huang, Sanjay Krishnan, Brandie Nonnecke, Kanji Uchino, Sam Joseph, Armando Fox, and Ken Goldberg. "M-CAFE: Managing MOOC student feedback with collaborative filtering." *Proceedings of the Second (2015) ACM Conference on Learning@Scale, pp. 309-312. ACM, 2015.*

- Mo Zhou, Alison Cliff, Sanjay Krishnan, Brandie Nonnecke, Camille Crittenden, Kanji Uchino, Ken Goldberg. "M-CAFE 1.0: Motivating and Prioritizing Ongoing Student Feedback During MOOCs and Large on-Campus Courses using Collaborative Filtering." *Proceedings of the 16th Annual Conference on Information Technology Education. Chicago, IL. September, 2015.*

- Mo Zhou, Sanjay Krishnan, Jay Patel, Brandie Nonnecke, Camille Crittenden, and Ken Goldberg. "M-CAFE 2.0: A Scalable Platform with Comparative Plots and Topic Tagging for Ongoing Course Feedback." *Proceedings of the 18th Annual ACM Conference on Information Technology Education. Rochester, NY. October 2017.*

DebateCAFE The Debate Collaborative Assessment and Feedback Engine (DebateCAFE) is a novel platform that seeks to mitigate selective exposure bias with an interface that scales to a large number of participants. DebateCAFE collects quantitative feedback to speculate the initial position of participants, encourages participants to enter convincing arguments on both sides of an issue, and rate the persuasiveness of other participants' arguments.
a) **Design** We present the interface and the incentive mechanisms of DebateCAFE so that it mitigates selective exposure and encourages strong arguments on various aspects on complex issues.
b) **Empirical Studies** A study with undergraduate students at UC-Berkeley on the issue of Apple vs. FBI was conducted to evaluate the performance of DebateCAFE. Greater exposure and articulation to adversarial arguments were observed.

This work was performed in collaboration with the California Report Card research group (The CAFE group) at UC-Berkeley. Details of this piece of work is presented in Chapter 3 of this dissertation and was submitted to:

- Mo Zhou, Sanjay Krishnan, Jay Patel, Brandie Nonnecke, Moonhyok Kim, Camille Crittenden, Nicholas Adams, Saul Perlmutter, and Ken Goldberg. "DebateCAFE v1. 0: Incentivizing Articulation and Consideration of Adversarial Arguments." *Proceedings of the 21st ACM Conference on Computer-Supported Cooperative Work and Social Computing.* Submitted.

CalFit CalFit is an iOS mobile application (app) to promote physical activity. It sets personalized, adaptive step goals using the behavioral analytics algorithm (BAA). This is the first mobile app that adopts machine learning techniques to set personalized step goals.
a) **Design** We present the interface and the human behavior modeling algorithm of CalFit. BAA first uses machine learning to construct a predictive quantitative model for each participant based on the historical step and goal data, and then, it uses the estimated model to generate challenging yet realistic step goals in an adaptive fashion by choosing step goals that, based on the estimated model, would maximize future physical activity.
b) **Empirical Studies** We conducted 2 studies to evaluate the performance of the BAA algorithm. The primary aim of these studies was to evaluate the efficacy of the automated mobile phone-based personalized, adaptive goal-setting intervention as compared with the active control with nonpersonalized, steady daily step goals of 10,000 steps. The main outcome measure was the relative change in objectively measured daily steps between the run-in period and 10 weeks.

This work was performed in collaboration with the University of California, San Francisco (UCSF). Details of this piece of work is presented in Chapter 4 of this dissertation and was published in:

- Mo Zhou, Yoshimi Fukuoka, Yonatan Mintz, Ken Goldberg, Philip Kaminsky, Elena Flowers, Anil Aswani. "Evaluating Machine Learning-Based Automated Personalized Daily Step Goals Delivered Through a Mobile Phone App: Randomized Controlled Trial." *JMIR mHealth and uHealth 6, no. 1 (2018): e28.*

- Mo Zhou, Yonatan Mintz, Yoshimi Fukuoka, Ken Goldberg, Elena Flowers, Philip Kaminsky, Alejandro Castillejo, Anil Aswani. "Personalizing mobile Fitness Apps using Reinforcement Learning" *Proceedings of the ACM Workshop on Theory-Informed User Modeling for Tailoring and Personalizing Interfaces (HUMANIZE).*

DiPS Discontinuation Prediction Score is a novel machine learning-based algorithm that predicts the probability of nonadherence in physical activity interventions.
a) **Algorithm** We use logistic regression (LR) and support vector machine (SVM) methods to design two versions of DiPS, which uses each individual's past data (e.g., physical activity duration, physical activity intensity and goal achievement) to assign a numeric value that quantifies their

likelihood of discontinuing physical activity in the upcoming week.
b) **Simulation** The potential utility of DiPS to provide guidance for provision of just-in-time interventions for individuals who are more likely to have an exercise relapse is demonstrated through a simulation in which I compare the cost-effectiveness of different schemes to allocate financial incentives that encourage selected individuals to increase their physical activity.

This work was performed in collaboration with the University of California, San Francisco (UCSF). Details of this piece of work is presented in Chapter 5 of this dissertation and was submitted to:

- Mo Zhou, Yoshimi Fukuoka, Ken Goldberg, Eric Vittinghoff, Anil Aswani. "Real-time Prediction of Future Adherence in Physical Activity Promotion Programs using Machine Learning: the Discontinuation Prediction Score (DiPS)." *Journal of the American Medical Informatics Association.* Submitted.

Other Publications

- Yoshimi Fukuoka, Mo Zhou, Eric Vittinghoff, William Haskell, Ken Goldberg, Anil Aswani. "Objectively Measured Baseline Physical Activity Patterns in Women in the mPED Trial: Cluster Analysis" *JMIR Public Health and Surveillance 4, no. 1 (2018): e10.*

- Sanjay Krishnan, Brandie Nonnecke, Alison Cliff, Angela Lin, Shir Nehama, Jay Patel, Mo Zhou, Laura Byaruhanga, Dorothy Masinde, Maria Elena Meneses, Alejandro Martin del Campo, Camille Crittenden, Ken Goldberg. "DevCAFE: A Customizable Participatory Assessment Platform for Development Interventions." *IEEE Global Humanitarian Technology Conference (GHTC)*. October 2015.

- Anne Miller, Anil Aswani, Mo Zhou, Matthew Weinger, Jason Slagle, Daniel France. "Using Telephone Call Rates and Nurse to Patient Ratios as Indicators of Resilient Performance under High Patient Flows" *Cognition, Technology & Work.* Accepted.

Part I

Collective Intelligence Platforms

Chapter 2

The M-CAFE Platform

2.1 Background and Related Work

Massive open online courses (MOOCs) have received widespread attention and excitement since 2012. But studies find that MOOCs have an extremely high dropout rate, due to limited interaction with instructors and peers, insufficient academic background and personal stress [78, 161, 210, 222, 224]. In the mean time, on-campus courses are experiencing growing class sizes, making individualized student-instructor interaction challenging.

A widely recognized indicator of teaching effectiveness is student interaction with instructors [10, 48, 101]. However, for MOOCs and large on-campus courses it is difficult to provide such interaction [77, 106, 33, 162]. Although most MOOCs and large on-campus courses include some form of end-of-course evaluation, new methodologies offer potential to facilitate student interaction and provide more frequent feedback to instructors.

Student Evaluation of Teaching

Student evaluation of teaching (SET) is widely used to evaluate instructor's teaching effectiveness. SET usually occurs at the end of the semester when students are asked to rate different aspects of the course on a numerical scale (usually from 1 to 10 "Very Low" to "Very High"). Many studies question the validity of naively aggregating quantitative ratings from SET [4, 60, 61, 131, 142, 184]. Stark and Freishtat [185] state that "SET are ordinal categorical variables and comparing an individual instructor's average to department average is meaningless, since there's no reason to believe that the difference between 3 and 4 means the same thing as the difference between 6 and 7 and that the difference between 3 and 4 means the same thing to different students." McCullough and Radson further introduce an alternative method to analyze SET data that uses categorical proportions instead of assigning a score to each category and numerically aggregating the results to robustly evaluate teacher's performance [136]. In contrast, Khong conducted a recent study of 200 students to suggest that the SET is a valid instrument in evaluating teaching effectiveness by measures such as internal consistency and correlations [98]. Surgenor suggests that SET can be valuable to measure dedication to teaching and improvement and to promote quality learning [194].

Furthermore, he identifies the need for easily obtainable feedback on modules. M-CAFE is a tool to help instructors improve teaching effectiveness by encouraging and examining feedback on a weekly basis. To ensure comparable SET responses, M-CAFE displays the average ratings for the past weeks as a benchmark for future ratings.

Student Satisfaction and Perceived Learning

Student satisfaction is a significant predictor of learning outcomes. Studies find that clarity of design, interaction with instructors and active discussion among course participants significantly influence students' satisfaction in online learning [56, 197]. Richardson and Swan further recognized that students with higher perceptions of social presence in the courses had significantly higher scores in perceived learning and perceived satisfaction with the instructor [174]. Therefore, most MOOCs and large courses include some version of discussion forums to collect qualitative input and to facilitate peer interaction and instructor engagement [44, 45, 46]. However, existing forums such as Piazza, stackExchange and Internet Relay Chat can be intimidating to students and instructors when the quantity of text is overwhelming. M-CAFE facilitates student interaction with instructors and among course participants, and provides timely feedback to the instructor, making in-time course design modifications possible. Furthermore, M-CAFE encourages students to assess and track their own motivation, enthusiasm, and performance over the duration of the course.

Collective Intelligence

Online discussion platforms such as Piazza and stackExchange are popular collective intelligence sites. Most MOOCs and large on-campus courses use one or more of these platforms to motivate interaction among peers and between students and instructors [127]. Gelman et al. investigated the emergence of interest-based subcultures in online communities and how they engage a large number of learners [72]. Sajjadi et al. adopt a peer-grading mechanism in a MOOC to explore the effective metric of aggregating peer assessments [169]. Krishnan et al. developed a self-organizing collective system called the Collective Discovery Engine (CDE) to collect insights from a diverse group on how social media can improve learning [112]. Woolley et al. quantify group performance by a collective intelligence factor and suggest that group performance exceeds individual performance [216]. Despite the various crowdsourcing applications, this approach has not been applied to student evaluations. Typical student evaluations involve anonymous individual responses to the questions where students are not allowed to discuss or view each other's response. This mechanism is inefficient because it requires extensive time from the instructors to read each of the responses, and the responses contain many repetitions. M-CAFE aims to overcome these challenges by adopting a collaborative filtering mechanism, where students provide peer-to-peer ratings on suggestions with the interested topic and assign a reputation score to each suggestion. The set of suggestions with the highest reputation are presented to the instructors each week.

Natural Language Processing (NLP)

One way to address the scale issue of qualitative data is to use Natural Language Processing (NLP). Adamopoulos applied two opinion mining tools [149], an orientation analysis mechanism and a sentiment analysis mechanism to course reviews to identify what course features affected the retention rate [1], which identifies popular topics in text, relationships between text and correlation patterns between topics to find insightful patterns in discussion forum text [166]. In the M-CAFE setting, however, insightful ideas may be rare and could vary greatly in content from week to week, making it difficult to infer from word-document structure. In course evaluations, NLP may identify popular topics, but the proposed insights are more important.

Collaborative Filtering (CF)

We explore an alternative approach for qualitative analysis called Collaborative Filtering (CF) [74, 75, 108, 154]. CF is rating-based and relies on the crowd to find insightful items in a large dataset. This is in contrast to prior approaches, which are content-based and rely on mathematical representations of items. CF has commonly been used to recommend books and movies. The idea is to combine subjective evaluations provided by humans to assign a numerical reputation to each item. Often, the reputation values are computed based on a local neighborhood for customized recommendations but similar techniques can also assign global reputations in a "peer-to-peer" approach [174]. Recent studies also demonstrate that social network information can improve accuracy of CF-based recommender systems [223]. For M-CAFE, we utilize peer ratings on each textual idea and combine standard error to solicit feedback with the Wilson metric to compute reputation values and rank ideas. Due to the self-selection nature of user rating, most CF systems suffer from sparse ratings matrix with many null values [223]. The popular usage of a list-based presentation of items to users can be responsible for the problem, in which case, highly rated items are shown on top of the list, enjoying greater exposure. In M-CAFE, we balance the exposure of items by simultaneously providing a subset of items (normally 6) with mixed rankings, permitting the system to collect feedback on all items.

Study Purpose

We developed the MOOC Collaborative Assessment and Feedback Engine (M-CAFE 1.0 and M-CAFE 2.0), a mobile and web-based platform designed to encourage students to check in weekly to quantitatively assess the course, their own performance, provide qualitative ideas about how the course might be improved, and rate ideas from other students. This platform builds on Opinion Space [113] and uses a combination of uncertainty sampling and statistical analysis to quickly and collaboratively identify valuable ideas. M-CAFE is complementary to existing platforms, such as Piazza and stackExchange and allows students to step back and consider their own performance and the performance of their instructors, filling the gap between voluminous transcripts from existing platforms and a one-time-only, end-of-course evaluation. M-CAFE is also independent of the registrar database to maintain student confidentiality.

2.2 M-CAFE 1.0 and 2.0 User Interfaces

We have developed two versions of M-CAFE and are describing the interface changes in the following paragraphs.

Upon entering M-CAFE 1.0 (Figure 2.1a), students are required to register by email and are given the option to provide their age, gender, home country, years of college-level education, and the primary reason for taking the course (Figure 2.1b). Then they rate five quantitative assessment topics (QAT) on a scale of 1 to 10: Course Difficulty, Course Usefulness, Self-enthusiasm, Self-performance and Homework Effectiveness (Figure 2.1c). Students click on mugs (Figure 2.1d) to view their peers' ideas, evaluate how valuable the ideas are on a scale of 1-10 (Figure 2.1e) and suggest new ideas (Figure 2.1f).

Compared with M-CAFE 1.0, the new version, M-CAFE 2.0 (Figure 2.2), adopts a more intuitive interface with a keypad design for ratings. In addition, M-CAFE 2.0 introduces comparative plots displaying the rating history of the current user against the course weekly average for each QAT, enabling users to track their own performance and quickly compare themselves to the course average (Figure 2.2d). As show in Figure 2.3, M-CAFE 2.0 further deploys topic tagging in the discussion phase to bring more structure to the textual suggestions. The list of tags includes Exams, Homework, Labs, Lectures, New Topic, Logistics, Projects and Other.

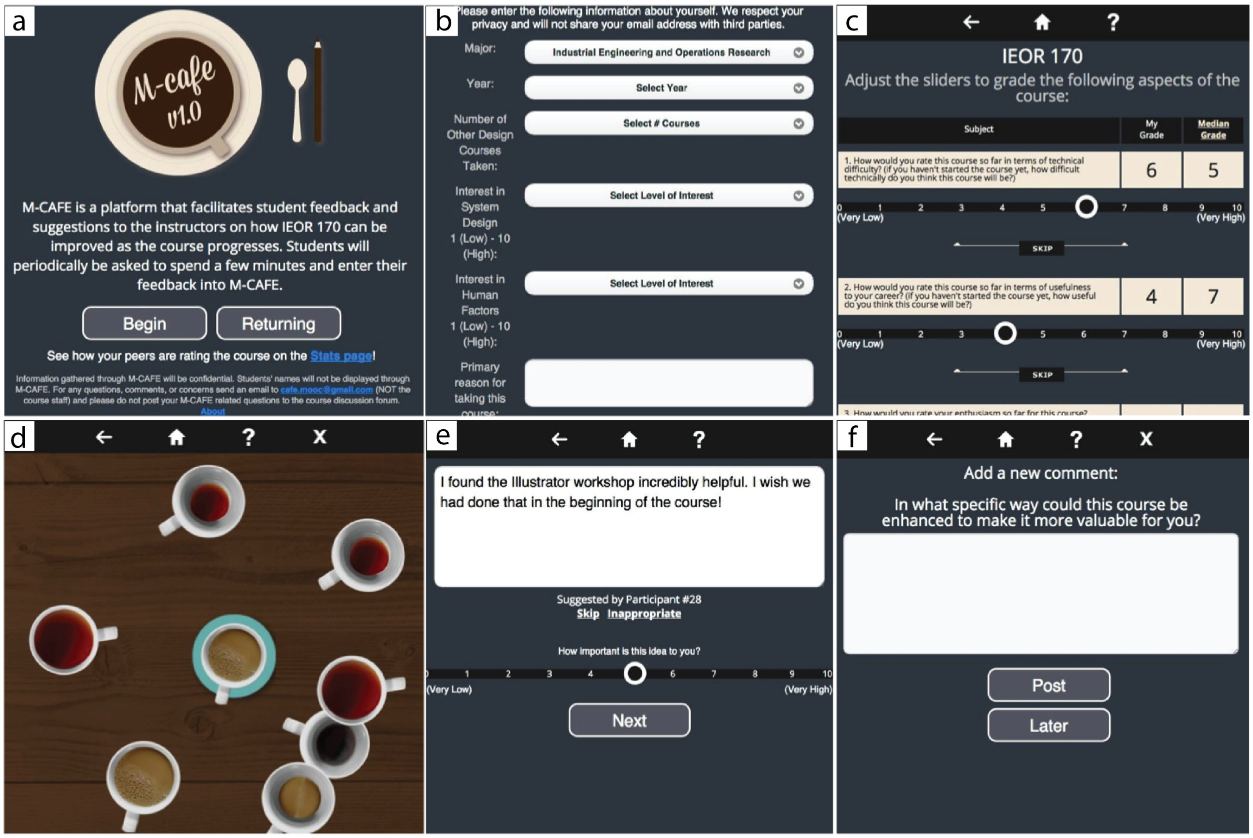
2.3 Case Studies and Analysis

M-CAFE 1.0 has so far been used in three courses: two MOOCs on edX: CS 169.2x and CS 169.1x, and a face-to-face classroom-based undergraduate course - IEOR 170: Industrial Design and Human Factor in Spring 2015 (S15), all offered through UC Berkeley. M-CAFE 2.0 has been used in two UC Berkeley undergraduate in-person courses to investigate its effectiveness: IEOR 115: Commercial Database Systems in Fall 2015 (F15) and IEOR 170: Industrial Design and Human Factor in Spring 2016 (S16). Students were invited to participate at the beginning of the course, and email reminders were sent on a weekly basis. Table 2.1 summarizes M-CAFE 1.0 and 2.0 participation statistics for the courses. M-CAFE is fully confidential and no individual identity is revealed on the platform or to the instructors. Since participation in M-CAFE is voluntary, it inherently suffers from self-selection bias, i.e., students who are more actively involved in the course tend to participate more often in M-CAFE. However, considering that M-CAFE aims to collect valuable feedback for the instructors, we expect the active students to provide more insightful ideas because they are more invested in the course and its outcomes.

Quantitative Analysis Topics

Graph Visualization of QAT Rating Changes

For each quantitative analysis topic, an average score and the associated standard error is computed each week. The changes in ratings over the weeks are then plotted to provide a straightforward



M-CAFE is a platform that facilitates student feedback and suggestions to the instructors on how IEOR 170 can be improved as the course progresses. Students will periodically be asked to spend a few minutes and enter their feedback into M-CAFE.

See how your peers are rating the course on the Stats page!

Information gathered through M-CAFE will be confidential. Students' names will not be displayed through M-CAFE. For any questions, comments, or concerns send an email to cafe.mooc@gmail.com (NOT the course staff) and please do not post your M-CAFE related questions to the course discussion forum.

Please enter the following information about yourself. We respect your privacy and will not share your email address with third parties.

Major: Industrial Engineering and Operations Research

Year: Select Year

Number of Other Design Courses Taken: Select # Courses

Interest in System Design 1 (Low) - 10 (High): Select Level of Interest

Interest in Human Factors 1 (Low) - 10 (High): Select Level of Interest

Primary reason for taking this course:

IEOR 170

Adjust the sliders to grade the following aspects of the course:

Subject	My Grade	Median Grade
1. How would you rate this course so far in terms of technical difficulty? (if you haven't started the course yet, how difficult technically do you think this course will be?)	6	5
2. How would you rate this course so far in terms of usefulness to your career? (if you haven't started the course yet, how useful do you think this course will be?)	4	7
3. How would you rate your enthusiasm so far for this course?		

I found the Illustrator workshop incredibly helpful. I wish we had done that in the beginning of the course!

Suggested by Participant #28

Skip Inappropriate

How important is this idea to you?

Add a new comment:

In what specific way could this course be enhanced to make it more valuable for you?

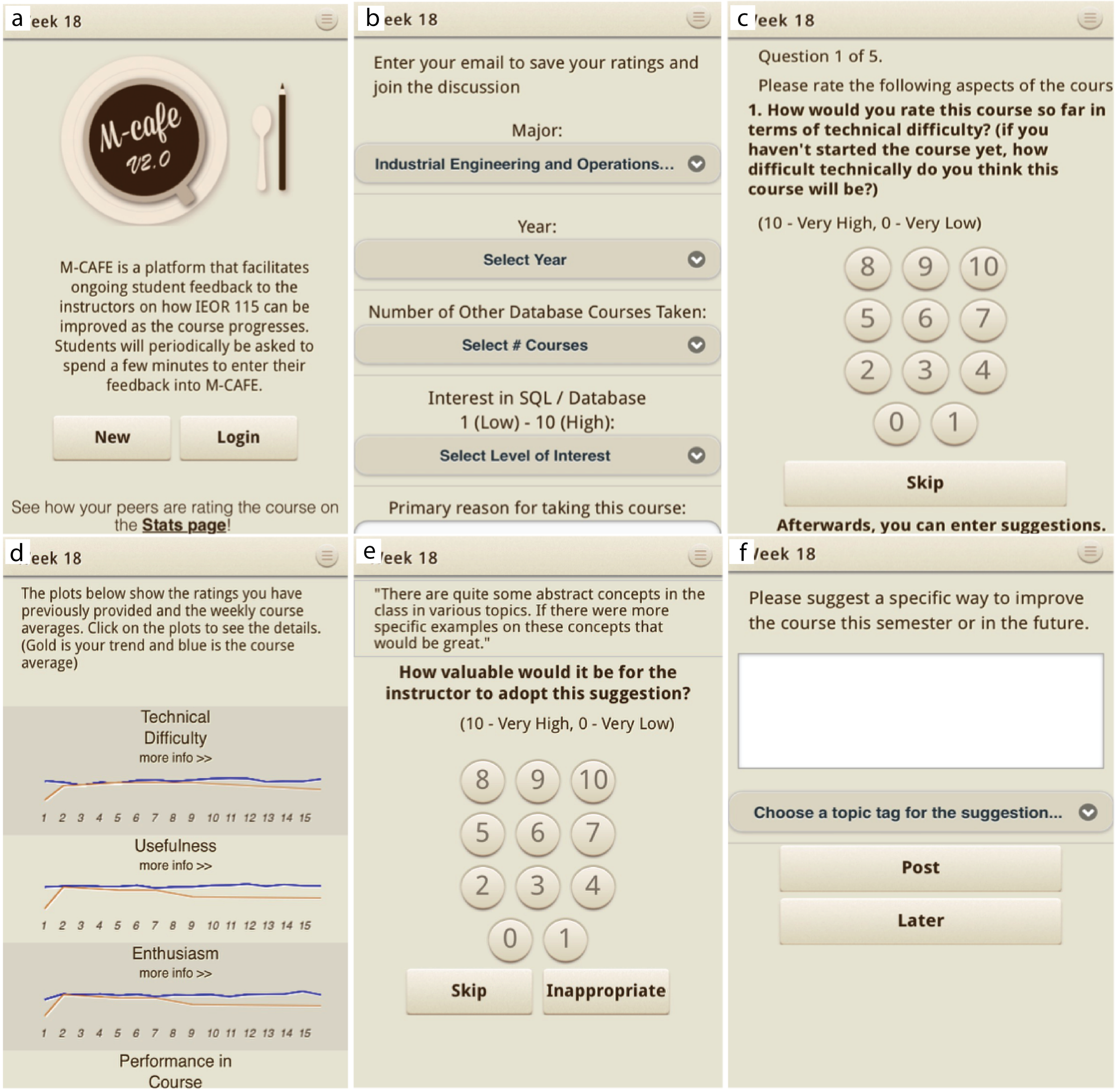
Figure 2.1: Screenshots of M-CAFE 1.0 interface.

visualization of the QATs to instructors. For example, Figure 2.4 is a plot of the course difficulty ratings over ten weeks, generated from M-CAFE 1.0 data collected in IEOR 170, S15.

Figure 2.4 provides a visualization of the changes in course difficulty over time. By viewing the plots, instructors can quickly identify the average changes from past weeks. The error bars indicate two standard errors (SE) above and below average, revealing the significance of changes at a 5% significance level. As we can see immediately from the plot, the course difficulty level increased gradually from the beginning of the semester to the latter half. It reached its peak in week 7 after the middle of the term, and at the time, the course was significantly more difficult than when it began. Revealing the relative changes in average QAT ratings provides instructors insights on the impacts of courses event changes, for example, how does this homework/exam affect students' perceptions of the course aspects? The visualizations quantify the effectiveness of individual course events and yield a decomposed evaluation on detailed course activities within fixed time intervals.

Understanding the Relationships Between QAT Rating Changes - a Validity Check

The quantitative feedback feature of M-CAFE also provides the possibility of assessing the relationships between weekly average QAT rating changes. In turn, the agreement of the relationships



M-CAFE is a platform that facilitates ongoing student feedback to the instructors on how IEOR 115 can be improved as the course progresses. Students will periodically be asked to spend a few minutes to enter their feedback into M-CAFE.

See how your peers are rating the course on the Stats page!

Enter your email to save your ratings and join the discussion

Question 1 of 5.

Please rate the following aspects of the cours

1. How would you rate this course so far in terms of technical difficulty? (if you haven't started the course yet, how difficult technically do you think this course will be?)

(10 - Very High, 0 - Very Low)

Afterwards, you can enter suggestions.

The plots below show the ratings you have previously provided and the weekly course averages. Click on the plots to see the details. (Gold is your trend and blue is the course average)

"There are quite some abstract concepts in the class in various topics. If there were more specific examples on these concepts that would be great."

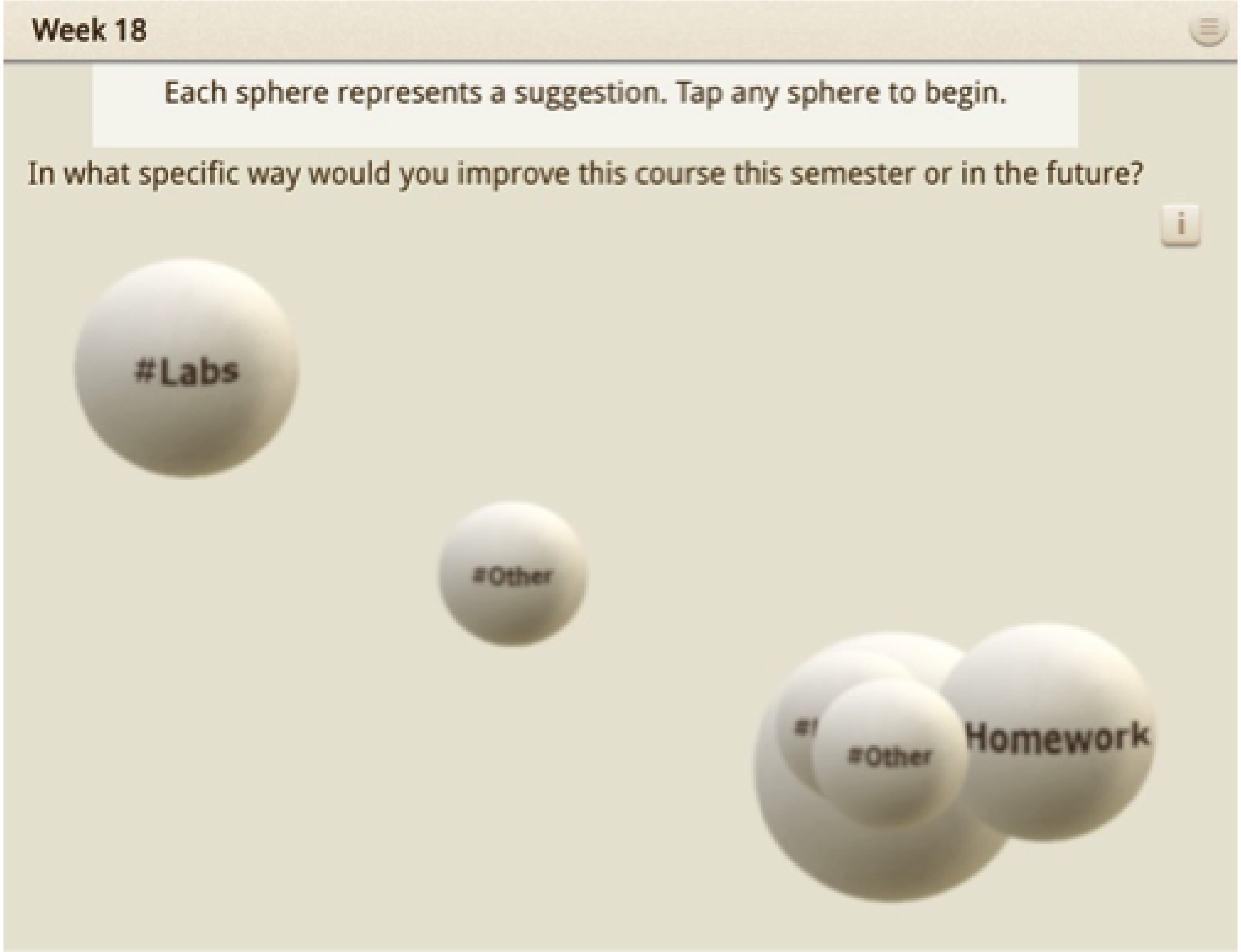
How valuable would it be for the instructor to adopt this suggestion?

(10 - Very High, 0 - Very Low)

Please suggest a specific way to improve the course this semester or in the future.

Figure 2.2: Screenshots of M-CAFE 2.0 interface.

between QAT rating changes and common beliefs could further demonstrate the reliability of the quantitative feedback from M-CAFE. Stark et al. [185] points out that quantitative scores in course evaluations suffer from validity concerns and may not be informative. Thus, we believe this analysis would be a valuable consistency check for the quantitative feedback obtained in M-CAFE. In all courses that implemented M-CAFE, we observe a negative correlation between course difficulty and self-enthusiasm, suggesting that difficult course materials lead to higher anxiety and thus



Week 18

Each sphere represents a suggestion. Tap any sphere to begin.

In what specific way would you improve this course this semester or in the future?

Figure 2.3: The space interface of course improvement suggestions with topic tags in M-CAFE 2.0.

reduced student enthusiasm. Course usefulness is positively correlated with homework effectiveness and self-enthusiasm. We speculate that as students become more enthusiastic and the homework assignments become more effective, the usefulness of the course would be rated more highly. The other relationships are not consistent among the three courses and are possibly dependent on the different student bodies and course materials involved. The agreement of the relationships between QAT rating changes and common beliefs is encouraging and is the building block of further analysis on the QAT ratings.

Social Influence Bias and Student Confidence

The visualizations (see Figure 2.4) for the QATs are available not only to instructors but can also be made available to students. Students can evaluate their ratings against the class average to get a better idea of where they stand among the peers. For example, if one student in CS 169.2x is

Stats	169.2x	169.1x	170, S15	115, F15	170, S16	115, F16
Version	1.0	1.0	1.0	2.0	2.0	Not used
User Count	348	253	58	57	54	36
QAT Set Rating Count	741	312	424	483	254	N/A
Suggestion Count	167	82	270	110	90	34
Peer Rating Count	4000	1715	2483	1759	979	N/A
Course Starting Time	Jun 2014	Oct 2014	Jan 2015	Sep 2015	Jan 2016	Sep 2016
Course Length	6 week	8 week	15 week	15 week	15 week	15 week

Table 2.1: Participation statistics in different stages of M-CAFE.

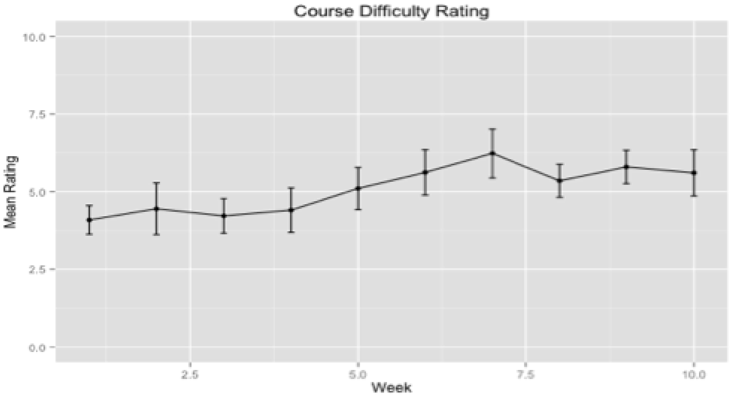


Figure 2.4: The average and two-standard errors of ratings on course difficulty over ten weeks of IEOR 170, S15.

finding the homework in week 4 particularly challenging and considers dropping the course, he might feel less stressed and gain some confidence if he learns that most of his classmates are in the same situation, i.e., the course difficulty rating is much higher in week 4 than the previous weeks. M-CAFE tries to summarize information in minimal volume and at the same time, provide a representative indicator of the course aspects that can be valuable to both instructors and students. Furthermore, M-CAFE 1.0 displays the median grade on each QAT after the student provides a rating. This feature and the QAT visualizations can potentially lead to less bias by reducing apple-to-orange comparisons in numerical scoring. One major concern about quantitative feedback on course evaluations is the varied scales among students, i.e., two students who feel the same difficulty level may provide different ratings because they don't have a common scale to refer to. M-CAFE 1.0 reduces the scale variability among students by giving the median, the average and the standard error of the ratings, allowing students to acquire knowledge of a "middle ground" rating on course aspects. However, unlike traditional paper evaluations, interactive systems are susceptible to social-influence bias, where students can change their ratings after seeing the median or rate the course close to the average rating.

Quantitative Evaluation Consistency

Comparing M-CAFE course ratings of two IEOR 170 courses offered in S15 and S16, we find that the absolute value of course assessment rating is not reliable. Figure 2.5 shows the mean weekly ratings of Course Difficulty from IEOR 170 in 2015 and 2016. Both trends increase gradually as the semester proceeds. Interestingly, there exists a gap between the two ratings from different years throughout the semester. Since the two courses were nearly identical in terms of instructor, material and schedule, there is no reason to believe that the course offered in 2016 was significantly harder than the same course offered in 2015. We suspect that the different rating scales of participating students in the two courses led to this gap, confirming that absolute course evaluation ratings can be unreliable when the population changes. For example, for some course material, one student may assign a Course Difficulty score of 6 but another student may assign a difficulty score of 8. Furthermore, displaying rating history to students provides a benchmark rating scale, which reinforces the difference between the two course ratings in later weeks. Thus we see a consistent relative change over time in average Course Difficulty ratings but a statistically significant difference in absolute ratings between the two courses. Comparing the weekly ratings from the two courses using a t-test, we see that with 5% significance level, the difference in means each week in the two years is not equal to 0, with an overall mean of 3.57 in 2015 and 5.845 in 2016. Similar gaps are found in the other 4 QATs between the two courses.

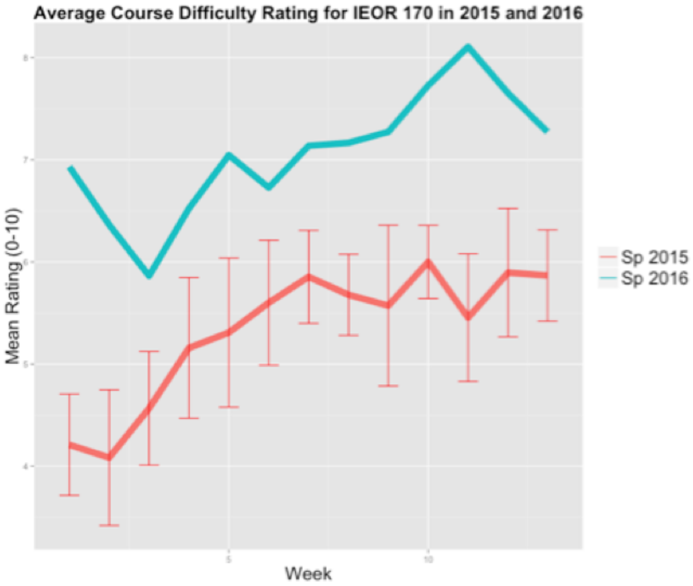


Figure 2.5: Course difficulty ratings for IEOR 170, S15 and IEOR 170, S16 over 15 weeks. Blue line: weekly mean in IEOR 170, S16; Red line: weekly mean in IEOR 170, S15.

Qualitative Evaluation with CF

Identifying the Most Valuable Ideas for Instructors

A shortcoming of qualitative data is its lack of structure. Natural Language Processing (NLP), although has been an active research field for years, is not effective in selecting a subset of insightful ideas from M-CAFE-generated qualitative data. Current text analysis of qualitative data hints at the important words or phrases, whereas the underlying sentence structure and word meanings are mostly ignored. As an alternative approach to NLP, Collaborative Filtering (CF) has gained popularity for ranking and recommending items in fields using peer-to-peer ratings.

The Ranking Metric

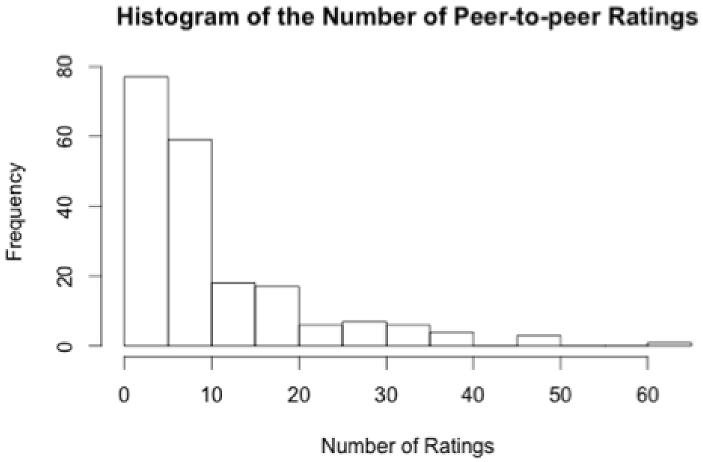


Figure 2.6: Histogram of the number of peer-to-peer CF ratings for ideas in IEOR 170, S15.

As can be seen in Figure 2.6, few ideas received more than 20 peer-to-peer CF ratings. Ranking ideas by their mean or median rating would not be reliable due to the small sample size and the variation in rating differences. Instead, we compute the Wilson score for each idea using the lower bound of the binomial proportion confidence interval. This incorporates the variance in ratings as follows. We took the mean grade g and then calculated the 95% confidence interval of g using standard error: g +/- 1.96*SE(g). We then rank the ideas by the lower bound g - 1.96*SE(g).

Topic Tagging and Incentive Analysis for M-CAFE 2.0

Two major challenges in the qualitative part of end-of-course evaluations are the analysis difficulty of unstructured data and the limitation of comment variety resulting from repetition. M-CAFE 2.0 addresses these challenges by collecting textual suggestions from students using topic tagging. For instance, after articulating a suggestion, students are required to choose the appropriate topic tag from a dropdown containing Exams, Homework, Labs, Lectures, Logistics, New Topics, Projects, Other. If the student chooses "Other," he/she is encouraged to suggest a new topic tag. For the following analysis, we define "Course Topic" as the topic tags associated with the suggestions, i.e., Exams, Homework, etc. and "Course Module" as the topic of the course materials, i.e., SQL, Relational Schema, etc. M-CAFE 2.0 organizes feedback by topic tags and encourages students to evaluate the course on a weekly basis when different course modules are covered.

Initial results suggest that:

(1) Students are more likely to provide a new suggestion for topics they have not considered in the peer-rating phase.

By requiring students to provide at least two peer ratings before supplying their own, M-CAFE 2.0 aims to reduce suggestion repetition. After investigating data from the two courses that used M-CAFE 2.0, we find that in both courses more than half of the students supplied a new suggestion with a topic different from the topics that they rated in the peer-rating phase, indicating an intention to articulate new suggestions that are different from those already in the system. Eighty percent of the students in IEOR 115 and 76% of students in IEOR 170 articulated a suggestion from a topic different from the topic of their last rated suggestion. For the students who rated at least one suggestion of the same topic, the content of the new suggestion is different from the rated suggestion. For example, a student in IEOR 115 rated 6 suggestions before providing his/her own, 3 on lectures, 1 on projects, 1 on homework and 1 on other. The suggestions on lectures he/she rated are:

1. "When drawing E-R Diagrams on the board–or any other diagrams that are complex–plan it out so that none of it has to be erased/moved to another board. It is way harder to fix diagrams on paper as we take notes."
2. "Slow down a little bit."
3. "The discussion sections could be clearer. It is difficult to see what kind of table manipulations are going on."

And the student provided the following new suggestion on lectures:

"If possible, posting an outline of every lecture will be very helpful considering the fast pace of the lecture. As a result, we can pay more attention to the explanation instead of putting too much effort in copying down everything in the notes."

Below is another example of a student who rated two suggestions on Homework and further provided a suggestion on Homework. The two suggestions he/she rated:

1. "I hope we receive plenty of feedback on what to improve from graders on our homework and exams."
2. "Re-evaluate homework length. The current homework on designing and prototyping a restraint was too much to ask for in one week – the design and report alone took me 5 hours. Prototyping was much harder this week due to reduced Jacobs access hours, and my friends and I were not able to make a physical prototype. This homework in particular should have its point balance changed to not take a physical prototype into account."

And the suggestion he/she provided afterwards:

"The fusion360 tutorial in class is not helpful for a first time user at all, and the assignment this week is not easy for students who never use 3D graphic software before."

In case of this student, we suspect that this student is interested in the topic "Homework". Thus he/she purposefully clicked on the spheres with the Homework tag to see if any other students have already suggested the same suggestion. After finding that these two existing suggestions are different from his/hers, he/she articulated his own.

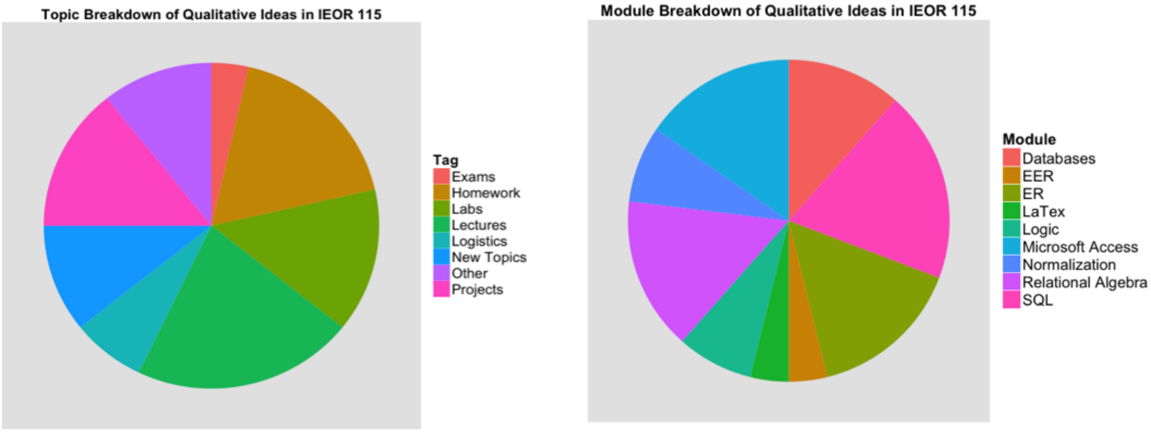


Figure 2.7: Left panel: Pie plot of IEOR 115 in 2015 indicating the breakdown of qualitative suggestions in terms of course topics; Right panel: Pie plot of IEOR 115 in 2015 indicating the breakdown of qualitative suggestions in terms of course modules.

(2) Course improvement suggestions from M-CAFE 2.0 are more diverse than those from traditional mid-course evaluations.

We compare two IEOR 115 courses offered in 2015 and 2016 to observe their suggestion diversity. IEOR 115, F15 used M-CAFE 2.0 and encouraged students to provide qualitative suggestions on a weekly basis, whereas the same course with the same instructor, materials and schedule offered in 2016 didn't use M-CAFE 2.0. Instead, a paper-based mid-term unofficial course evaluation was conducted in lecture on Oct. 3, 2016. We compare the paper results to the qualitative suggestions on M-CAFE 2.0 suggested before Oct. 3, 2015. For IEOR 115, F15, M-CAFE 2.0 collected a total of 50 unique suggestions with topic distribution shown in the pie plot (Figure 2.7a), with 13 on Homework, 5 on Labs, 16 on Lectures, 2 on Logistics, 4 on New Topics, 1 on Policies, 5 on Projects and 4 on Other, covering most topics of the course. Out of the 50 suggestions, 35 suggestions mention a specific course module. Figure 2.7b displays the proportion of suggestions on each course module and results show that most modules received improvement suggestions and the number of suggestions per module is positively correlated with the number of lectures the instructor spent on the module. The suggestions range from "I hope we receive plenty of feedback on what to improve from graders on our homework and exams." to "Despite the pace of the lecture, the examples given in the lecture so far are helpful to understand the overall concept of entity-relationship diagram. For the future, I think it would be better if the Professor can elaborate more why he does a particular step." For IEOR 115, F16, we manually analyzed the mid-term evaluations from 36 students and summarized the textual suggestions. Out of all the suggestions, 18 are unique within the list, covering multiple course aspects such as Lectures, Homework, Labs, Logistics, etc. However, many students provided the same suggestion: for example, 5 students requested the instructor "to provide a study guide/notes for the lectures." Seventeen of the 18 suggestions are general statements that do not refer to any particular course module, thus making it impossible for the instructor to understand how the students perceive each course module.

Anonymous Extra Credit Design

Since the M-CAFE platform is completely anonymous, to reward the students who were main contributors to the course, a different anonymous extra credit system was developed. This system allows students to claim extra credit while remaining anonymous. The process is as follows:

The platform first generated unique code for each student that was awarded and sent the code to his/her email address (the email they used to signup for the account). The list of codes were shared with the instructors but the associate M-CAFE account was not exposed. Then the student have the freedom to decide if he/she would like to claim the credit. He/she can share the code with the instructors and the code is valid when it was claimed the first time. In this case, the students remain anonymous but are awarded for contributing to the platform.

2.4 Conclusion

M-CAFE collects ongoing student course evaluations and effectively identifies valuable suggestions. Pilot studies suggest that visualizing relative changes in course assessment over time and topic tagging encourages a more diverse set of course suggestions.

2.5 Future Work

The current versions of M-CAFE are separate, standalone website applications. In the future, we would like to integrate M-CAFE to MOOC platforms such as Coursera and edX. This integration can enable instructors to self register for M-CAFE and start using the tool for his/her courses. Wide adoption of M-CAFE will also generate more data that we could use to further iterate on the platform design.

Chapter 3

The DebateCAFE Platform

3.1 Background and Related Work

Online Deliberation Platforms

Online deliberation systems support structured discussion and debate among participants to reach informed decisions that usually involve complex issues and difficult trade-offs [52, 109, 180, 183, 200]. Compared with face-to-face public deliberation, online deliberation has the potential to enable wider participation in civic engagement opportunities and to facilitate group decision-making and problem solving [173].

With growing numbers of researchers and government officials recognizing the effectiveness of public deliberation, many online deliberation platforms have been created [35, 42, 71, 76, 86, 102, 110]. The Deliberatorium structures input into issues/problems, ideas/solutions, and for/against arguments [102]. Chilton et al. introduced a collaborative application, Frenzy, to gather the entire program committee for conference session-creation [42]. This crowd-sourced approach significantly reduced the time needed to make a conference program. Kriplean et al. developed a plug-and-play platform called Consider.it, which enables participants to view and articulate arguments on a linear scale of a particular issue. Participants are exposed to pro/con arguments and are encouraged to adopt arguments they find persuasive and articulate pro/con points [110]. Consider.it then requests participants to position themselves on a linear scale to reflect their stand on the issue. Debategraph is another deliberation tool that visually organizes complex debates into graphs, where each node represents an argument and each edge reflects arguments' relationship [20]. Participants are encouraged to provide evidence for each sub-argument. While these platforms are valuable in organizing arguments and collecting feedback, arguments on both sides can be greatly lopsided and the large volume of textual arguments makes it difficult for these systems to harvest valuable insights.

Selective Exposure

One potential drawback of online deliberation platforms is selective exposure, when participants seek out confirming information and resist changing their opinions. An inherent characteristic of the internet is the freedom to select from a range of content. However, this freedom enlarges the concern that individuals only view information that aligns with their personal attitudes. A number of studies acknowledge this issue [6, 41, 53, 73, 104, 119, 120, 155, 170] and it has been shown that people differ in their willingness to be exposed to adverse information, especially political content [160, 177, 193] and that exposure to online content is tailored by algorithms, creating what Pariser refers to as "filter bubbles" that limit exposure to divergent viewpoints [150]. This phenomenon of unbalanced exposure could largely hinder people from making rational and informed decisions and can impede political tolerance [105]. For instance, Knoblock-Westerwich and Meng show in an experiment that participants are inclined to view articles that are consistent with their own position disregarding the target issue, with 36% more reading time [104]. Salehi et al. found that the decentralizing characteristics of the internet can restrict the discussion to a narrow path in collective intelligence systems [170]. Due to such inherent bias, a small set of initial users with similar positions can dominate discussion, skew the ratings of arguments and eventually reduce the effectiveness of online deliberation. Therefore, greater care is needed when designing such systems to increase awareness of alternative perspectives [29].

To mitigate selective exposure, Graells-Garrido et al. built a platform that mixes a visual interface and a recommender algorithm to recommend politically diverse profiles to each user and found that an indirect approach in developing systems that reduce user behavior bias can be beneficial [73]. Gao et al. designed a social forum interface that displays controversial social opinions with user reactions to various stances. Results suggest that showing different stances for each topic and providing user reactions to each topic can effectively mitigate selective exposure [55]. However, when the crowd starts to grow, their interface will rely greatly on moderators to categorize new insights, which could quickly become unwieldy. Moreover, directly revealing summarization of stances can be intimidating to those users in the minor group and discourage further participation.

Collaborative Filtering

Scaling is another challenge faced by online deliberation platforms. One approach to handle the scale issue is to leverage the crowd using collaborative filtering. Collaborative filtering aggregates subjective ratings provided by humans to assign a numerical reputation to each item [74]. In many CF systems, such as Amazon and Netflix, the reputation of each item is based on the aggregated ratings from a neighborhood of similar items [123]. CF can also be applied globally when the reputation of an item depends on the aggregated ratings from ALL participants, reflecting a common opinion of the crowd. Most CF systems adopt a list-based presentation, resulting in unbalanced exposure of items, where highly rated items are shown on top [223]. Though the system does not set out to bias any item, the self-selection nature of humans could hinder the presentation of new items due to limited exposure [145]. This can be particularly counterproductive for online deliberation platforms, where arguments on one side could have greater exposure, discouraging participants to articulate and rate arguments with contradicting views.

Theory of Incentives

Selective exposure and biased discussion on existing online deliberation platforms are results of conflicting objectives, where users maximize their persuasiveness on the position they agree with and the platform wishes to optimize the persuasiveness of all positions. The theory of incentives has identified conflicting objectives and decentralized information as two basic factors that lead to inefficient outcome [117]. In particular, the principal-agent model is an abstraction of our use case [55, 93]. In the principal-agent model, the "agent" is able to make decisions on behalf of the impact of the "principal", where both parties (the agent and the principal) have contrary interests and asymmetric information [28]. Oftentimes, it is costly for the agent to perform the activities that are useful to the principal. This model has wide adoption in corporate management and is a core motivation of contract theory. A game theory approach of the problem suggests that an introduction of rules of the game so that the agent interest coincides with that of the principal can be beneficial [165].

Study Purpose

Novel online deliberation platforms are needed to automate and/or assist the moderation task of filtering out valuable items at scale while mitigating selective exposure bias.

We developed the Debate Collaborative Assessment and Feedback Engine (DebateCAFE v1.0), a new platform that seeks to mitigate selective exposure bias with an interface that scales to a large number of participants. DebateCAFE collects quantitative feedback to speculate the initial position of participants, encourages participants to enter convincing arguments on both sides of an issue, and rate the persuasiveness of other participants' arguments. Determining the persuasiveness of the argument is crowdsourced with peer-to-peer ratings, as participants are shown peer arguments (based on an uncertainty sampling algorithm). Participants then receive points for the persuasiveness of their weakest argument. The combination of arguing both sides and a sampling interface avoids allowing a small set of initial participants with similar positions to dominate discussion (a problem highlighted by Salganik [171]) Similar to M-CAFE, DebateCAFE also uses Collaborative Filtering (CF) to solve the scale issue. CF allows DebateCAFE to scale to an expanding population of participants and process unstructured textual data on subjective access (i.e., how persuasive is this argument). While collaborative filtering is effective in bringing some structure to lists of items by assigning reputation scores, this should be coupled with incentive mechanisms and an interface design that helps to avoid the biases of selective exposure [29].

In this chapter, we describe a case study on the issue of "Apple vs. FBI," inspired by the debate of whether private companies (i.e., Apple Inc.) should cooperate with government (i.e., FBI) requests to have backdoor access to encrypted information technologies. We report data and system performance from a preliminary study with 94 University of California, Berkeley students who entered 170 arguments on both sides and conducted 1754 peer-to-peer ratings of the persuasiveness of arguments. Results offer valuable insights into platform mechanism design. However, because the participants reflect a narrow demographic (i.e., university students), results may not be reflective of the general population.

This chapter is structured as follows: we first evaluate the current state of online deliberation platforms, summarize the design choices that mitigate selective exposure, and survey the theory of incentive to aid the design of our platform. Then we describe the interface and algorithm design of DebateCAFE, in response to the shortcomings of the existing platforms. Finally, we present results from a user study on the "Apple vs. FBI" issue to measure selective exposure and validate the efficacy of the platform.

3.2 The DebateCAFE Platform

In this section, we first describe the user interface of DebateCAFE. Next we introduce the incentive mechanism implemented in the platform to encourage participants to articulate strong arguments for both sides of the issue. At last, we describe an instance of the platform centered on the "Apple vs. FBI" issue.

Interface

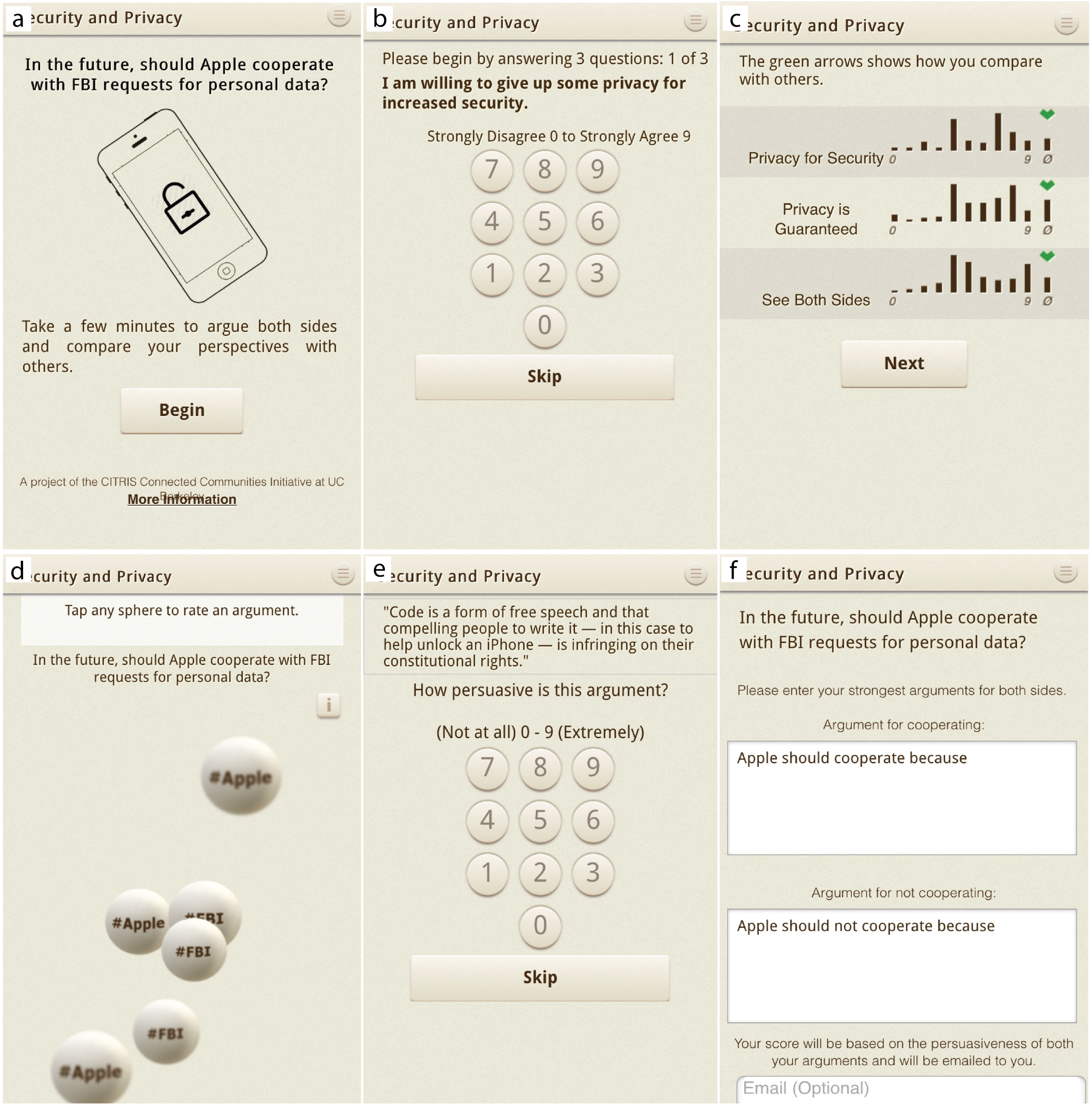
DebateCAFE guides participants through three stages: assessment, argument articulation, and peer-to-peer argument evaluation.

Assessment Phase

Participants first assess their current beliefs of the discussed issue by rating three Initial Bias Assessment questions (see the Apple vs. FBI issue section for an example). These statements should be able to summarize participants' initial stance. Participants have the option to skip any question they choose not to answer by either pressing the skip button or leaving the response blank (Figure 3.1b). Participants are then asked to provide their zip code. We find zip code to be an informative demographic statistic, while not being so intrusive as to hinder further participation in the system. After that, DebateCAFE displays the histograms of the responses to the three IBAs so that the participant learns his/her position on these questions relative to all participants (Figure 3.1c).

Argument Articulation Phase

In this phase, participants are prompted to formulate their own arguments in response to the central discussion question (Figure 3.1d). Compared with previous CAFE instances, DebateCAFE is novel in that participants are prompted to enter two arguments: one for each side. The interface strongly encourages, but does not require that a participant fill in both arguments. DebateCAFE also requests that participants supply their email address so we can update them with peer-rating scores on their arguments (explained further in the next section).



In the future, should Apple cooperate with FBI requests for personal data?

Take a few minutes to argue both sides and compare your perspectives with others.

A project of the CITRIS Connected Communities Initiative at UC Berkeley

Please begin by answering 3 questions: 1 of 3

I am willing to give up some privacy for increased security.

Strongly Disagree 0 to Strongly Agree 9

The green arrows shows how you compare with others.

Tap any sphere to rate an argument.

In the future, should Apple cooperate with FBI requests for personal data?

"Code is a form of free speech and that compelling people to write it — in this case to help unlock an iPhone — is infringing on their constitutional rights."

How persuasive is this argument?

(Not at all) 0 - 9 (Extremely)

In the future, should Apple cooperate with FBI requests for personal data?

Please enter your strongest arguments for both sides.

Argument for cooperating:

Apple should cooperate because

Argument for not cooperating:

Apple should not cooperate because

Your score will be based on the persuasiveness of both your arguments and will be emailed to you.

Email (Optional)

Figure 3.1: DebateCAFE mobile interface screenshots. The Initial Bias Assessment phase asks Likert scale questions to assess initial bias of each participant and displays the results with histograms based on all previous participants. The platform then requests input of two adversarial arguments followed by a graphical display to collect peer-to-peer numerical evaluation of the arguments from other participants.

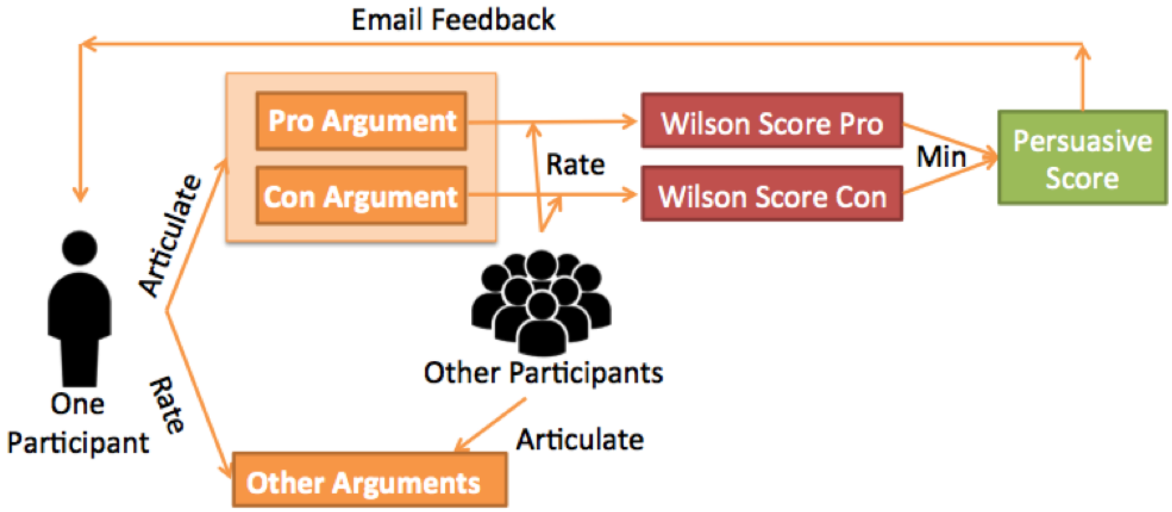


Figure 3.2: Information flow of DebateCAFE. Each participant articulate two arguments for opposing positions of an issue. Other participants rate the persuasiveness of the two arguments. Then, ratings are aggregated using the Wilson metric and the persuasive score for the participant is computed as the lower of the two argument scores and is sent back to the participant via email. In addition, participants are encouraged to rate other arguments on an uncertainty-sampling interface.

Peer-to-Peer Argument Evaluation Phase

Next, participants enter the "discussion space", a 2D visualization where other participants' arguments are represented by spheres arranged across the space (Figure 3.1e). This discussion space displays 8 argument spheres at a time, 4 "pro" and 4 "con" arguments, and the stance of the argument is labeled on the sphere. Notice the actual arguments are not displayed in this page and participants have the freedom to click on whichever arguments they wish to view. In order to better ensure that all arguments are seen and rated, DebateCAFE prioritizes the display of arguments that have high uncertainty in their evaluation grades. We quantify uncertainty of argument i using the standard error:

$$SE_i = \frac{SD(R_i)}{\sqrt{N_i}} \tag{3.1}$$

where R_i are a list of ratings for argument i and N_i is the number of ratings argument i receives.

The spheres are placed in the 2D space according to the first two dimensions of a Principal Component Analysis (PCA) applied to participants' responses during the assessment phase [215]. (Skipped questions are assigned the mean response rating for that question.) Thus spheres that are closer to the center of the space are provided by more similar participants in terms of IBAs. Participants then click on the spheres in the 2D space to read other participants' arguments. Finally, using a rating interface similar in design to that used during the assessment phase (Figure 3.1f), participants evaluate their peers' arguments on the question "How persuasive is this argument?" using a scale from 0 (Not at all Persuasive) to 9 (Extremely Persuasive).

Persuasive Scoring Mechanism

A key challenge of online deliberation platforms is the conflict of interest between users and the platform, where users optimize their persuasiveness of their position and the platform aim to collect valuable arguments for all positions. To resolve this conflict, DebateCAFE not only provides an uncertainty-sampling interface in the argument evaluation phase, but also introduces a "persuasiveness score" for each participant. We first describe how we compute the score of each argument and then explain the aggregation procedure. In crowdsourced rating systems, using the average rating for assessment is not robust due to small sample size. Therefore, in DebateCAFE, we calculate argument score with the lower bound of the binomial proportion confidence interval (also called the Wilson Score) [214]. Intuitively, this approach is more robust because it incorporates information about the uncertainty of the score estimate. For example, an argument that receives ratings 10, 0 is ranked lower than one that is rated 5, 5. Now each participant receives two scores s_1, s_2 for his/her two arguments on opposing positions. We adopt the insights from the theory of incentive to define the overall persuasiveness of a participant as the minimum of these two scores $\min(s_1, s_2)$. We can view our problem as a principle-agent problem where the platform is the principal and users are agents. We assume that users aim to maximize the persuasiveness of their own position and providing a strong argument for the opposing position lowers their utility. Conversely, the platform tries to optimize argument quality for all positions. Under this setup, we can derive the utility optimization problem of a user in the absence of the persuasive score mechanism as:

$$\begin{aligned} \max\ U_{Init} &= |s_1 - s_2| \\ \text{s.t. } s_1 &\in [0, v_1] \\ s_2 &\in [0, v_2] \end{aligned} \tag{3.2}$$

where v_1,v_2 are the scores of the user's most persuasive arguments on the two positions. Without loss of generality, we can assume that this user believes in position 1, then (3.2) simplifies to:

$$\begin{aligned} \max\ U_{Init} &= s_1 - s_2 \\ \text{s.t. } s_1 &\in [0, v_1] \\ s_2 &\in [0, v_2] \end{aligned} \tag{3.3}$$

Therefore, the optimal solution is $s_1 = v_1$ and $s_2 = 0$, i.e., this user will provide his/her strongest argument for position 1 and provide his/her weakest argument for position 2. Now we introduce the persuasiveness scoring mechanism and further assume that users gain utility from their overall persuasiveness. Notice also that there is asymmetric information since the platform does not know

v_1, v_2. If we let the value of persuasiveness to be λ, then the principal-agent model becomes:

$$\begin{aligned} \max\ & s_1 + s_2 \\ \text{s.t. } & s_1, s_2 = \arg\max\ (|s_1 - s_2| + \lambda \min(s_1, s_2)) \\ & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned} \tag{3.4}$$

and each user solves the updated utility optimization problem as follows:

$$\begin{aligned} \max\ & U = |s_1 - s_2| + \lambda \min(s_1, s_2) \\ \text{s.t. } & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned} \tag{3.5}$$

Let's again assume this user believes in position 1, then we can simplify (3.5) to:

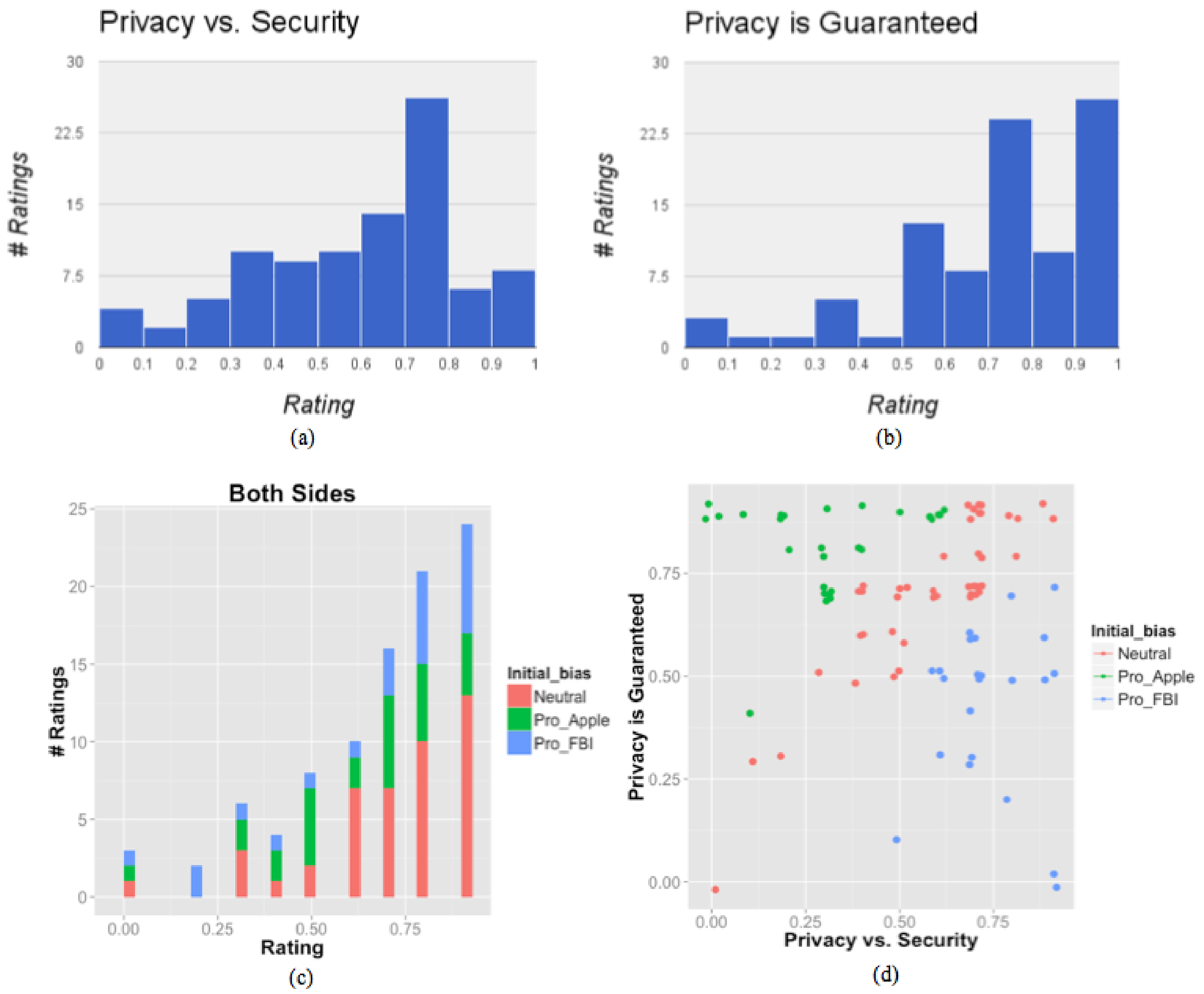
$$\begin{aligned} \max\ & U = s_1 - s_2 + \lambda s_2 \\ \text{s.t. } & s_1 \in [0, v_1] \\ & s_2 \in [0, v_2] \end{aligned} \tag{3.6}$$

Now when $\lambda > 1$, the optimal solution is $s_1 = v_1$ and $s_2 = v_2$, i.e., the user will provide his/her strongest argument for both positions, which aligns with the incentive of the platform. Notice that even though the value of λ is user-specific (as users can have different valuations on being persuasive), it can be enhanced by additional mechanism designs. For instance, the platform can limit the exposure of arguments provided by users with a low persuasiveness score or having a public leaderboard showing the most persuasive users. We plan to explore the impact of these additional designs in the next version of DebateCAFE.

The Apple vs. FBI Issue

To evaluate platform performance, we chose the topic of "Apple vs. FBI." The issue arose when the FBI requested that Apple help investigators gain access to an iPhone used by Syed Rizwan Farook in a December 2015 mass shooting in San Bernardino, CA [13]. Apple refused the request because it would require writing new software to bypass encryption features of the iPhone and would create "the potential to unlock any iPhone in someone's physical possession" [97]. This particular incident attracted public attention and led to debates on social media platforms and related online forums. Starting with this incident, the discussion has expanded to a more general theme of "personal privacy vs. national security". We perceive this issue as a highly complex and controversial topic that may be clarified through deliberation, allowing the public to articulate ideas side-by-side. For this issue, we encourage participants to rate the following three IBA questions on a 10-point scale from 0 (Strongly Disagree) to 9 (Strongly Agree).

1. I am willing to give up some privacy for increased security.
2. Personal privacy should be guaranteed by the US Constitution.
3. There are reasonable arguments on both sides of the security and privacy debate.



(a)

(b)

(c)

(d)

Figure 3.3: IBA summary plots. Panel (a), (b), (c): Histogram of IBA 1-3 respectively; Panel (d): Plot of IBA2 versus IBA1, color coded by the estimated initial position of the participant.

These statements were chosen as they succinctly summarized participants' initial stance (pro-Apple vs. pro-FBI) on digital privacy, as well as their degree of open-mindedness to opposing arguments. In the argument articulation phase, the main discussion question is, "In the future, should Apple cooperate with FBI requests for personal data?" To help bootstrap participant's writing, we prepopulate the textboxes with "Apple should cooperate because", and "Apple should not cooperate because." Furthermore, in the peer-to-peer argument evaluation phase, each sphere is labeled either "Apple" or "FBI" to reflect the stance of the argument.

3.3 Case Studies and Analysis

In this section, we present results from a preliminary study of undergraduate students at University of California, Berkeley who were assigned to participate in DebateCAFE. The resulting dataset contains responses from 94 students with 284 IBA responses, 170 new arguments on both sides and 1754 peer-to-peer ratings. Note that the population in this study is not a representative sample but this informal study provides insights into the performance of DebateCAFE and gives suggestions for future design choices. For analysis purpose, we scale the raw ratings to 0-1. We first reveal the distribution of the IBA responses and the initial positions of the participants. Next we present the top arguments for both positions to evaluate their quality. After that, we quantify selective exposure and measure the subsequent biases. Finally we summarize the participant scores to evaluate the incentive mechanism.

The Initial Bias Assessment Questions

The first two questions provide a rough indication of the participant's initial attitude toward the issue and the third question captures how open the participant is to deliberation. Figure 3.3 panels (a), (b) and (c) show the histogram of each of the three IBAs. The response to IBA1 has a wider spread than IBA2 and IBA3 with a mean at 0.54, indicating people have varied preference in giving up privacy for security. IBA2 is skewed left suggesting that most participants value personal privacy and regard it as a constitutional right. For IBA3, responses are clustered at 0.7-1.0, reflecting an appreciation of arguments on both sides of an issue. The relation between IBA1 and IBA2 is interesting. Figure 3.3d shows the scatter plot of IBA2 vs. IBA1. Observe that most points lie above the y=1-x line, demonstrating that participants who were not willing to give up privacy for security had a strong pursuit of personal privacy. Participants on the top left of the plot were likely to favor Apple's position, whereas participants on the bottom right were likely to favor the FBI. The participants' initial estimated positions were determined by the following metric: If IBA2 - IBA1 > 0.3, this participant is "pro-Apple"; If IBA1 - IBA2 > 0.1, this participant is "pro-FBI"; These criteria were not symmetric because the ratings received for IBA1 and IBA2 were asymmetric and we wanted the two groups to have similar number of participants. This metric resulted in 27 "pro-Apple" participants, 23 "pro-FBI" participants and 44 "neutral" participants as shown in three colors in Figure 3.3d.

Arguments on Both Sides

Out of a total of 170 arguments, after ranking with the Wilson metric, 17 of the top 20 were "pro-Apple" arguments and 3 were "pro-FBI" arguments. This outcome may be a result of the demographic characteristics of the respondents, most of whom were university students. They were more likely to own an Apple product and may be more skeptical of government agencies. Here we present the top 3 arguments on both sides of the issue. Pro-Apple (personal security) arguments:
1. Apple should not cooperate because it sets a dangerous precedent with regard to privacy and the FBI and other investigative groups. It also has the potential for issues with the same opening being

exploited by hackers.
2. Apple should not cooperate because allowing the FBI access to these systems opens a door that cannot be closed again at will. Creating a cyber-security loophole allowing access for the FBI means that anyone with the technological knowledge can also exploit this access; a weakness in the fundamental security creates vulnerability. Our personal information would be vulnerable to sophisticated attacks by hackers and terrorists themselves.
3. Apple should not cooperate because it violates customer's privacy. If customers do not trust Apple with their data anymore then sales will drop and Apple will no longer be relevant in tech. One thing that the government is asking is to create a backdoor for a particular OS, which could lead to hackers exploiting a bug in subsequent OS's. This is a huge security red light because Apple does not and should not knowingly create a backdoor which could lead to major security concerns in the future.
Pro-FBI (national security) arguments:
1. Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.
2. Apple should cooperate because if there is a little compromise involved in guaranteeing the security of the nation, then as a resident and/or citizen of this nation, people should be willing to make that compromise.
3. Apple should cooperate because there is an increased security risk when they do not cooperate. If potential terrorist information is on the phones of those like the San Bernardino shooters, it will jeopardize the entire safety and security of the United States.
These arguments are rather well articulated with a clear position and concrete reasoning, demonstrating the capability of DebateCAFE in harvesting persuasive arguments for opposing stances. However, argument duplication is an issue: we observe many arguments conveying the same idea with slightly different wording. For example, the top 2 "pro-Apple" arguments both mention that complying with the FBI's request could lead to potential cyber-attack from hackers and the top 2 "pro-FBI" arguments point out the tradeoff between personal privacy and increasing national security from terrorist acts.

Clustering

Measuring Selective Exposure

As mentioned earlier, selective exposure is a major concern for online deliberation platforms, and DebateCAFE is able to quantify the extent of this behavior. Recall that DebateCAFE's peer-to-peer argument evaluation interface presents 8 arguments covering both sides of an issue. Each argument is represented by a sphere in the space with a tag of either "Apple" or "FBI" indicating the position of the argument. Participants are free to click on any sphere in the space when they first land on this page. This design enables us to observe which side of the argument a participant chooses to view first given his/her initial bias estimated by the IBAs. 20/27 pro-Apple participants first selected an argument for "Apple" and 11/23 pro-FBI participates first selected an argument for "FBI." Participants were more inclined to first view an argument for "Apple" despite their initial position and a greater percentage of pro-Apple participants were more inclined to first view an argument that agreed with their position compared to pro-FBI participants. The difference was, however, not significant. Furthermore, 25/27 pro-Apple participants and 22/23 pro-FBI participants viewed at least one argument for the opposing position. This high percentage suggests that indirectly exposing arguments on both sides via an uncertainty-sampling interface can prompt participants to proactively view/rate adversarial arguments.

Articulation Bias

Here we would like to compare the quality of participants' arguments for and against their initial position (as estimated from the results of their IBAs). Peer-ratings of argument persuasiveness reveal that among the pro-Apple participants who provided rated arguments, only 3/20 had a higher rated argument for "FBI." However, among the pro-FBI participants who provided rated arguments, 8/21 had a higher rated argument for "FBI." Some participants did not provide arguments and some arguments were not rated. Results indicate that pro-Apple participants were less likely to articulate persuasive arguments for the opposing view, while pro-FBI participants had little trouble crafting persuasive arguments for the opposing view, but the difference is not significant.

Peer Rating Bias

We received 1754 valid peer ratings, among which 851 rated "pro-FBI" arguments and 903 rated "pro-Apple" arguments. By adopting the Welch two-sample t-test, with a t-value of 4.289 and p-value <0.01, we conclude that the "pro-Apple" arguments were, on average, receiving significantly higher ratings than "pro-FBI" arguments. We conjecture this difference came from the inherent bias of participants, who tended to rate more highly those arguments that align with their position. Furthermore, we classify the arguments as either "Consistent" or "Adversarial" with respect to initial bias. From participants in the pro-Apple and the pro-FBI groups, "Consistent" arguments received an average rating of 0.523 with a Standard Deviation (SD) of 0.25 and a Standard Error (SE) of 0.0118, while "Adversarial" arguments received an average rating of 0.485 with a SD of 0.25 and a SE of 0.0119. Although the absolute difference was under 5%, with a p-value of 0.022, the Welch two-sample t-test indicates that there existed a true difference in means, suggesting a consistency between initial bias and ratings bias. However, a controlled experiment is warranted to confirm this.

Participant Score

DebateCAFE defines participant score as the minimum of the two argument scores: $min(s_1, s_2)$, i.e., based on the weaker of the two arguments entered, to reflect a participant's ability to articulate adversarial arguments. We scaled all the scores to 0 to 1 for more intuitive comparison. From Figure 3.4, we observe that participant scores followed a normal distribution, where few participants were extremely capable or incapable of articulating adversarial arguments and most participants were mediocre. Among the top-scoring participants, we confirm they were able to provide strong

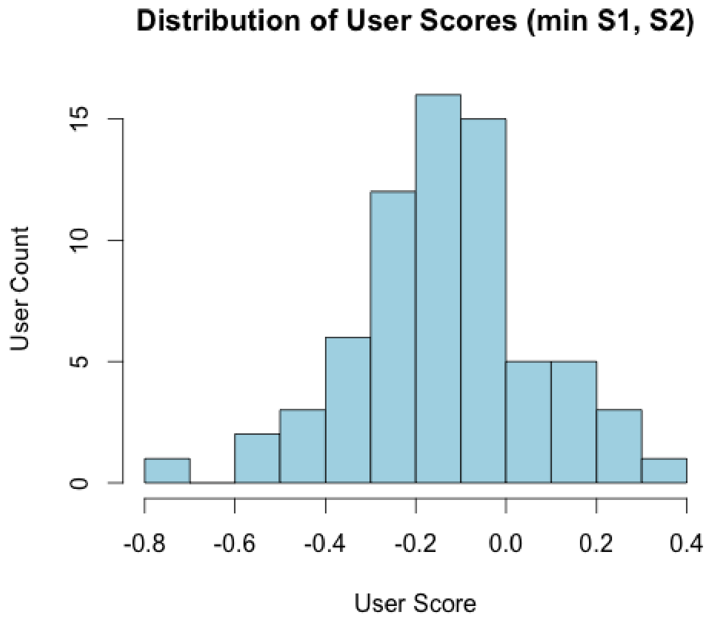


Figure 3.4: Histogram of scaled scores of all participants.

adversarial arguments. For example, the two arguments from a top-scoring participant were:
(Pro-FBI argument) Apple should cooperate because it is a small price to pay for the increased understanding of this terrorist act.
(Pro-Apple argument) Apple should not cooperate because if this gets scaled to all cases, it could seriously diminish the customers' privacy. Note the first argument was among the top 3 pro-FBI arguments and the second argument focused on the consequence of scaling such action to all cases. Among the low-scoring participants, the Wilson scores of their two arguments were highly lopsided. The participant with the lowest score only provided a pro-Apple argument, leaving the pro-FBI argument as "Apple should cooperate because." Another participant with a low score gave a "pro-FBI" argument as "Apple should cooperate because public security?" while providing a well-articulated "pro-Apple" argument: "Apple should not cooperate because it has paid such amount of advertisement and technology to increasing security, so it should not yield all to FBI".

3.4 Conclusion

We present a novel deliberation platform, DebateCAFE, designed to encourage users to review and articulate adversarial arguments on contentious issues. DebateCAFE implements collaborative filtering to identify persuasive arguments and to balance exposure to polarized viewpoints using uncertainty sampling. By introducing a scoring system to users, DebateCAFE incentivizes them to articulate persuasive arguments on both sides of the issue.

We describe an application of DebateCAFE to the Apple (personal privacy) vs. FBI (national security) issue and report results. By presenting an equal number of arguments on both positions using uncertainty sampling, our platform is successful in motivating participants to view and rate arguments on opposing positions, mitigating selective exposure and achieving a more balanced discussion around the issue. Even though a small sample size did not permit us to show that DebateCAFE helped participants change their minds on this particular issue, results suggest that DebateCAFE can reduce selective exposure bias without relying on high-cost human moderation.

3.5 Future Work

De-duplication

During the deployment, we found that argument duplication was a significant problem. Many participants provided similar arguments with slightly different wording. For example, among the "pro-Apple" arguments, many identified the potential that hackers could gain access to all Apple devices. Duplicate arguments should be consolidated to optimize the effectiveness and efficiency of participants' peer-ratings. One possible solution is to introduce a moderator, who reads and consolidates arguments. However, when the platform scales, it would be infeasible for the moderator to view every single argument. An alternative approach is to identify potentially similar arguments using Natural Language Processing techniques, and then enlist participants (in an interface similar to that used in the Argument Articulation Phase) to de-duplicate or synthesize those arguments.

Enhancing Value of Persuasiveness

As we mentioned before, the value for being persuasive varies for different individuals. To further improve the effectiveness of the persuasive scoring mechanism, we would like to tie this score back to the design of the platform. One approach is to introduce a positive relationship between argument exposure and user persuasiveness, where arguments provided by more persuasive users have a higher probability of being shown. Alternatively, we may completely hide those arguments provided by users with low persuasive score. Another approach is to add a public leader board with a list of top users ranked by their persuasive score.

Measuring Changes in Opinion

When participants rate persuasiveness of arguments, they might give high ratings to arguments that are logically coherent but are peripheral to the issue or of marginal importance. To better measure "persuasiveness" as opposed to "logical validity," we will experiment with a slider indicating the participant's position on the issue. It will be available throughout the discussion phase, allowing the participant to record her/his opinion change after viewing each argument. The persuasiveness of each argument will then be captured by the total opinion changes of other participants.

Part II

Healthcare Intervention Platforms

Chapter 4

The CalFit Mobile Phone App and the Cal Fitness Studies

4.1 Background and Related Work

Physical Inactivity

Physical inactivity is the fourth leading risk factor for mortality, causing an estimated 3.2 million deaths worldwide [217]. It is associated with cardiovascular disease, certain types of cancer, type 2 diabetes, and depression [103, 199, 181, 192]. Moderate- to vigorous-intensity physical activity, such as brisk walking or running, has significant health benefits across all age groups. The 2008 National Physical Activity Guideline for Americans recommends at least either 150 min of moderate-intensity physical activity or 75 min a week of vigorous-intensity physical activity for adults [83]. However, approximately half of American adults, particularly women and minorities, do not meet this physical activity guideline [186, 84].

Mobile Health Interventions

Several lifestyle modification programs that promote physical activity have been demonstrated to be effective, but these programs are costly and labor-intensive because they require substantial in-person counseling [54, 125, 124]. To lower costs, researchers have conducted randomized controlled trials (RCTs) to investigate the feasibility of mobile health (mHealth) interventions (e.g., mobile phone apps and digital pedometers) with reduced number of in-person counseling sessions [63, 59, 68, 88, 100, 148, 204, 187, 57]. Prior mHealth interventions implemented various goal-setting strategies to induce efforts, for example, to achieve and maintain 10,000 steps per day [14, 87, 176, 122, 158, 168] or meet adaptively increasing step goals [68, 164, 3, 2]. These studies demonstrated that mHealth interventions with goal setting can increase physical activity relative to baseline levels of activity.

Mobile Fitness Apps

Mobile fitness apps have the potential to be a scalable way of disseminating behavior change interventions in a cost-effective manner. In addition to being able to deliver interventions through wireless internet and messaging connectivity, smartphones can also leverage in-built tools like GPS, digital accelerometers, and cameras to objectively measure (as opposed to self-reported data) health parameters. Though many smartphone apps for fitness have been developed, systematic reviews [18, 24, 138, 207] of mobile fitness apps found an overall lack of persuasive attributes that are needed for the general public to maintain exercise motivation through continued use of the app. These reviews [24, 207] also identified a lack of experimental validation for the efficacy of specific features implemented in mobile fitness apps. The low efficacy of current mobile fitness apps is due primarily to this lack of inclusion of important features based on behavioral theory [18, 24, 138, 207]. Examples of key behavior change features include: objective outcome measurements, self-monitoring, personalized feedback, behavioral goal-setting, individualized program, and social support. In particular, researchers recommend that self-monitoring should be conducted regularly and in real-time, so as to target activity with precise tracking information and emphasize performance successes. In addition, personalized feedback is most effective when it is specific, such as in comparing current performance to past accomplishments and previous goals.

Goal Setting

Goal setting is known to be an important factor for facilitating behavior change [178, 126, 21], and effective goal setting requires self-monitoring to better enable attainment of goals and increase self-efficacy [68, 178, 126, 65]. There are three considerations regarding goal setting: (1) self-set goals versus assigned goals versus participatory goals, (2) adaptive goals versus fixed goals, and (3) personalized goals versus nonpersonalized goals. Despite the fact that self-set goals are of higher personal importance, a review of the goal-setting literature [178] reveals that assigned goals are more effective compared with self-set goals because self-set goals require regular input from participants, which is more difficult to maintain. Furthermore, more recent RCTs reveal that increases in physical activity through mobile-only programs with fixed, nonpersonalized physical activity goals are often substantially lower than increases in physical activity through programs that include adaptive goals [3, 2, 91, 95] or personalized goal setting provided during in-person counseling [91, 59, 179, 201, 212, 40]. For instance, one study [3] found that setting adaptive step goals resulted in an increase of 1130 more steps between baseline and 6 months, compared with setting fixed step goals of 10,000. Studies suspect that assigning nonpersonalized, fixed goals to all participants can lead to unrealistically high goals for some participants and unchallenging goals for other participants, which reduces goal-setting effectiveness [201, 159]. Therefore, assigning adaptively personalized goals can be a favorable alternative to better induce efforts and increase physical activity [159, 32, 50]. Personalized, adaptive goal setting allows changing goals over time based on prior individual behavior. For example, future daily step goals can be assigned based on step totals from the previous days to ensure that the goals are challenging yet realistic for each individual. Two trials [2, 3] used the same approach by combining financial incentives for meeting goals with an adaptive approach that set goals for the next day to be the 60th percentile of the steps taken in the past 10 days. Although this simple adaptive goal algorithm was modestly effective [2], a computer simulation study [140] for a weight loss intervention involving physical activity goal-setting and in-person counseling sessions found that a more sophisticated algorithm using statistics and machine learning to set goals by learning participants' responsiveness to goals could provide greater effectiveness (as compared with simple rules such as goal setting using a fixed percentile of steps taken in the past few days) in encouraging individuals to increase their physical activity and lose weight. In particular, the simulation showed that (when each participant received four counseling sessions) the more sophisticated machine learning algorithm would encourage almost half of the participants to have 5% or more body weight loss, whereas the use of goal setting using a fixed percentile of steps taken in the past few days would encourage only about one-quarter of the participants to have 5% or more body weight loss. Furthermore, previous studies have found that financial incentives may be effective during the intervention period, but in the maintenance period, participants are more likely to not adhere when no financial incentive is given [58, 141, 153].

Study Purpose

The purpose of our study was to test a sophisticated algorithm for personalized, adaptive goal setting that uses statistics and machine learning [140, 16], and specifically to examine its efficacy in two fully automated mobile phone-based intervention with no in-person contact or counseling sessions during the trial. It is important to note that goal setting is only one component of a behavior change intervention, and our study is designed to isolate the impact of goal setting from other components to evaluate the efficacy of goal setting alone. We developed an automated mobile phone-based iPhone operating system (iOS, Apple Inc) app named CalFit, which sets personalized, adaptive step goals using the behavioral analytics algorithm (BAA) [140, 16], and conducted two RCTs (called the Cal Fitness Studies) using this mobile phone app in the United States. BAA first uses machine learning to construct a predictive quantitative model for each participant based on the historical step and goal data, and then, it uses the estimated model to generate challenging yet realistic step goals in an adaptive fashion by choosing step goals that, based on the estimated model, would maximize future physical activity. The primary aim of these RCTs was to evaluate the efficacy of the automated mobile phone-based personalized, adaptive goal-setting intervention as compared with the active control with nonpersonalized, steady daily step goals of 10,000. The main outcome measure was the relative change in objectively measured daily steps between the run-in period and 10 weeks. Secondary outcome measures included the following: step goal attainment (i.e., fraction of step goals achieved by each participant), weight and height, self-reported sociodemographic information, self-reported medical history, Barriers to Being Active Quiz [39], and the short version of the international physical activity questionnaire [49]. We collected these survey results to investigate if the goal-setting component alone is capable of changing participants' survey responses before and after the study.

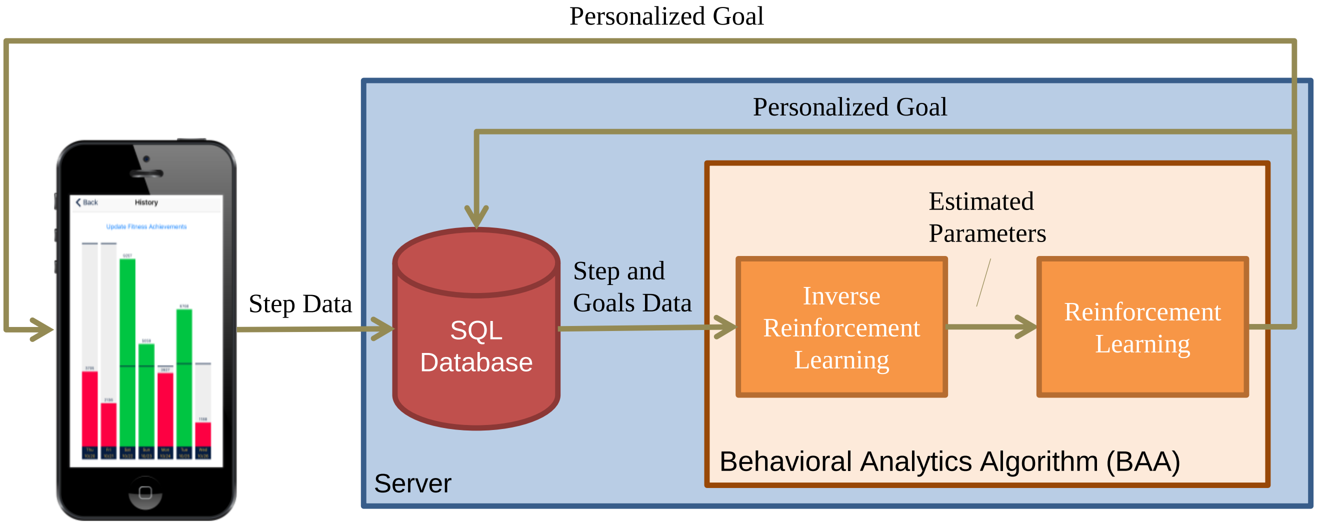


Figure 4.1: The CalFit app information flow. The CalFit app interface uploads step data to a SQL database on a server, and the stored step and goals data is accessed by the Behavioral Analytics Algorithm (BAA) comprised of inverse reinforcement learning to estimate model parameters describing the user and followed by reinforcement learning to compute personalized step goals that will maximize the user's future physical activity. The personalized step goals are stored in the SQL database and communicated to the user via the CalFit app interface.

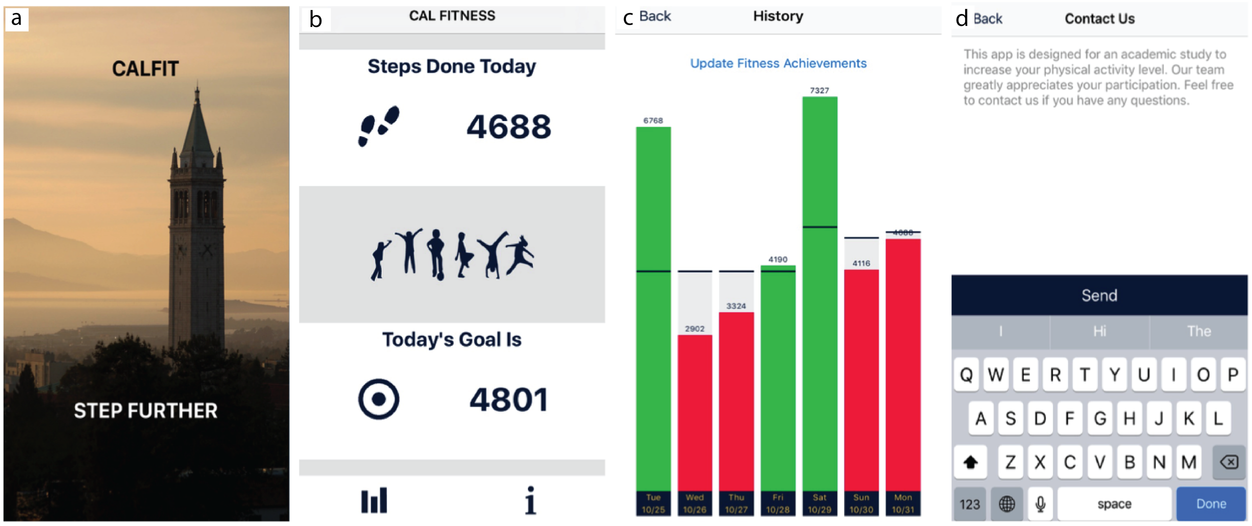
4.2 The CalFit App

CalFit is a mobile fitness app that uses key behavior change features to improve effectiveness. It combines a personalized goal setting algorithm and a structured interface with regular self-monitoring and feedback to provide an adaptive and individualized physical activity intervention. This section discusses the design of the interface, communication, and computation elements of our app, which are shown in Figure 4.1.

Interface

The CalFit app interface is built for the iOS platform. Upon opening the app interface, the user first sees the splash screen (Figure 4.2a) and then lands on the home tab (Figure 4.2b). On the home tab, the user can find his/her step goal for the day and the steps done so far today. The steps are tracked in real-time using the built-in health chip on the iPhone and are updated every 10-minutes. This design facilitates direct comparison between daily step goals and objectively measured daily steps in order to enhance self-monitoring.

There are two icons at the bottom of the home tab. If the left icon on the home tab is clicked, the user is shown the history tab (Figure 4.2c) that displays a barplot outlining the user's performance in the past 7 days. The black lines on each bar represent the step goal, and the height of each



This app is designed for an academic study to increase your physical activity level. Our team greatly appreciates your participation. Feel free to contact us if you have any questions.

Figure 4.2: Screenshots of the main tabs of the CalFit app, including the (a) splash screen, (b) home tab, (c) history tab, and (d) contact tab.

bar represents the actual measured steps. If the user achieved the goal, then the bar is green. If the user did not achieve the goal, then the bar is red. This tab is designed to provide a quick, yet comprehensive, visualization of the user's past performance, allowing the user to quickly identify days of successes and failures. If the right icon on the home tab is clicked, the user is directed to the contact tab (Figure 4.2d), where they can type in a message and send it to the research team regarding their concerns, app bugs, etc.

The built-in health chip in the iPhone collects the step data, and the accuracy of step counts collected by the iPhone health chip has been validated in a number of studies to have comparable accuracy to an ActiGraph [19, 38, 7, 8, 118, 211]. One of these studies [38] conducted a large number of experiments and concluded that iPhones are accurate for tracking step counts, with a relative difference in the mean step count of -6.7% to 6.2% compared with direct observation. Another study [211] compared iPhone pedometer measurements with measurements from wearable devices in a free-living setting and concluded "measurements of number of steps and distance were excellent and could provide reliable judgment on the individuals' activity amount." Our app first saves the step and goal data locally on the phone and then syncs with the server every 10 min when the phone is active. The push notification for the app is also activated, and the standard iOS push notification is used. The push notification is visible in the landing page and in the recent notifications tab on the phone.

Behavioral Analytics Algorithm (BAA)

Automated goal setting is a crucial component of the CalFit app. To set personalized goals that are challenging yet attainable for each user, we use a reinforcement learning algorithm [16, 140] that we have adapted to the context of physical activity interventions. The Behavioral Analytics Algorithm

(BAA) uses inverse reinforcement learning to construct a predictive quantitative model for each participant, based on the historical step and goal data for that user; then, it uses the estimated model with reinforcement learning to generate challenging but realistic step goals in an adaptive fashion.

Below, we elaborate upon the mathematical formulations underlying these steps of BAA. Since the BAA algorithm does calculations for each user independently of the calculations for other users, our description of the algorithm (and accompanying models) is focused on calculations for a single user.

Stage 0 – Predictive Model of User's Step Activity

Our predictive model is based on a model from [16, 140] for predicting weight loss based on steps and diet, and we have adapted that model to the specific case of only predicting step activity. Let the subscript t denote the value of a variable on the t-th day of using the app, and define the function $(x)^-$ as

$$(x)^- = \begin{cases} x, & \text{if } x \leq 0 \\ 0, & \text{if } x > 0 \end{cases} \tag{4.1}$$

Our predictive model for the number of steps that the user takes on the t-th day is

$$u_t = \arg\max_{u \geq 0} \; -(u - u_b)^2 + p_t \cdot (u - g_t)^-, \tag{4.2}$$

where u_t is the number of steps the user (subconsciously) decides to take, $u_b \in \mathbb{R}_+$ is a parameter describing the user's natural (or baseline) level of steps in a day, and $p_t \in \mathbb{R}_+$ is a parameter that quantitatively characterizes the user's responsiveness to the goal $g_t \in \mathbb{R}_+$.

The general idea of (4.2) is that users make decisions to maximize their utility or happiness related to several objectives. The $-(u-u_b)^2$ term means a user has an ideal level of steps they prefer to take in a day, wherein the user is implicitly trading off a small number of steps in a day (and the dissatisfaction accompanied by physical inactivity) with a large number of steps in a day (and the effort and time required to achieve many steps). The parameter u_b quantifies this baseline number of steps that achieves this tradeoff for the user. The $p_t \cdot (u - g_t)^-$ term means a user gets increasing happiness the closer their steps are to the goal g_t, and p_t describes the rate of increase in happiness as the steps get closer to the goal; however, this model says that exceeding the goal results in no additional happiness. A more complex model would include a term to describe an increase in happiness as the goal is exceeded, but a detailed study [16] found that not including this additional term still produced a model with high prediction accuracy.

There is one additional component to our predictive model. Equation (4.2) describes how a user decides the number of steps to take on the t-th day. The theory of goal setting [22, 126] recognizes that the effectiveness of goals can increase or decrease over time, depending on the level of the goals and whether or not an individual was able to meet the goals. To quantify these effects, our predictive model includes

$$p_{t+1} = \gamma \cdot p_t + \mu \cdot \mathbf{1}(u_t \geq g_t), \tag{4.3}$$

where $\gamma \in (0,1)$ characterizes the user's learned helplessness, $\mu \in \mathbb{R}_+$ quantifies the user's self-efficacy, and $\mathbf{1}(\cdot)$ is the indicator function. Self-efficacy is defined as a user's beliefs in their capabilities to successfully execute courses of action, and it plays an essential role in the theory of goal setting [22, 126]. Self-efficacy influences a variety of health behaviors, including physical activity [96, 134]. Though γ will be different for each individual, the past study [16] found that setting $\gamma = 0.85$ generated models with high prediction accuracy.

There are several points of intuition about (4.3). The term $\mu \cdot \mathbf{1}(u_t \geq g_t)$ describes the relationship between self-efficacy and meeting goals. When a user achieves a goal, $\mathbf{1}(u_t \geq g_t)$ is one and p_{t+1} increases by μ. Achieving a goal increases the user's self-efficacy, leading to increased steps on future days. But if the user misses a goal, then $\mathbf{1}(u_t \geq g_t)$ is zero and p_{t+1} does not increase. Not achieving a goal decreases the user's self-efficacy, leading to lower steps in the future. The term $\gamma \cdot p_t$ describes the phenomenon whereby learned helplessness reduces the utility or happiness an individual achieves for achieving goals. Consequently, (4.3) captures the interplay between increasing self-efficacy from meeting specific goals with the decrease in self-efficacy from learned helplessness.

Stage 1 – Inverse Reinforcement Learning

The BAA algorithm first uses inverse reinforcement learning to estimate the parameters u_b, p_t, μ in the predictive model (4.2), (4.3) for a user. Denoting n measurements of the user's step counts at times t_i as $\tilde{u}_{t_i}$, for $i = 1, \ldots, n$, our measurement model $\tilde{u}_{t_i} = u_{t_i} + \varepsilon_i$ is that the observed step counts $\tilde{u}_{t_i}$ deviate from the step counts chosen in the predictive model u_{t_i} by an additive zero mean random variable ε_i. The study [16] found that assuming ε_i has a Laplacian distribution led to an easily computable formulation and generated accurate predictions.

Under the above setup, the inverse reinforcement learning problem [15, 81, 111, 146] is equivalent to estimating the model parameters u_b, p_t, μ. This problem can be formulated as a log-likelihood maximization [16, 140]. If we define H to be the duration of the intervention, then we can write this estimation problem as a bilevel optimization problem

$$\begin{aligned} \min \ & \sum_{i=1}^{n} \left| u_{t_i} - \tilde{u}_{t_i} \right| \\ \text{s.t. } & u_t = \arg\max_{u \geq 0} \ -(u - u_b)^2 + p_t \cdot (u - g_t)^- \\ & p_{t+1} = \gamma \cdot p_t + \mu \cdot \mathbf{1}(u_t \geq g_t) \\ & 0 \leq p_t, \mu \leq \mathsf{UB}_p \\ & 0 \leq u_t, u_b \leq \mathsf{UB}_u \end{aligned} \tag{4.4}$$

where the constraints hold for $t = 1, \ldots, H$, and $\mathsf{UB}_p, \mathsf{UB}_u$ are constants that are upper bounds on the possible values. Existing numerical optimization software is not able to solve the above problem, but we can rewrite it as a mixed-integer linear program (MILP) [16, 140]. Let δ be a small positive constant, and M be a large positive constant. The above optimization problem can be rewritten as the following MILP:

$$
\begin{aligned}
\min\ & \sum_{i=1}^{n_u} a_{t_i} \\
\text{s.t.}\ & -a_{t_i} \leq u_{t_i} - \tilde{u}_{t_i} \leq a_{t_i} \\
& u_t = \tfrac{1}{2}(\lambda_{1,t} + \lambda_{3,t}) + u_b \\
& 0 \leq \lambda_{3,t} \leq p_t \\
& (g_t - \delta) - Mx_{1,t} \leq u_t \leq g_t - \delta + M(1 - x_{1,t}) \\
& (g_t - \delta) - M(1 - x_{2,t}) \leq u_t \leq g_t + \delta + M(1 - x_{2,t}) \\
& (g_t + \delta) - M(1 - x_{3,t}) \leq u_t \leq g_t + \delta + Mx_{3,t} \\
& p_t - M \cdot (1 - x_{t,1}) \leq \lambda_{3,t} \leq M \cdot (1 - x_{3,t}) \\
& u_t \leq My_{u,1} \\
& \lambda_{1,t} \leq M \cdot (1 - y_{u,t}) \\
& p_{t+1} \geq \gamma \cdot p_t \\
& p_{t+1} \leq \gamma \cdot p_t + M \cdot (1 - x_{t,1}) \\
& p_{t+1} \geq \gamma \cdot p_t + \mu - M \cdot x_{t,1} \\
& p_{t+1} \leq \gamma \cdot p_t + \mu \\
& x_{t+1,1} \geq x_{t,1} - \mathbf{1}(g_{t+1} - g_t < 0) \\
& x_{t+1,2} \leq x_{t,2} + \mathbf{1}(g_{t+1} - g_t < 0) \\
& x_{t+1,3} \leq x_{t,3} + \mathbf{1}(g_{t+1} - g_t < 0) \\
& x_{t,1} + x_{t,2} + x_{t,3} = 1 \\
& y_{u,t}, x_{t,1}, x_{t,2}, x_{t,3} \in \{0, 1\} \\
& \lambda_{1,t} \geq 0 \\
& 0 \leq p_t, \mu \leq \mathsf{UB}_p \\
& 0 \leq u_t, u_b \leq \mathsf{UB}_u
\end{aligned}
\tag{4.5}
$$

where the constraints hold for $t = 1, \ldots, H$ and $i = 1, \ldots, n$. The above MILP can be easily solved using standard optimization software [12, 80, 89].

Stage 2 – Reinforcement Learning

Under our setup, the reinforcement learning problem [140, 195, 196] for computing an optimal set of personalized goals for the user is equivalent to performing a direct policy search using the estimated model parameters $\hat{u}_b, \hat{p}_0, \hat{\mu}$ computed by solving (4.5). Adapting the solution in [140] to the current context of choosing an optimal sequence of step goals leads to a MILP:

$$
\begin{aligned}
\max\ & u_{min} \\
\text{s.t. } & u_{min} \leq u_t, \text{ for } t > T \\
& -\delta \leq u_t - \hat{u}_t \leq \delta, \text{ for } t \leq T \\
& -\delta \leq p_t - \hat{p}_t \leq \delta, \text{ for } t \leq T \\
& u_t = \tfrac{1}{2}(\lambda_{1,t} + \lambda_{3,t}) + \hat{u}_b \\
& 0 \leq \lambda_{3,t} \leq p_t \\
& (g_t - \delta) - M x_{1,t} \leq u_t \leq g_t - \delta + M(1 - x_{1,t}) \\
& (g_t - \delta) - M(1 - x_{2,t}) \leq u_t \leq g_t + \delta + M(1 - x_{2,t}) \\
& (g_t + \delta) - M(1 - x_{3,t}) \leq u_t \leq g_t + \delta + M x_{3,t} \\
& p_t - M(1 - x_{t,1}) \leq \lambda_{3,t} \leq M(1 - x_{3,t}) \\
& u_t \leq M y_{u,1} \\
& \lambda_{1,t} \leq M(1 - y_{u,t}) \\
& p_{t+1} = \gamma p_t + \hat{\mu}(1 - x_{1,t}), \text{ for } t > T \\
& x_{t+1,1} \geq x_{t,1} - g_{ind,t}, \text{ for } t > T \\
& x_{t+1,2} \leq x_{t,2} + g_{ind,t}, \text{ for } t > T \\
& x_{t+1,3} \leq x_{t,3} + g_{ind,t}, \text{ for } t > T \\
& x_{t,1} + x_{t,2} + x_{t,3} = 1 \\
& g_{t+1} - g_t \leq M(1 - g_{ind,t}), \text{ for } t > T \\
& g_{t+1} - g_t \geq -M g_{ind,t}, \text{ for } t > T \\
& y_{u,t}, x_{t,1}, x_{t,2}, x_{t,3}, g_{ind,t} \in \{0, 1\}, \text{ for } t > T \\
& \lambda_{1,t} \geq 0 \\
& 0 \leq p_t \leq \mathsf{UB}_p \\
& 0 \leq u_t \leq \mathsf{UB}_u
\end{aligned}
\tag{4.6}
$$

where T is the current time, and the remaining constraints hold for $t = 1, \ldots, H$ and $i = 1, \ldots, n$. The intuition is that the above MILP picks future goals in order to maximize the smallest number of steps on any given day in the future, and the reason for this choice is that in our simulations we found that this objective function choice led to the largest increases (as compared to other possible objective function choices) in physical activity. Moreover, the above MILP can be easily solved using standard optimization software [12, 80, 89].

4.3 The Cal Fitness Studies

We conducted two 10-week long Studies to evaluate the efficacy of the CalFit app. The first study was conducted with university staff members whereas the second study was conducted with university students. We will refer to the studies as Cal Fitness1 and Cal Fitness2.

Cal Fitness1

Methods

Study Design

The Cal Fitness1 study was a 10-week RCT with 2 groups: (1) the intervention group received automated personalized daily step goals, and (2) the control group received fixed daily step goals of 10,000 steps per day. The study was approved by the Committee for Protection of Human Subjects at the University of California, Berkeley (UCB; institutional review board number 2016-03-8609), in July 2016 and was registered with the clinicaltrials.gov (NCT02886871) in August 2016. All participants provided written informed consent before study enrollment. This RCT was conducted in 2016 and analyzed in 2017.

Participant Recruitment

A total of 64 adult staff employees of UCB were recruited via email announcements. Recruitment commenced in August 2016 and ended in September 2016. The study ended in December 2016 to allow a 10-week period to all participants. Potential participants were contacted through email and then directed to a Web-based screening survey to assess eligibility. Those participants who met all the inclusion criteria were then contacted by trained study personnel via email to arrange an in-person session. Ineligible participants were informed by email to advise them that they are ineligible, and corresponding data were deleted. The inclusion and exclusion criteria for the Cal Fitness study are below:

Inclusion criteria

- Staff member of University of California, Berkeley
- Intent to become physically active in the next 10 weeks, which was evaluated by asking potential participants if they wanted to increase their physical activity beyond their self-assessed current level
- Own an iPhone 5s (or a newer model)
- Willing to keep the iPhone in pockets during the day
- Willing to install and use the study app (which requires internet connection) every day for 10 weeks
- Ability to speak and read English

Exclusion criteria

- Known medical conditions or physical problems that require special attention in an exercise program

- Planning an international trip during the next 3 months, which could interfere with daily server uploads of mobile phone data
- Pregnant or gave birth during the past 6 months
- Severe hearing or speech problem
- History of an eating disorder
- Current substance abuse
- Current participation in lifestyle modification programs or research studies that may confound study results
- History of bariatric surgery or plans for bariatric surgery in the next 12 months

Study Procedure

Eligible participants were asked to attend two 15-min in-person sessions (initial and 10-week post intervention visits) at UCB. The first in-person session occurred in September 2016, and the second session occurred in December 2016. During the first in-person session, a trained research staff member installed the CalFit app on participants' phones and advised the participants to keep the phone in their pocket or purse for the following 10-week period. A trained research staff member measured height (cm) and weight (kg) in both the sessions using a Seca 700 Physician's Balance Beam Scale with Height Rod, and body mass index (BMI) was also calculated. Participants were then instructed to complete the sociodemographic survey, the medical history survey, the Barriers to Being Active Quiz [201], and the short version of the international physical activity questionnaire [212]. During the second in-person session, a trained research staff removed the CalFit app from participants' phones. Participants were then instructed to complete the Barriers to Being Active Quiz [201] and the short version of the international physical activity questionnaire [212]. Participants received a US $50 Amazon gift certificate at completion if they completed all study requirements.

Run-In Period and Randomization

A total of 64 eligible participants started a 1-week run-in period after completing the initial in-person session. The purpose of the run-in period was to collect run-in daily steps, and assess if the participant was able to comply with the requirements needed to regularly use the CalFit app. During the run-in period, all participants in the control and the intervention groups received the identical set of daily step goals for day 1 to day 7 as 3000, 3500, 4000, 4500, 5000, 5500, and 6000 steps, respectively. The BAA algorithm was not used to compute step goals for participants in the intervention group during the run-in period. Dynamically increasing step goals were used in the run-in period to engage participants in using the app regularly. In addition, all participants received a push notification at 8 AM that provided today's step goal, and if the participant accomplished the goal before 8 PM, then another push notification was sent to congratulate the participant on reaching their step goal for the day. The identical goals between the 2 groups during the run-in period is used to establish a reference level of initial physical activity, which we used in our statistical analyses to compare the difference in daily steps between run-in and 10 weeks for the 2 groups. Data collected during the run-in period were used by the BAA algorithm to compute step goals for the intervention period. This is a valid approach because run-in data were indicative of the preference of different participants. All 64 participants were randomized to one of the 2 groups with a one-to-one ratio by a computer-based random number generator using the simple randomization approach. A one-to-one ratio means that each participant had a 50% probability of being assigned to one of the 2 groups, and the number of participants in each group may differ due to chance. The randomization to groups was implemented by the CalFit app after the run-in period, and the participants were aware of the 2 groups.

Control

After the 1-week run-in period, participants in the control group were provided with constant daily step goals that were set to 10,000 steps per day through the CalFit app. Participants received a push notification at 8 AM every day that provided that day's step goal (10,000 steps), and if the participant achieved the goal before 8 PM, then another push notification was sent to congratulate the participant on reaching their step goal (of 10,000 steps) for the day.

Intervention

After the 1-week run-in period, participants in the intervention group received adaptively personalized step goals through the CalFit app. The daily step goals were computed using the BAA [140, 16] on the complete history (past steps and goals) of the user. The BAA algorithm was applied every week to reduce variance in future steps and goals. Participants received a push notification at 8 AM every day that provided today's step goal, and if the participant accomplished the goal before 8 PM, then another push notification was sent to congratulate the participant on reaching their step goal for the day. A rigorous mathematical formulation of the BAA algorithm that we used is provided in 2 studies [140, 16]. This algorithm uses statistics and machine learning to adaptively compute personalized step goals that are predicted to maximize future physical activity for each participant based on all the past steps' data and goals of each participant. The BAA algorithm is applied to each participant individually, and it consists of two main steps. The first step is to use all of the participant's data to construct a quantitative model that predicts how many steps the participant will take in the future, given a prescribed set of step goals, and an important aspect of the model is a component that describes how achieving goals in the present can increase the likelihood of achieving goals in the future. The second step is to use this quantitative model to select a sequence of step goals that maximizes the predicted future number of steps. To make the process of updating step goals adaptive, the BAA algorithm is applied each week (using all the users' past data) to generate step goals for the coming week; moreover, the step goals computed by the BAA algorithm for the coming week are not constant, but increase or decrease based on the model prediction. A computer simulation study [140] found that applying the algorithm weekly is as effective as applying the algorithm daily (because steps can vary significantly on a day-to-day basis), and so we applied the algorithm weekly.

Outcome Measures

The primary outcome of the study was the relative change in daily steps from run-in to the 10-week follow-up, measured objectively by the participants' iPhones. The daily step values were compared in the manner described in the statistical analysis section. Step count data were stored in a database on a private computer server at UCB. Data were automatically synced with the iPhone once every 10 min during the study. At the 10-week in-person session, complete step data were downloaded from the iPhone to store step count data that were unable to be transmitted. Data were unable to be transmitted if the app was turned off or no Internet connection was available. Other measures included weight and height, self-reported sociodemographic information, self-reported medical history, Barriers to Being Active Quiz [39] (which consists of 21 questions on a 10-point Likert scale on 7 subareas: lack of time, social influence, lack of energy, lack of willpower, fear of injury, lack of skill, and lack of resources), and the short version of the international physical activity questionnaire [49].

Statistical Analysis

Assuming an expected loss to follow-up of 10%, a target sample size of 30 participants per group was selected to give 80% power to detect between-group difference of 1500 steps with a pooled standard deviation (SD) of 2000 using a two-sided test and an alpha of .05. Differences between groups in run-in and 10 weeks were assessed using Student t test. The statistical analysis of the primary outcome of daily steps was performed using a linear mixed-effects model (LMM) with piecewise linear growth curve [144, 198, 157] with random effects for each individual of random slope and random intercept, and fixed effects of time, treatment group, and interaction term of time and treatment group. Our statistical analysis of the secondary outcome of step goal attainment (i.e., fraction of step goals achieved by each participant) was performed by a similar LMM but with an additional specification of a binary response variable (i.e., goal is either attained or not attained by an individual on a particular day). Means with 95% confidence intervals were obtained from the LMM. Sensitivity analysis was performed to obtain adjusted estimates of the effect of the treatment with the missing data on primary outcome, evaluated at $p < .05$. The primary cause of missing step data was failure to turn on the app. LMM implicitly imputes missing data by interpolation and is a common approach to deal with missing data in physical activity interventions [144, 198, 157, 206, 114]. (We did not use the common imputation method of "last observation carried forward "because it would increase bias in this context and lead to potentially false conclusions by inflating step counts at 10 weeks.) For accurate comparison between the control and the intervention groups, the weekly average steps in run-in were adjusted by adding the coefficient corresponding to each group (i.e., control or intervention) computed by the LMM model. In addition, weekly moving average steps were computed by taking the average of each moving window with length 7, to reduce noise for better visualization. To quantify app use for the intervention, a participant was categorized as a

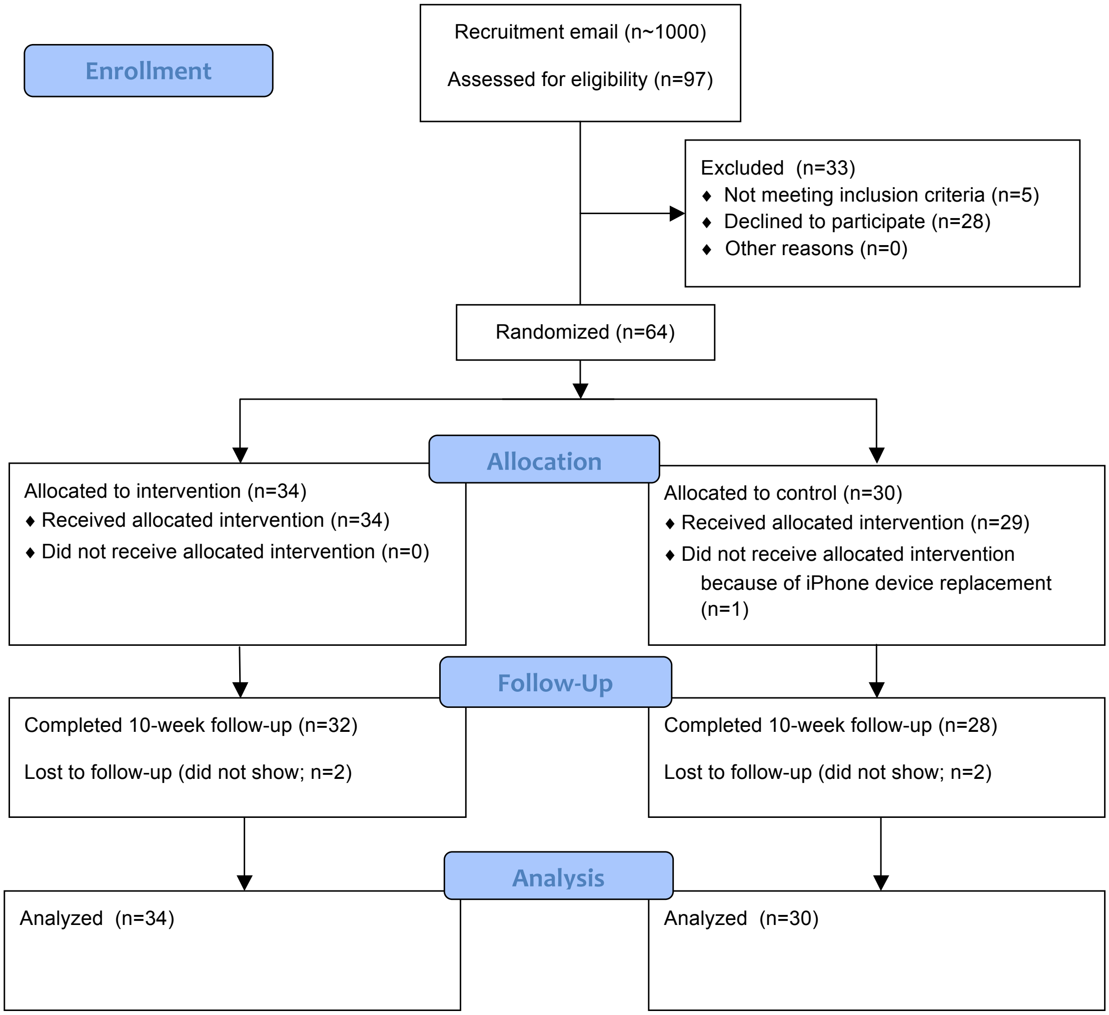

Enrollment
Recruitment email (n~1000)
Assessed for eligibility (n=97)
Excluded (n=33)
♦ Not meeting inclusion criteria (n=5)
♦ Declined to participate (n=28)
♦ Other reasons (n=0)
Randomized (n=64)
Allocation
Allocated to intervention (n=34)
♦ Received allocated intervention (n=34)
♦ Did not receive allocated intervention (n=0)
Allocated to control (n=30)
♦ Received allocated intervention (n=29)
♦ Did not receive allocated intervention because of iPhone device replacement (n=1)
Follow-Up
Completed 10-week follow-up (n=32)
Lost to follow-up (did not show; n=2)
Completed 10-week follow-up (n=28)
Lost to follow-up (did not show; n=2)
Analysis
Analyzed (n=34)
Analyzed (n=30)

Figure 4.3: Screening, randomization, and assessments of study participants for Cal Fitness study 1.

nonfrequent app user if the app was not used for a consecutive period of 7 days. By this criterion, 17 participants out of 34 in the intervention group and 16 participants out of 30 in the control group were frequent app users. Per-protocol analysis was performed on the 33 frequent app users, and intention-to-treat analysis was performed on all 64 subjects. Although the power for the per-protocol analysis will be low, the reason for conducting this analysis is that we want to investigate the impact of the CalFit app on an active subgroup, which could be more representative for its true performance if adopted in other full interventions that include additional components of a behavior change intervention. Intention-to-treat analysis was performed for the primary and secondary outcomes, and per-protocol analysis was performed only for the primary outcome. Missing survey response data resulting from lost to follow-up was imputed by the latest available survey response of the subject. The statistical analysis was performed in MATLAB (MathWorks, Massachusetts, USA) version 9.0 [133] and R (R Core Team, Vienna, Austria) version 1.0.136 [163] in year 2017.

Results

Recruitment Results

As shown in Figure 4.3, 97 potential participants were screened for eligibility by an online form, and 64 completed the initial in-person session.

Baseline Characteristics

Table 4.1 shows the baseline characteristics of the participants. A total of 34 participants were randomly assigned to the intervention group, and 30 participants were randomly assigned to the control group. All participants were included in the analysis based on the original assigned groups. Overall mean age was 41.1 (SD±11.3) years, and 83% (53/64) participants were female. In addition, 55% (35/64) of the participants self-identified as a member of a racial minority group. The baseline mean weight of participants was 77.2 kg (SD±18.7 kg) and the mean BMI was 27.3 kg/m^2(SD±6.1 kg/m^2). The mean height and weight for male and female participants were 177.5 cm and 82.6 kg and 165.9 cm and 76.1 kg, respectively. Furthermore, 20% of the participants reported at least one medical condition (i.e., high blood pressure, type 2 diabetes, type 1 diabetes, coronary heart disease, or hypercholesterolemia). No baseline characteristics differed between the control and intervention groups. The run-in mean daily steps in the control and intervention groups were similar (7427 steps vs 7237 steps, respectively; $p = .79$) and are in line with baseline steps in other similar studies [69, 50, 201, 191]. As shown in Table A.1, the self-reported survey results did not differ considerably between the 2 groups except for the lack of resources, which is a subscale of the Barriers to Being Active measure. The intervention group had a significantly higher rating of lack of resources than the control group ($p = .03$). We suspect this significant difference for lack of resources occurred due to chance.

Table 4.1: Baseline characteristics between the control and intervention groups in Cal Fitness study 1.

Baseline characteristics	All participants (N=64)	Control (N=30)	Intervention (N=34)	P
Run-in daily average steps, mean (SD)	7326 (2907)	7427 (2398)	7237 (3326)	0.79
Age, years, mean (SD)	41.1 (11.3)	40.5 (10.5)	41.6 (12.2)	0.72
Weight, kg, mean (SD)	77.2 (18.7)	77.8 (21.3)	77.0 (17.1)	0.87
BMI, kg/m2, mean (SD)	27.3 (6.1)	27.1 (6.7)	27.4 (5.8)	0.82
Gender, n (%)				0.82
Male	11 (17)	6 (20)	5 (15)	
Female	53 (83)	24 (80)	29 (85)	
Ethnicity, n (%)				0.86
Asian	13 (20)	7 (23)	6 (18)	
Black or African American	8 (13)	3 (10)	5 (15)	
Hispanic or Latino	9 (14)	5 (17)	4 (12)	
White or non-Hispanic	29 (45)	13 (43)	16 (47)	
Other	5 (8)	2 (7)	3 (9)	
Marital status, n (%)				0.2
Currently married or cohabitating	36 (56)	15 (50)	21 (62)	
Never married	21 (33)	13 (43)	8 (24)	
Divorced or widowed	7 (11)	2 (7)	5 (15)	
Education, n (%)				0.3
Completed some college	5 (8)	1 (3)	4 (12)	
Completed college (4 years)	28 (44)	12 (40)	16 (47)	
Completed graduate school	31 (48)	17 (57)	14 (41)	
Work hour (per week), n (%)				0.17
1-20 hours	3 (5)	3 (10)	0 (0)	
21-40 hours	16 (25)	7 (23)	9 (27)	
>40 hours	45 (70)	20 (67)	25 (74)	
Own a dog, n (%)				0.99
Yes	16 (25)	8 (27)	8 (24)	
No	48 (75)	22 (73)	26 (77)	
Transportation to work, n (%)				0.49
Car	28 (44)	10 (33)	18 (53)	
Public transportation	25 (39)	14 (47)	11 (32)	
Walk	4 (6)	2 (7)	2 (6)	
Bicycle	6 (9)	3 (10)	3 (9)	
Other	1 (2)	1 (3)	0 (0)	
Gym membership, n (%)				0.45
Yes	32 (50)	13 (43)	19 (56)	
No	32 (50)	17 (57)	15 (44)	
Self-reported medical history, n (%)				
High blood pressure				0.88
Yes	5 (8)	3 (10)	2 (6)	
No	59 (92)	27 (90)	32 (94)	
Type 2 diabetes				0.43
Yes	5 (8)	1 (3)	4 (12)	
No	59 (92)	29 (97)	30 (88)	
Type 1 diabetes				0.62
Yes	0 (0)	0 (0)	0 (0)	
No	64 (100)	30 (100)	34 (100)	

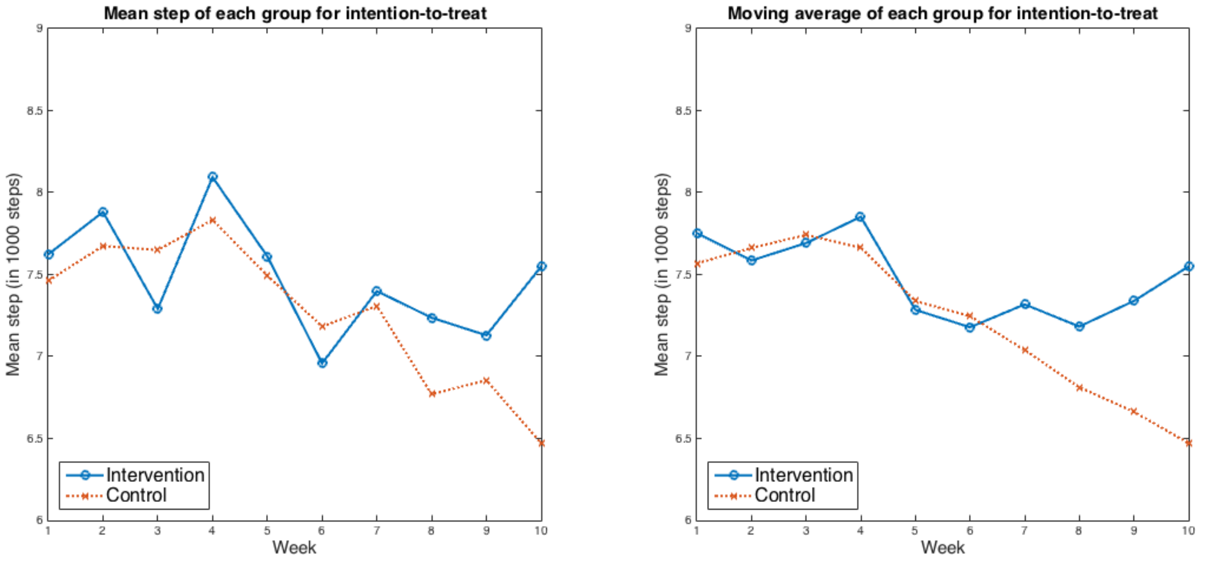


Figure 4.4: Weekly average and moving average steps for the 2 groups over the course of the study for intention-to-treat analysis after run-in adjustment. Left panel: Mean weekly steps for intention-to-treat; Right panel: Weekly moving average for intention-to-treat.

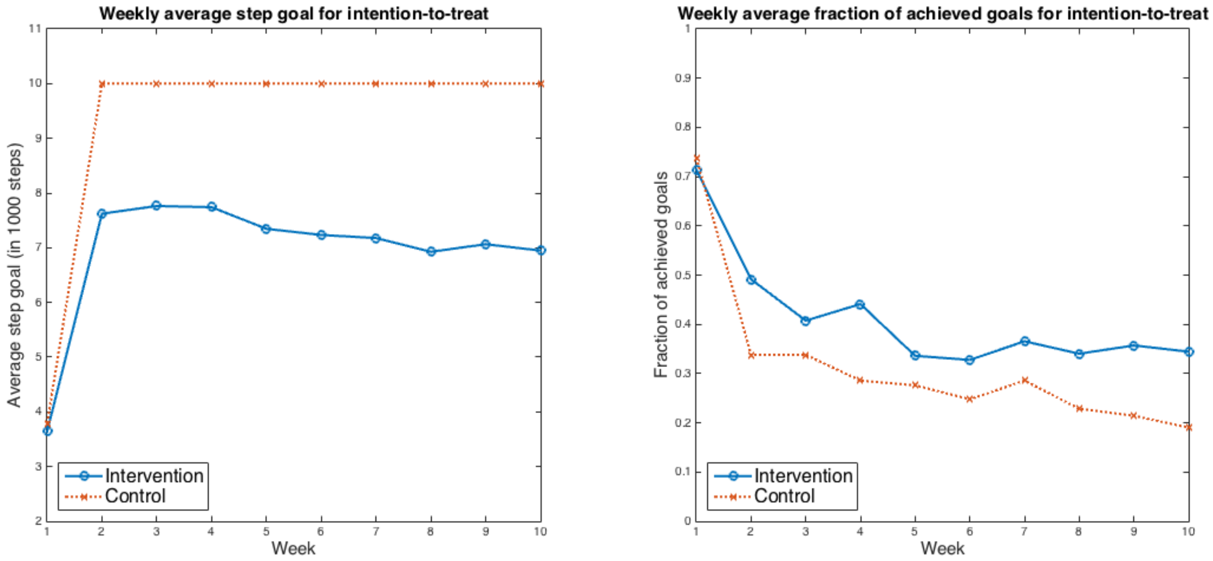


Figure 4.5: Weekly average step goals and average fraction of goals achieved for the 2 groups for intention-to-treat analysis. Left panel: Weekly average step goals for intention-to-treat; Right panel: Weekly average fraction of achieved goals for intention-to-treat.

Table 4.2: Run-in adjusted objectively recorded (using iPhone) physical activity in Cal Fitness study 1.

Week	Mean number of steps	
	Control (N=30)	Intervention (N=34)
Week 1 (Run-in)	7462	7623
Week 2	7674	7882
Week 3	7650	7290
Week 4	7834	8094
Week 5	7494	7611
Week 6	7183	6958
Week 7	7308	7399
Week 8	6770	7237
Week 9	6855	7129
Week 10	6471	7549

Table 4.3: Fraction of achieved daily step goals in 10 weeks in Cal Fitness study 1.

Week	Control (N=30)	Intervention (N=34)
Week 1 (run-in)	0.74	0.71
Week 2	0.34	0.49
Week 3	0.34	0.41
Week 4	0.29	0.44
Week 5	0.28	0.34
Week 6	0.25	0.33
Week 7	0.29	0.37
Week 8	0.23	0.34
Week 9	0.21	0.36
Week 10	0.19	0.34

Efficacy of Intervention

Main Analysis Intention-to-treat analyses indicated that the intervention group had a decrease in mean (SD) daily step count of 390 (SD±490) steps between run-in and 10 weeks compared with a decrease of 1350 (SD±420) steps among controls (P=.03). The net difference in daily steps between the groups was 960 steps (95% CI 90-1830 steps). Table 4.2 shows the run-in adjusted objectively measured raw average weekly steps for both the groups without missing data imputation. 4.4 shows the run-in adjusted weekly average steps and moving average steps for intention-to-treat. The average step goals for the first week are the same for both the control and the intervention groups because both received the same goals during the first week. Table 4.3 gives the fraction of achieved step goals for the 2 groups. Intention-to-treat analysis indicated that the intervention group had a decrease in mean fraction of achieved step goals of 0.34 (SD±0.05) between run-in and 10

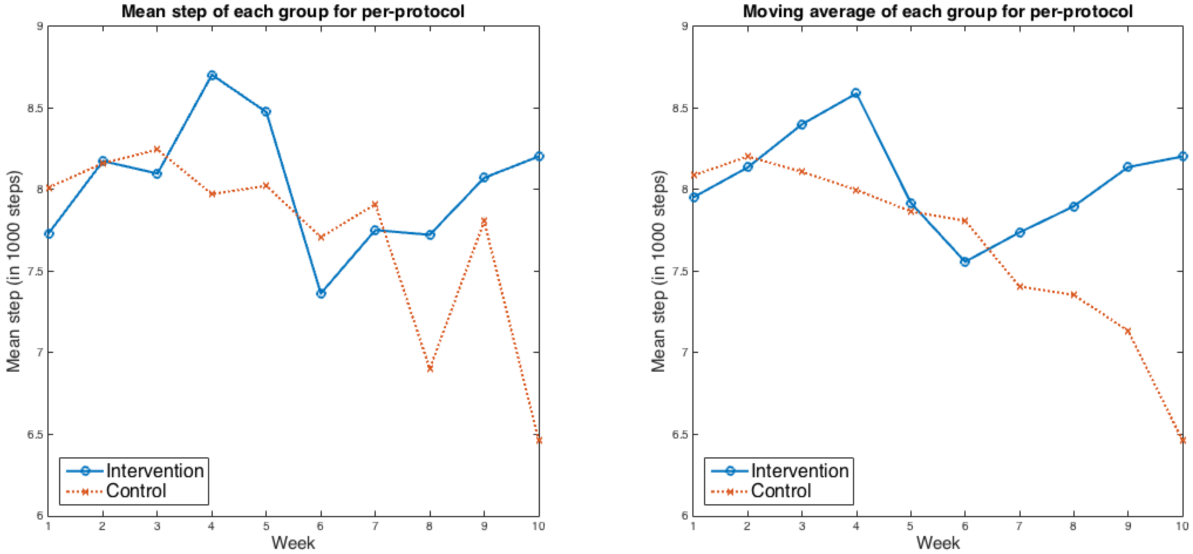


Figure 4.6: Weekly average and moving average steps for the 2 groups over the course of the study for per-protocol analysis after run-in adjustment. Left panel: Mean weekly steps for per-protocol; Right panel: Weekly moving average for per-protocol.

weeks compared with a decrease of 0.49 (SD±0.04) among controls (P=.003). The net difference in fraction of achieved step goals between the groups was 0.15 (95% CI 0.02-0.25). 4.5 details the intention-to-treat weekly average step goals and the fraction of achieved step goals for the 2 groups.

Sensitivity Analysis Per-protocol analysis (among the 33 frequent app users: 16 in control and 17 in intervention groups) indicated that the intervention group had a decrease in mean (SD) daily step count of 0 (SD±420) steps between run-in and 10 weeks, whereas the control group had a decrease of 1500 (SD±550) steps ($p = .03$). The net difference in daily steps between the groups was 1500 steps (95% CI 130-2900 steps). Figure 4.6 shows the run-in adjusted weekly average steps and moving average steps for per-protocol. Per-protocol analysis also indicated that the intervention group had a decrease in mean (SD) fraction of achieved step goals of 0.27 (SD±0.08) between run-in and 10 weeks compared with a decrease of 0.46 (SD±0.06) among controls ($p = .02$). The net difference in fraction of achieved step goals between the groups was 0.19 (95% CI 0.02-0.38). Figure 4.7 details the per-protocol weekly average step goal and the fraction of achieved step goals for the 2 groups.

Cal Fitness2

To experimentally evaluate the efficacy of the CalFit app and personalized goal setting using the BAA algorithm with different audience, we conducted the CalFitness 2 study (also called the Mobile Student Activity Reinforcement (mSTAR) study) with college students in University of

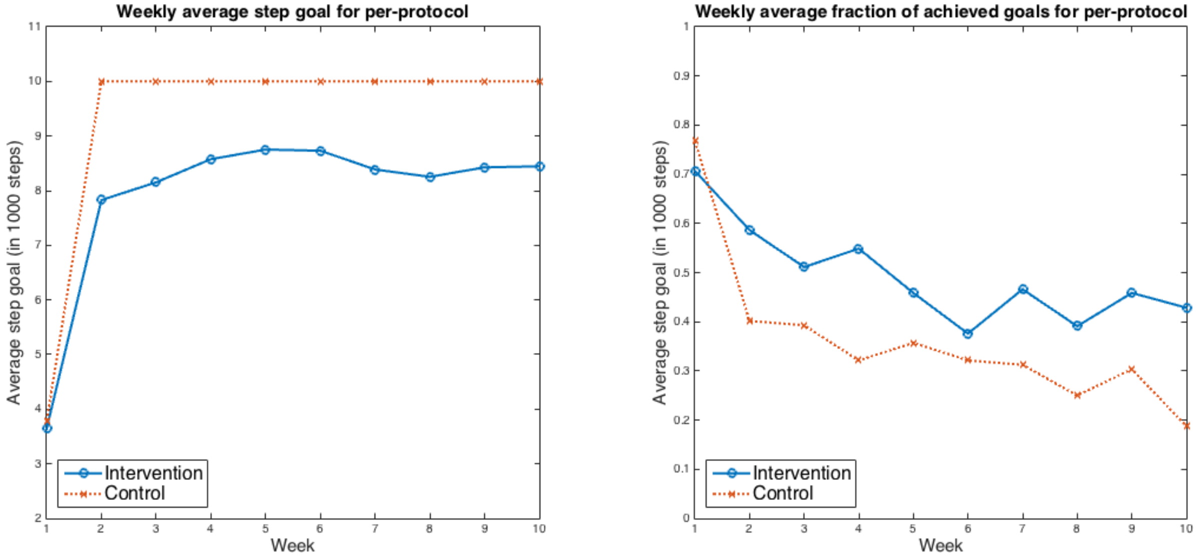


Figure 4.7: Weekly average step goals and average fraction of goals achieved for the 2 groups for per-protocol analysis. Left panel: Weekly average step goals for per-protocol; Right panel: Weekly average fraction of achieved goals for per-protocol.

California, Berkeley (UCB). The main research question was: Does setting personalized step goals increase user's steps compared to fixed step goals? The secondary research question was: Does setting personalized step goals improve adherence? The study was approved by the Committee for Protection of Human Subjects of the University of California, Berkeley (IRB Number 2016-03-8609) in July 2016. All participants provided written informed consent prior to study enrollment.

Methods

The methodology and the study procedure of Cal Fitness2 were identical to Cal Fitness1. We excluded students who took 20,000 steps per day because it is not possible to increase activity by using our app if they were at that activity level (since the BAA algorithm uses 20,000 steps as the upper bound for the goal), and the procedure was that students satisfying the other criteria were enrolled and then excluded if 20,000 steps was observed in the step data collected. After the 1-week run-in period, the daily step goals for users in the control group (N=7) were set to 10,000 steps/day through the CalFit app, whereas the daily step goals for users in the intervention group (N=6) were set by the BAA algorithm. The BAA algorithm was applied every week (to mitigate the impact of large step variance), and it computes the step goals for the following 7 days. Both groups received morning and evening push notifications. The study lasted for 10-weeks, and participants could earn up to a $25 Amazon gift card for completing all parts of the study, including attending a final in-person session.

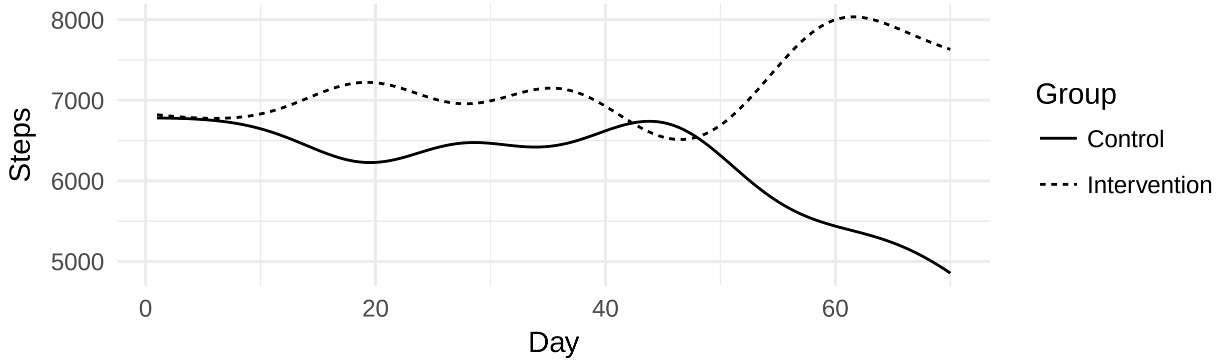


Figure 4.8: The objectively measured daily steps of the control group and the intervention group over the 10-week study period show the statistically significant difference in the number of daily steps at the end of the study. The plotted values are computed by averaging the raw data over each user in the corresponding group, adjusting the baseline value based on the value computed from the LMM model, and then smoothing the data using a standard (nonparametric) Nadaraya-Watson estimator.

Results

Baseline Characteristics

Table 4.3 shows the baseline characteristics of the participants. The overall mean age was 22.2 (SD±2.9) years and 77% of the participants were female. The baseline mean daily step in the control group was slightly higher than that in the intervention group, but the difference is not statistically significant (6,829 steps versus 5,387 steps, respectively; $p = 0.16$). The p-values in Table 1 were computed using t-tests for continuous variables and χ^2-tests for categorical variables.

Efficacy of Intervention

The primary outcome of the study is the objectively measured daily steps from baseline to 10-weeks. We conducted our statistical analysis of the primary outcome of daily steps using a linear mixed-effects model (LMM) with random effects for each individual of random slope and random intercept, and fixed effects of time, intervention group, and interaction term of time and intervention group. This analysis found that the control group had a decrease in daily step count of 1520 (SD±740) steps between baseline and 10-weeks, compared to an increase of 700 (SD±830) steps in the intervention group. The difference in daily steps between the two groups was 2220 ($p = 0.039$) with a 95% confidence interval of (100, 4480), which is a statistically significant difference. The step goals computed by the BAA algorithm were on average between 6,000 steps and 8,000 steps. They varied between different users and days resulting from its adaptive and personalized nature.

Figure 4.8 shows the change in daily steps over the 10-week study period, and for fair comparison we baseline-adjusted the plotted steps by adding the coefficient corresponding to each group (i.e.,

	All Users (N=13)	**Control (N=7)**	**Intervention (N=6)**	*p*-value
	Mean (± SD)	**Mean (± SD)**	**Mean (± SD)**	
Baseline daily average steps	6,163 ± 1,822	6,829 ± 2,023	5,387 ± 1,309	0.16
Age (years)	22.2 ± 2.9	21.6 ± 2.3	23.0 ± 3.5	0.40
Weight (kg)	70.4 ± 23.9	73.7 ± 31.9	66.5 ± 20.8	0.61
	% (N)	**% (N)**	**% (N)**	
Gender				0.88
Male	23.1 (3)	14.3 (1)	33.3 (2)	
Female	76.9 (10)	85.7 (6)	66.6 (4)	
Ethnicity				0.85
Asian	23.1 (3)	28.6 (2)	16.7 (1)	
Hispanic/Latino	15.4 (2)	14.3 (1)	16.7 (1)	
White (non-Hispanic)	23.3 (3)	28.6 (2)	16.7 (1)	
Other	38.5 (5)	28.6 (2)	50.0 (3)	
Marital Status				1.00
Currently Married	7.7 (1)	14.3 (1)	0.0 (0)	
Never Married	92.3 (12)	85.7 (6)	100.0 (6)	
Divorced/Widowed	0.0 (0)	0.0 (0)	0.0 (0)	
Year in School				0.52
Freshman	0.0 (0)	0.0 (0)	0.0 (0)	
Sophomore	15.4 (2)	28.6 (2)	0.0 (0)	
Junior	30.8 (4)	28.6 (2)	33.3 (2)	
Senior	23.1 (3)	14.3 (1)	33.3 (2)	
Graduate	30.8 (4)	28.6 (2)	33.3 (2)	
Own a Dog				1.00
Yes	7.7 (1)	14.3 (1)	0.0 (0)	
No	92.3 (12)	85.7 (6)	100.0 (6)	
Transportation to Work				0.43
Car	23.1 (3)	28.6 (2)	16.7 (1)	
Public Transportation	7.7 (1)	14.3 (1)	0.0 (0)	
Walk	61.5 (8)	42.9 (3)	83.3 (5)	
Bicycle	7.7 (1)	14.3 (1)	0.0 (0)	

Table 4.4: Comparison of baseline characteristics shows that the differences between participants in the control and intervention groups were not statistically significant, which is expected since participants were randomly assigned to groups.

control or intervention) computed by the LMM model. Despite the slightly higher steps in the intervention group, the daily steps of the two groups did not differ substantially in the first 5 weeks. However, in the last 3 weeks, the intervention group had an average increase of 1,000 steps and the control group had an average decrease of 2,000 steps. We suspect that we fail to see differences in the early weeks due to the initial stimulation of participating in a fitness program. As time went by, the excitement from participation cooled down and the impact of the BAA algorithm started to dominate.

We further defined adherent users to be those who used the CalFit app for 80% of the days during the study period. Under this criterion, 2 of the 7 users in the control group and 1 of the 6 users in the intervention group were identified as non-adherent. However, the difference in adherence percentage was not statistically significant ($p = 0.61$) between the two groups, primarily due to the small sample size.

During the second in-person session at 10-weeks, a trained research staff member interviewed the users on their experience. All users agreed that the CalFit app was easy to navigate, required minimal effort on the user side, and the number of push notifications was about right. One user in the intervention group told us, “I am excited to know my step goal every morning! I know I am doing well if my goal increases, and I know I need to keep up when my goal decreases.” Another user in the control group, however, stated, “The goals are always the same. It’s impossible for me to get that many steps so I stopped tracking.”

4.4 Discussion

These two studies evaluated the efficacy of mobile phone-based physical activity interventions that provided adaptively personalized daily step goals. The interventions both led to a statistically significant difference in the intervention group compared with the control group over 10 weeks, in line with similar studies [2, 3]. Although both groups in the studies had reduced daily steps at 10 weeks as compared with run-in, we speculate this was caused by run-in step counts being higher than the natural baseline. We believe this inverse relationship was a result of participants receiving step goals and monitoring step count through the CalFit app or the built-in iPhone Health app during the run-in period. This is supported by the observations that during the run-in period, all participants received daily step goals of 3000, 3500, 4000, 4500, 5000, 5500, and 6000 steps and initially over-responded to these goals, and that the trends in daily steps between the control and intervention groups began to diverge in the 6th week of the study when enthusiasm of study participants wore out. Thus, later in the studies, the personalized daily step goals seemed to be more effective in engaging participants and maintaining daily step counts compared with constant step goals.

The health literature has identified that setting goals is effective in lifestyle modification and physical activity promotion [54, 68, 179, 11]. One analysis found that the importance of goal attainment and self-efficacy are the two main factors that contribute to goal commitment [21]. More recent studies [25, 147, 90] showed that individuals with higher self-efficacy are more likely to achieve activity goals and that failing to achieve activity goals reduces individuals’ self-efficacy.

Therefore, activity goals need to be set with care. Past studies [43, 139, 202] and most persuasive technologies [209, 91] either adopted a steady goal of 10,000 steps or allowed self-set goals. To our knowledge, these are the first study to use machine learning to automatically set adaptively personalized step goals and deliver the step goals using a mobile phone technology. The RCT outcomes show that adaptively personalized goals were important in promoting physical activity relative to constant step goals. The adaptive step goals were set to be challenging yet attainable; thus, the average step goals for the intervention group were lower than the average step goals for the control group. As the adaptive step goals were designed to be challenging, the goal achieving percentage for the intervention group was not 100%. Instead, we observed the goal achieving percentage for the intervention group was 30%-40%, which was 15% more compared with the goal achieving percentage for the control group. Being able to achieve more daily step goals can enhance participants' self-efficacy, which further promotes physical activity in the days to follow [21, 23, 9, 62]. The significantly higher (but not too high) rate of achieving step goals and significantly more steps of the participants in the intervention group demonstrate that the BAA algorithm computed adaptively personalized step goals that were capable of being both challenging and manageable for participants, and these goals effectively promoted physical activity.

Nonadherence is another challenge in mobile phone-based lifestyle modification programs. As a result, many past mHealth interventions involve regular in-person counseling sessions besides the mobile intervention to motivate adherence [54, 68, 205]. However, in-person counseling sessions are costly and put a burden on both the participants and the research staff [51, 34, 221]. Our studies were intentionally designed to have only two in-person sessions (each of 15 min) at run-in and at 10 weeks to better simulate the environment of a completely mobile phone-based physical activity intervention. Note that the two in-person sessions in our studies were necessary in-person contacts for the purpose of assessment in the study; they are different from in-person counseling sessions that serve as an essential part of an actual intervention. Despite the absence of coaching sessions, the percentage of frequent users observed over 10 weeks in our study was better than that reported in similar trials [202, 17, 37]. Our results indicate that a mobile phone-based intervention without coaching sessions is still effective in promoting physical activity. In-person contact and coaching sessions are therefore not necessary requirements for effective physical activity interventions, and there is potential to replace those contacts with better-designed physical activity apps. Chapter 6 discusses an analytical approach to predict nonadherence. Incorporating such predictive models into intervention programs has the potential to improve program efficacy.

An additional advantage of these studies is that it only relied on one device for both data collection and intervention delivery. Similar studies either used a pedometer or accelerometer besides the mobile phone or requested regular data inputs from the participants, requiring greater efforts on the participant side, which was shown to be burdensome and could lead to declining use of the app [5]. In our studies, step data were objectively measured by the iPhone, and participants were only requested to carry their mobile phones with them (in their pocket or their purse). No other manual data entry was needed on the participant side. Moreover, the CalFit app is designed in a flexible way that is compatible with other data collection devices, such as wearable step trackers, as long as the step data can be synced with the iPhone.

In addition to objectively measured outcomes, it is of interest to investigate if self-reported survey results differ between this study and full behavioral interventions (with many behavior change components). Barriers to Being Active quiz and the international physical activity questionnaire are popular surveys that have been widely adopted [68, 67, 37, 115, 175, 99]. Researchers found that there exist significant differences in survey responses before and after a full behavioral intervention [68, 30, 172, 67]. However, we failed to observe such difference. We suspect that goal setting alone may not be strong enough to change participants' opinion on self-reported surveys and that other behavior change components are required (e.g., coaching sessions).

These two studies tested one single component of behavior change (i.e., goal setting), and the purpose of this design was to isolate the impact of goal setting from other behavior change components. Beyond goal setting, there are many other components of behavior change that can be beneficial for fitness apps. For instance, customized messages and social interactions have the potential to further improve the efficacy of fitness apps. Our studies are not designed to be a stand-alone intervention but rather to provide evidence on the efficacy of evaluating a single design component to motivate future evaluations on other design components. We believe there is great potential for better-designed fitness apps that can contribute to more effective physical activity intervention.

4.5 Design Implications

There are two major challenges associated with providing fully automated smartphone-based physical activity interventions. The first challenge is supporting users through key behavior change features and effective goal-setting in order to increase their level of physical activity. The second challenge is ensuring sustained maintenance of any increases in physical activity initiated by an intervention. Typical physical activity interventions address these challenges through frequent in-person coaching sessions, which are effective in initiating and maintaining behavior change. Since in-person coaching is expensive, mobile physical activity interventions seek to lower costs by reducing the amount of coaching. As a result, meeting these two challenges is substantially more difficult for fully automated smartphone-based physical activity interventions.

Our studies demonstrates the potential of adopting behavior-change features and using personalization in mobile physical activity interventions to address these challenges. In particular, we found that sending one or two push notifications serves as a useful reminder. Furthermore, users prefer apps that do not require too much time and effort. Features that require regular user input, such as setting personal goals or keeping a diary to record steps/food intake, can create a burden on app adherence. Another main design choice is personalization. The BAA algorithm that sets personalized step goals for users is shown to be effective in increasing daily steps. Providing challenging but yet attainable goals can induce goal-achieving incentives, and giving daily feedback on performance (i.e., reminder push notification on daily goal and congratulating push notification) further reinforces exercise motivation. Conversely, fixed steps goals (10,000 steps/day) with no personalization can be unrealistically high or too easy to achieve and hinders users from progressing to be active.

Future designs of mobile fitness apps should consider personalized interventions, including but not limited to goals, push notifications, and displays. In addition, algorithms for goal-setting should take the complete history of the user as the basis to generate future interventions, particularly when the input and target metrics have high day-to-day variations. Implementing behavior change features, such as self-monitoring and summary feedback on performance, can further motivate physical activity. Overall, the app should be easy to navigate and require minimum manual inputs from users, particularly by using algorithms to automate personalization.

4.6 Limitations

The first limitation of these studies is the relatively small sample size, which only contained UCB adult staff workers (students in Cal Fitness2) with a dominant proportion of females. The results may not generalize to the general public. The relatively high education level of the participants may also limit the generalizability. In addition, the CalFit app was only available on the iOS platform, which could bias results. Second, the daily steps assessment during the run-in period was not able to establish a natural baseline. Therefore, our trials could only determine the relative (to the control) benefit of the intervention, but could not determine the absolute (compared with the natural baseline) benefit of the intervention. Blocked display of step counts with no step goal during the run-in period may provide additional insights to the natural baseline. Third, the iPhone was not able to collect data when it was being turned off or was not with the participant, and it was not able to distinguish the carrying method (purse vs pocket). However, the chance of the above happening was the same for the control and the intervention groups because of randomization; so, these factors do not impact the relative step differences between the 2 groups. Fourth, these studies did not assess the underlying behavior skills (self-efficacy, goal setting, etc) that may impact individual's response to interventions. Finally, the study was conducted for 10 weeks, which is a relatively short time. Studies that span a longer period are needed to evaluate the long-term effect of such personalized step goal-setting intervention delivered via mobile phones.

4.7 Conclusion

Our RCTs indicate that mobile phone-delivered adaptively personalized step goals are promising in promoting physical activity. The interventions both led to a statistically significant more steps in the intervention group compared with the control group over 10 weeks. The higher (but not too high) percentage of goal achievement in the intervention group confirms that the adaptively personalized step goals computed by the BAA algorithm used in this trial are capable of creating challenging yet attainable goals. The significant step difference between the two groups suggests that a mobile phone-based physical activity intervention with reduced in-person sessions is feasible. The results obtained in this study can guide the design of future mobile phone-based physical activity interventions.

4.8 Future Work

In the future, we would like to extend our observations further by studying hypotheses in three directions. Firstly, how do different goal setting sources (i.e, self-set, trainer-set, and machine set) impact the intervention outcome? Secondly, how do different dynamic goal setting algorithms impact the intervention outcome? In particular, it would be beneficial to unveil if the success of our studies is due to the BAA algorithm or due to the fact that step goals are not steady. We would like to compare the BAA algorithm to simpler analytical algorithms, such as, for example, setting the goal to be the 60th percentile of the steps in the past week. Thirdly, we would like to isolate the impact of the various design features (i.e., push notification, history tab, etc.) to provide recommendations on the most effective features to future fitness app designers.

Chapter 5

The Discontinuation Prediction Score (DiPS)

5.1 Background and Related Work

As we mentioned in Chapter 4, low adherence reduces the efficacy of physical activity promotion programs. Adherence is affected by factors like self-efficacy, exercise history, health condition, stress, and cognitive activities [79, 31, 82, 92], and adherence can vary on a day-to-day basis with even usually-adherent individuals having temporary relapses [189]. Inadequate coping skills for high-risk situations are a leading factor in temporary relapse, and such relapse often results in an "abstinence violation effect" that leads to a perceived loss of control and eventually total relapse [130, 182, 128]. Thus, accurate predictions of temporary relapse can be valuable to improve adherence to medical treatments. Such prediction, when combined with automation and mobile technologies, can assist in more precise targeting of just-in-time behavioral interventions (e.g., messages or goals) to reduce relapse.

Much work on predicting adherence has focused on the use of sociodemographics and self-reported questionnaire data [135, 190, 121]. However, newly-available, objective, quantitative data, such as from electronic health records (EHR), wearable devices, or mobile phones can potentially increase the prediction accuracy of adherence. One approach [132] used EHR data to construct a Markov chain model to predict medication adherence, where the model states were frequency of taking medication. Unfortunately, this kind of model is not applicable to personalized interventions where adherence is measured relative to a baseline that varies for each individual. Another approach [16] used mobile data to construct a utility-function model of behavior in weight loss programs. Though the model is personalized to the baseline weight and physical activity of each individual, this model predicts future weight loss and not adherence. Given the potential value of predicting adherence to medical treatments using EHR and mobile data, there is a need for the development and validation of new models that can make such predictions.

Review of Mobile Technologies and Physical Activity

As mentioned in Chapter 4, physical activity promotion programs suffer from exercise relapses and low adherence, which hinders subjects from meeting the suggested guidelines. Despite the potential of leveraging mobile technologies with activity trackers and wearable devices to provide accurate real-time measurements of physical activity and deliver interventions to encourage adherence [7, 38, 64, 85], the capacity of these technologies in automating and personalizing physical activity promotion programs is only beginning to be explored. Compared to typical programs that involve only in-person coaching sessions [129], recent studies have found that mobile-based behavior modification programs with a reduced number of coaching sessions can achieve statistically significant increases in physical activity [26, 27, 36, 59, 67, 68, 88, 94, 100, 152, 167, 66].

Encouraged by this success, a more recent question is whether it is feasible to use mobile technologies to deliver fully-automated or nearly-fully-automated physical activity promotion programs. Though [26, 27, 36, 59, 67, 68, 88, 94, 100, 152, 167] reduced the number of coaching sessions, in-person interactions are still an essential component of those programs. Moreover, these programs lack substantial personalization and automation. For instance, several studies used a small set of pre-programmed textual messages, and the same messages were sent to all participants at the same day/time to encourage certain behaviors. One study [26] sent personalized messages based on self-reported assessments, but did not use objectively measured data for personalization. Additional levels of automation will enable further scaling of these programs to larger populations and improve the efficacy of these programs in a cost-effective way. Accurate prediction of adherence is potentially important to scaling these programs through automation and personalization.

Review of Warning Scoring in Healthcare

Several warning scores have been developed in healthcare for the purpose of predicting adverse medical events at an early stage so that medical interventions can be delivered before significant patient state deterioration occurs. Early Warning Scoring (EWS) is a popular system for bedside patient assessment [116, 137, 151]: It is based on the physiologic assessment of multiple vital signs (e.g., respiration, heart rate, body temperature, etc.) and abnormal observations, which trigger immediate notifications that lead to early interventions to prevent critical events from happening. Validated EWS algorithms are also used to provide guidance for optimizing patient management and guiding resource allocation within healthcare organizations. Current versions of EWS includes Modified Early Warning Score (MEWS), National Early Warning Score (NEWS), and Pediatric Early Warning Score (PEWS) [208, 213].

Despite the wide adoption of EWS systems in critical care medicine, similar types of warning scores have not been developed for or applied to physical activity interventions. Since exercise relapse tends to hinder further participation in physical activity, an early warning scoring system with accurate predictions on exercise relapse can be used to guide the provision of immediate interventions and has the potential to increase adherence.

Study Purpose

In this chapter, we use logistic regression (LR) and support vector machine (SVM) methods to design two versions of a Discontinuation Prediction Score (DiPS), which uses each individual's past data (e.g., physical activity duration, physical activity intensity and goal achievement) to assign a numeric value that quantifies their likelihood of discontinuing physical activity in the upcoming week. The potential utility of DiPS to provide guidance for provision of just-in-time interventions for individuals who are more likely to have an exercise relapse is demonstrated through a simulation in which we compare the cost-effectiveness of different schemes to allocate financial incentives that encourage selected individuals to increase their physical activity.

This chapter is organized as follows: We first describe the dataset of the Mobile Phone-Based Physical Activity Education program (mPED), a randomized controlled trial with 210 women. Our first step in designing DiPS is to perform *feature engineering*, which is an important step in machine learning whereby the raw data for each individual is converted into a set of summary statistics that characterize the data for each individual. Next, we define two versions of DiPS using logistic regression and SVM, and we quantify the prediction accuracy of DiPS using an out-of-sample evaluation methodology. Lastly, we discuss how DiPS can be integrated into physical activity promotion programs and present a simulation to demonstrate the potential of using DiPS-based interventions to increase adherence in a cost-effective way.

5.2 Methods

Data Description

Mobile-based physical activity promotion programs collect real time objectively measured health data through devices such as digital accelerometers and smartphones. Digital accelerometers are capable of collecting steps and metabolic equivalents (METs) data at the minute-level (i.e., steps or METs per minute). Most of these programs also implement a goal setting feature and monitor goal achievement through either a smartphone application or an online platform. Summarizing statistics (known as *features* in machine learning) can be extracted from this data such that these features accurately predict future exercise relapse for each participant. In this chapter, we use such data from a single mobile-based physical activity promotion program to design and evaluate a DiPS score. Note that our approach can be applied to similar datasets with objectively measured physical activity data.

The mPED Dataset

In this chaper, we used data from the Mobile Phone-Based Physical Activity Education (mPED) program with 210 community dwelling female adults, age 25 to 69 years. The study protocol was approved by the University of California, San Francisco Committee on Human Research (CHR) and the mPED Data and Safty Monitoring Board. A detailed description of the study was described in [69]. The mPED was a randomized controlled trial (RCT) with 3 groups and participants were randomized into one of the three groups with a 1:1:1 ratio. (1) the CONTROL group received accelerometer only, with no intervention provided; (2) the PLUS intervention group received a 3-month (12 weeks) mobile phone and accelerometer based physical activity intervention and a 6-month (24 weeks) mobile phone diary maintenance intervention; and (3) the REGULAR intervention group received a 3-month (12 weeks) mobile phone and accelerometer based physical activity intervention, but kept only accelerometer during the 6-month (24 weeks) maintenance period. Prior to the intervention, all participants participated in a 3-week run-in period to collect baseline average daily steps. To avoid providing any feedback and to collect clean baseline activity data, neither the step counts nor METs were displayed. The CONTROL group did not receive any daily step goals during the 9-month (36 weeks) period. In contrast, the PLUS and REGULAR groups received an identical 3-month (12 weeks) physical activity intervention, and their weekly daily step goals were set to increase at a 20% rate from the participant's average baseline daily steps. Once their daily step goals reached 10,000 steps, they were asked to maintain at least 10,000 steps per day, seven days a week during the remaining study period.

In the mPED trial, physical activity was measured using a triaxial accelerometer (HJA-350IT, Active style Pro, Omron Healthcare Co., Ltd.). It collects the following types of physical activity data:

METs data: The mean intensity value of a 1-min epoch is calculated as the average value of six 10-s epochs. Based on the METs recordings, physical activity is automatically classified as no measurement, lifestyle activity and walking activity. Moderate to vigorous intensity physical activity (MVPA) is METs ≥ 3.
Steps data: The accelerometer provides information on the steps value of a 1-hour epoch and daily steps.

Features

The performance of machine learning techniques largely depends upon the set of features used. We extracted a set of interpretable features from the objectively measured physical activity data. In detail, for a participant in a particular week (for example, week t), we defined the following set of features:

1. Average daily steps: average of daily steps from the first day of the run-in period to the last day of week $t-1$.

2. Initial average daily steps: average of daily steps in the run-in period.

3. Last week average daily steps: average of daily steps in week $t-1$.

4. Average goal achieving percentage: percentage of step goals achieved from the first day of the run-in period to the last day of week $t-1$.

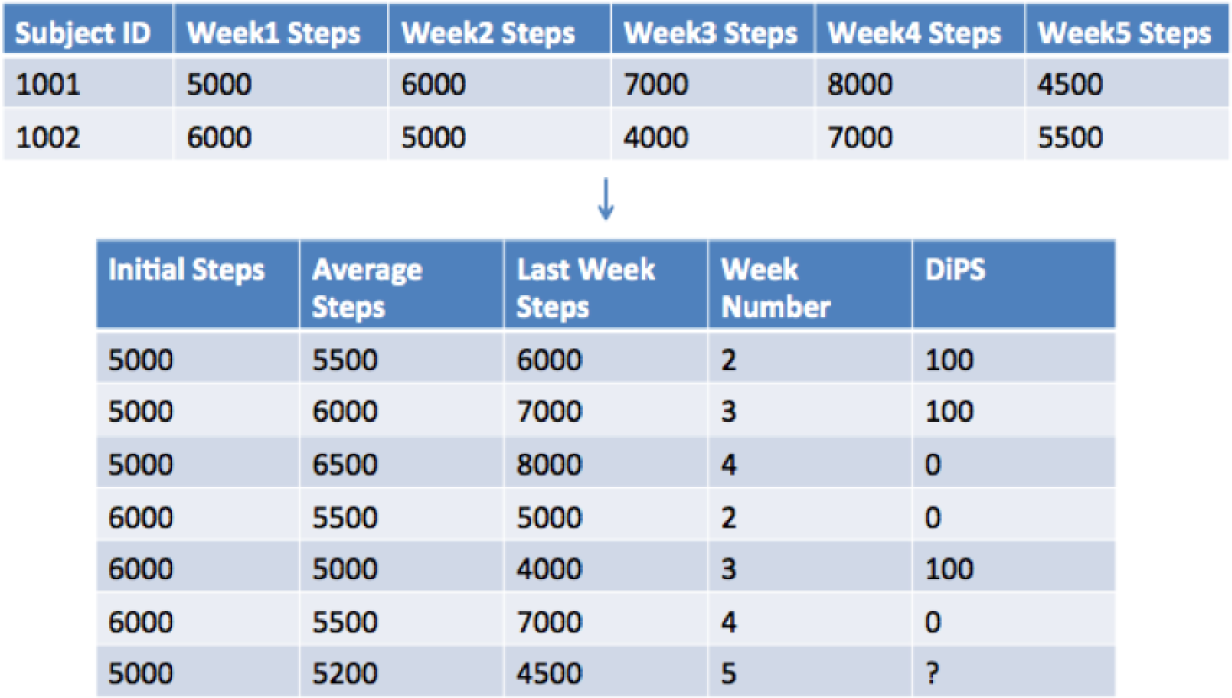
5. Last week goal achieving percentage: percentage of step goals achieved in week $t-1$.

6. Average MVPA minutes in the morning: average number of minutes with METs ≥ 3 in the morning (3:00 - 9:59) from the first day of the run-in period to the last day of week $t-1$.
7. Initial MVPA minutes in the morning: average number of minutes with METs ≥ 3 in the morning (3:00 - 9:59) in the run-in period.
8. Last week MVPA minutes in the morning: average number of minutes with METs ≥ 3 in the morning (3:00 - 9:59) in week $t-1$.
9. Average MVPA minutes in the afternoon: average number of minutes with METs ≥ 3 in the afternoon (10:00 - 14:59) from the first day of the run-in period to the last day of week $t-1$.
10. Initial MVPA minutes in the afternoon: the average number of minutes with METs ≥ 3 in the afternoon (10:00 - 14:59) in the run-in period.
11. Last week MVPA minutes in the afternoon: average number of minutes with METs ≥ 3 in the afternoon (10:00 - 14:59) in week $t-1$.
12. Average MVPA minutes in the evening: average number of minutes with METs ≥ 3 in the evening (15:00 - 3:00) from the first day of the run-in period to the last day of week $t-1$.
13. Initial MVPA minutes in the evening: average number of minutes with METs ≥ 3 in the evening (15:00 - 3:00) in the run-in period.
14. Last week MVPA minutes in the evening: average number of minutes with METs ≥ 3 in the evening (15:00 - 3:00) in week $t-1$.
15. Average MVPA intensity: average METs readings for METs ≥ 3 from the first day of the run-in period to the last day of week $t-1$.
16. Initial MVPA intensity: average METs readings for METs ≥ 3 in the run-in period.
17. Last week MVPA intensity: the average METs readings for METs ≥ 3 in week $t-1$.
18. Week number t: the number of weeks in the study.

Daily steps reflect the participant's overall daily physical activity. Goal-achieving percentage demonstrates the participant's response to step goals. MVPA in different time in day expresses the preferred time in day of performing MVPA, and MVPA intensity is coarsely indicative of the type of physical activity performed. We separately a day into three intervals: morning (3:00 - 9:59), afternoon (10:00 - 14:59), and evening (15:00 - 3:00) because prior clustering analysis on this dataset [70] identified three clusters of individuals who tend to do physical activity in the morning (3:00 - 9:59), afternoon (10:00 - 14:59), and evening (15:00 - 3:00), respectively.

Restated, our defined set of features includes daily steps, goal-achieving percentage, MVPA in the morning, MVPA in the afternoon, MVPA in the evening and MVPA intensity: The complete set of features includes all 18 features listed above, where we included the individual set of features for the run-in period, in the last week and over the entire study period. The set of features for the run-in period are included to account for the initial differences between participants. The features on last week behavior capture the immediate past performance, which has been shown to be most predictive for the immediate future. The features on average behavior demonstrate the overall performance of the participant so far during the study. The week number is included to model the changes in behavior over time.

Pre-processing



Subject ID	Week1 Steps	Week2 Steps	Week3 Steps	Week4 Steps	Week5 Steps
1001	5000	6000	7000	8000	4500
1002	6000	5000	4000	7000	5500

Initial Steps	Average Steps	Last Week Steps	Week Number	DiPS
5000	5500	6000	2	100
5000	6000	7000	3	100
5000	6500	8000	4	0
6000	5500	5000	2	0
6000	5000	4000	3	100
6000	5500	7000	4	0
5000	5200	4500	5	?

Figure 5.1: Simplified example of training data augmentation. The first table shows the raw physical activity data for two different participants 1001 and 1002 before augmentation and the second table shows the resulting data after augmentation, where the first six rows are the training data and the last row is the testing data.

Recall that the mPED dataset contains 210 participants. In our use of machine learning methods, we first assume that the relationship between week i and $i+1$ is independent of the relationship between week j and $j+1$ for $i \neq j$ given the week number. Under this assumption, we can augment the training data to include features for each participant for each observed week. For instance, assume we are at week n of the study and we want to train the model, then instead of using a single observation for each participant using data in week $n-1$ as the response variable, our new approach creates a set of observations for the participant, where each observation uses data in week

$3, 4, .., n-1$ as the response variable and the corresponding features are from weeks prior to that week. For example, suppose we are at the end of week 5 of the study and would like to generate observations for participant 1001. Then the augmented training data contains 3 observations for participant 1001, i.e., the complete set of features for week 2, 3 and 4; and the response variable is the observed DiPS of participant 1001 in week 3, 4 and 5 respectively, i.e., 0 if the participant's average step in that week is below the participant's average step in the run-in period, 1 otherwise. Figure 5.1 below illustrates a simplified example of this training data augmentation procedure.

Analytic Models

In this section, we first define the Discontinuation Prediction Score (DiPS) in the context of a clinical trial. Then we move on to introduce the statistical models (i.e., logistic regression and SVM) used to develop this score.

DiPS Definition

DiPS aims to predict the probability of having exercise relapse within a pre-specified time interval for a particular participant. In this dissertation, we use [0,1] as the interval for DiPS since the model output is a probability. We further define that a participant is having an exercise relapse in a given week if the average steps in that week is lower than the average steps in the run-in period. The reason we use the run-in period average is that the mPED trial was designed so that true baseline steps data is collected from each participant during the run-in period. Since the aim of these programs is to increase participants' daily steps, comparison with the run-in average serves as a useful signal reflecting the progress of the participant.

Recall that the 2008 National Physical Activity Guidelines use a week as the unit to measure activity time for different physical activity intensity [203, 219]. We adopt a similar approach to use a week (rather than a day) as an assessment unit. Furthermore, this granularity can mitigate the impact of large day-to-day fluctuations of daily steps. Note that for a particular week in the past (thus with known data), a participant has a DiPS of 0 if his/her weekly average step is lower than his/her run-in average steps and a DiPS of 1 if his/her weekly average steps is higher than his/her run-in average steps. Therefore in the training phase, DiPS can be regarded as a binary variable (since it is either 0 or 1). But in the prediction phase, DiPS is a continuous variable in the range of [0,1], indicating the likelihood of achieving an above run-in step in the following week.

Logistic Regression

Since our response variable (exercise relapse or not) is binary, our problem is essentially a classification problem. Therefore, logistic regression is a favorable statistical approach since it has high interpretability and works well in practice. Logistic regression models the log-odds using an affine function. Let X be the feature matrix and Y be the response variable (with 1 indicating no exercise relapse and 0 indicating exercise relapse). For a generalized linear model parameterized by θ:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^t x}}$$

the probability that y is 0 or 1 can be expressed as

$$Pr(y|x;\theta) = h_\theta(x)^y(1-h_\theta(x))^{(1-y)}.$$

The maximum likelihood estimate (MLE) of the model parameters can be computed by standard algorithms. The output of a prediction of the logistic regression model for a given set of estimated parameters is a numerical value indicating the likelihood of exercise relapse. In addition to prediction, a fitted logistic regression model can be valuable for interpretation, because we can identify the importance of features by considering their corresponding coefficients. We used the `glm` function with the binomial family in R [163] for implementation.

SVM

Support Vector Machine (SVM) is a classification method that uses separating hyperplanes. SVM selects the hyperplane that gives the largest maximum distance to the training examples. Formally, if we assume we want to classify data points into two classes: -1 for exercise relapse and 1 for no exercise relapse (note this definition of class is different from logistic regression), then we can define a hyperplane:

$$f(x) = \beta_0 + \beta^T x.$$

Notice that we can scale β_0 and β so that $|\beta_0 + \beta^T x| = 1$. Then if we define the training data that have minimum distance to the hyperplane as support vectors, we can write out the distance to the support vectors as:

$$\text{distance}_{\text{ support vectors}} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}.$$

With some manipulations, we see that finding the optimal separating hyperplane is the same as solving the following optimization problem:

$$\min_{\beta,\beta_0}\{||\beta||^2 \mid y_i(\beta^T x_i + \beta_0) \geq 1 \ \forall i\}$$

Note that the above problem is infeasible if the data points are not linearly separable. Therefore, we relax the constraints and adopt the soft-margin SVM approach. In detail, we solve the following optimization problem:

$$\min_{\beta,\beta_0}\{||\beta||^2 + \lambda \textstyle\sum_i \xi_i \mid y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i \ , \xi_i \geq 0 \ \forall i\}$$

where λ is a constant that controls the trade-off between fit-to-data and the size of the margin. This problem can be solved easily by standard algorithms. SVM is a popular classification algorithm and [47] provides more theoretical details about this method. We used the `svm` function in the e1071 package in R [163] for implementation.

Train and Test Data

We train the models using the preprocessed data collected in the first 15 weeks of the study. Then we use the trained model to predict exercise relapse for weeks 16-30. We compare our logistic regression (LR) model with 18 extracted features to the SVM model to demonstrate prediction accuracy.

5.3 Results

Model Evaluation

Table 5.1: Test AUC for predicting weeks 16-30 on all subjects using the fitted DiPS model.

Week	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
LR	0.932	0.861	0.900	0.893	0.892	0.925	0.884	0.916
SVM	0.905	0.866	0.886	0.879	0.882	0.900	0.862	0.894
Week	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**Mean**
LR	0.876	0.920	0.912	0.900	0.900	0.915	0.899	0.902
SVM	0.825	0.900	0.904	0.885	0.905	0.889	0.899	0.886

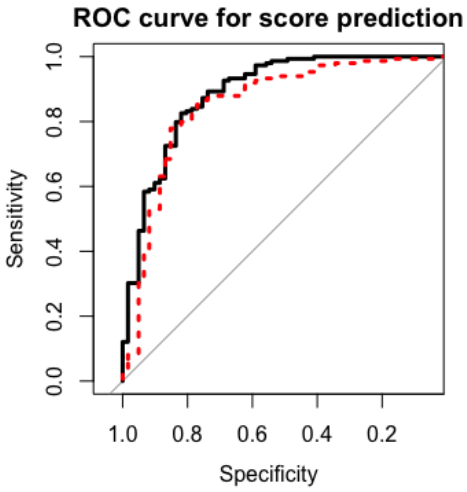


Figure 5.2: Receiver Operating Characteristics (ROC) curve of the predictions for week 20 using Augmented Logistic Regression and SVM. Black solid line is Logistic Regression and red dash line is SVM.

We evaluate the performance of the models by comparing their Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC), where an AUC close to 1 indicates better performance of the classification task. Table 5.1 shows the AUC of the predictions for weeks 16-30 using the model trained by data from the first 15 weeks (including the 3 weeks of run-in and 12 weeks of intervention).

Results show that the LR model has a higher average test AUC of 0.9016. The SVM model has a slightly lower average test AUC of 0.8855. In addition, we observe that the LR model is more robust than the SVM model with lower variance of AUC when predicting for different weeks. Overall, the high accuracy of both models indicates the robustness of the selected features in predicting DiPS. We show an example of the AUC curves of the two models in week 20 (the AUC curves of other weeks are similar) in Figure 5.2 and observe that the optimal thresholds for the two models both have high accuracy (>80%) and high specificity (>80%). Regularized versions (i.e., lasso, ridge, and elastic net) of LR and SVM was also tried, but we did not observe significant improvement in model performance on the test data, primarily due to the small number of features.

Table 5.2: Confusion matrix of the observed and predicted class for week 20 using the augmented Logistic Regression approach with a threshold of 0.5.

		True Class		
		Relapse	Not Relapse	Total
Predicted class	Relapse	53	49	102
	Not Relapse	8	100	108
	Total	61	149	210

Table 5.2 displays the confusion matrix of the observed and predicted class for week 20 using the LR model with a threshold of 0.5 (this is an example threshold, and not necessarily the optimal). DiPS obtains an accuracy of 85% and a specificity of 67%. The prediction accuracies in other weeks are comparable to week 20, confirming the overall robustness of the model. Note the above analysis was conducted on all participants, disregarding the randomization group. Therefore, we conducted additional analysis to evaluate whether the performance of the algorithm is consistent with different randomization groups. We conducted the ROC and AUC analysis for each of the individual groups on their test data from weeks 16-30 and present the results in Table 5.3. The outcome indicates that in the beginning weeks of the maintenance period, the test AUC of the CONTROL Group is lower than that of the other two groups. But toward later weeks, the test AUC of the CONTROL group increases. The PLUS group has the highest test AUC for most of the weeks during the maintenance period. The test AUC on the REGULAR group is between those of the CONTROL group and the PLUS group.

Table 5.3: Test AUC for predicting weeks 16-30 on each intervention group using the fitted DiPS model.

Week	Control Group	Regular Group	Plus Group
16	0.619	0.878	0.968
17	0.804	0.888	0.986
18	0.810	0.908	0.945
19	0.784	0.891	0.983
20	0.867	0.911	0.889
21	0.904	0.916	0.945
22	0.894	0.850	0.954
23	0.871	0.902	0.976
24	0.846	0.809	0.963
25	0.908	0.914	0.943
26	0.875	0.888	0.912
27	0.907	0.882	0.929
28	0.882	0.857	0.947
29	0.848	0.864	0.957
30	0.893	0.831	0.959

Model Interpretation

An advantage of using LR and SVM as the machine learning methods are their interpretability. For a fitted LR and SVM model, the importance of each feature can be assessed using the coefficients. Table 5.4 shows the feature importance for the fitted LR model using data collected during the first 15 weeks. Week number is highly significant, and the negative coefficients indicate that as the study progressed, participants were more likely to have exercise relapses. Initial steps, mean steps and last week steps are all highly significant in predicting exercise relapse. In addition to steps, physical activity intensity turns out to be a predicative feature. The positive coefficient of last week intensity indicates that if a participant was doing higher intensity physical activity, he/she was less likely to have an exercise relapse. Overall, DiPS is largely affected by week number, daily steps and overall physical activity intensity, but less dependent on other features, like preferred MVPA time in day.

Simulation

Next, we show how DiPS could be used to personalize and adapt interventions. In this section, we use simulation to compare a DiPS-based intervention to a random intervention and a steps-based

Table 5.4: Feature importance for the fitted augmented Logistic Regression model for week 20.

Feature	Coefficient	P-value
Intercept	1.378	0.03
Week number	-0.122	<0.001
Initial average daily steps	-0.001	<0.001
Average daily steps	0.0008	<0.001
Last week average daily steps	0.0004	<0.001
Initial MVPR minutes morning	-0.029	0.138
Initial MVPR minutes afternoon	0.027	0.04
Initial MVPR minutes evening	0.015	0.242
Average MVPR minutes morning	0.067	0.02
Average MVPR minutes afternoon	-0.037	0.103
Average MVPR minutes evening	0.005	0.805
Last week MVPR minutes morning	-0.032	0.058
Last week MVPR minutes afternoon	0.001	0.920
Last week MVPR minutes evening	-0.003	0.797
Initial MVPA intensity	-0.158	0.479
Average MVPA intensity	-0.354	0.280
Last week MVPA intensity	0.515	<0.001
Average goal achieving percentage	0.161	0.849
Last week goal achieving percentage	0.294	0.340

intervention, by assuming a simple dynamic step model with financial incentives. Our model assumes that for each participant, his/her steps for day $n+1$ is determined as follows:

$$steps_{n+1} = \alpha \cdot steps_n + C + \varepsilon$$

where C is a constant. Here, ε is a random variable that captures day-to-day fluctuations in physical activity, and we use the model that ε is a uniform distribution with range $-E$ to E for a constant E. We further assume that $steps_0$ is the average steps during the run-in period. We used the mPED data to fit the above model for each participant so that the i-th participant has model parameters $\{\alpha_i, C_i, E_i\}$. We used these parameters when conducting our simulation.

In the simulation, we compare the number of adherent participants after a 3-month intervention period, where adherence is as defined in the methodology section: a participant is adherent if his/her steps in the latest week is greater or equal than his/her steps in the run-in period. We consider the

following three policies for an intervention to increase adherence:
1. Random intervention: the probability of giving intervention for a given day and a given participant is p^*.
2. Step based intervention: give intervention if observed daily step is below some threshold $step^*$.
3. DiPS based intervention: give intervention if predicted DiPS score is below some threshold $DiPS^*$.
For simplicity, we assume the intervention is a financial incentive (i.e., some dollar value for each intervention) and that giving the intervention will lead to an increase of 500 steps for that day. (The actual responsiveness to a fixed financial incentive will vary for each participant, and this sensitivity can be adaptively estimated for each participant using machine learning [16, 225]. We did not estimate this sensitivity for the simulation because financial incentives were not used in the mPED trial, and so the data needed to be able to estimate sensitivity is not available.) Therefore, after the parameter estimation phase, we use the resulting parameters to simulate data for a new study using one of the three intervention policies. Formally, we have:

$$steps_{n+1} = \alpha \cdot steps_n + C + \varepsilon + 500 \cdot u_n$$

where u_n follows one of the three intevention policies described above:
1. u_n has a Bernoulli distribution with success probability p^*
2. $u_n = \begin{cases} 1, & \text{if } step_{n-1} < step^* \\ 0, & \text{otherwise} \end{cases}$
3. $u_n = \begin{cases} 1, & \text{if } DiPS_n < DiPS^* \text{ and n mod 7 is 1 (i.e., first day of a week)} \\ 0, & \text{otherwise} \end{cases}$

For this simulation, DiPS is computed using the steps data and the week number since we do not observe the other features. Also, the first two policies are assessed daily, while the DiPS policy is assessed weekly and its intervention occurs only on the first day of the week when the predicted DiPS is smaller than the threshold. We select a sequence of values for $p^*, step^*, DiPS^*$ so that total spending in the three policies is comparable. Figure 5.3 shows the percentage of adherent participants versus number of interventions per participant for the three policies.

5.4 Discussion and Implications

Accuracy and Interpretation of DiPS

The Discontinuation Prediction Score (DiPS) uses each individual's past data to assign a numeric value that quantifies their likelihood of discontinuing physical activity in the upcoming week, and it based on a Logistic Regression model or a SVM model with interpretable features. The results of our model validation suggest that DiPS has a test accuracy around 80% and makes robust predictions across different weeks. This suggests that DiPS is a potentially effective score that clinicians could use to tailor and adapt physical activity interventions to prevent exercise relapses.

The most predictive features in DiPS are: steps data, including initial average daily steps, average daily steps, last week steps, and physical activity intensity data. In contrast, preferred

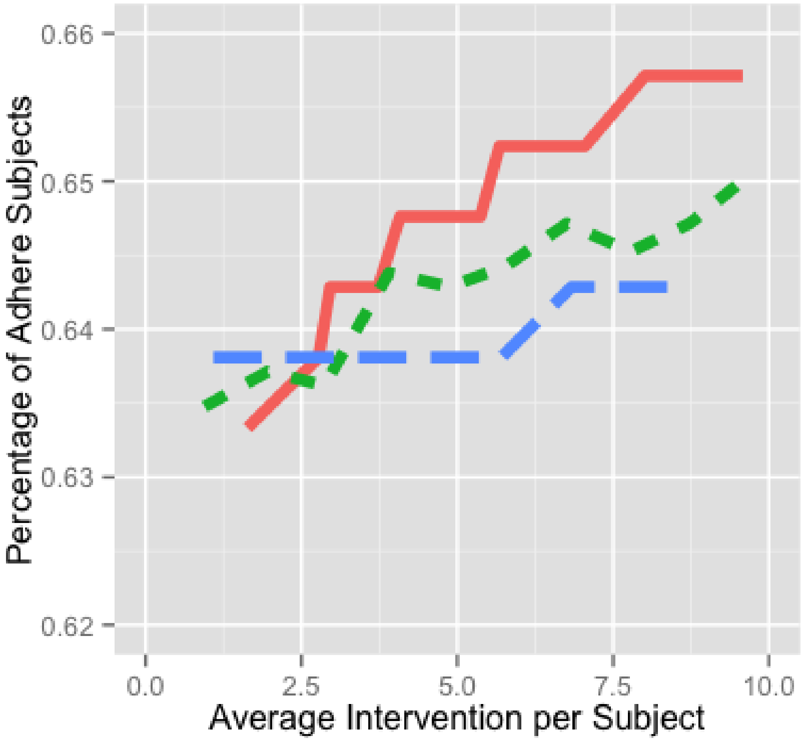


Figure 5.3: Simulation outcome of number of adhere participants after a 3-month trial with increasing spending under the three intervention policies. Red solid line is DiPS-based intervention; Green shorter dash line is random intervention; Blue dash line is step-based intervention.

MVPA time in day was not significant. Also, the coefficient of initial average daily steps is negative, indicating that participants with a higher physical activity level during the run-in period tended to have exercise relapse more often on average. This is intuitive since we define exercise relapse to be a comparison between current week steps and initial steps, thus higher initial steps means a more difficult baseline to beat. Furthermore, last week daily steps and average daily steps have positive coefficients, indicating that participants who were more active in the last week and over the entire study period were less likely to have exercise relapses.

Efficient Resource Allocation using DiPS

The ability of DiPS to provide real-time feedback for generating just-in-time interventions for individuals likely to have an exercise relapse was demonstrated through a simulation that compared the cost-effectiveness of different policies to allocate financial incentives to encourage selected individuals to increase their physical activity. The results of our simulation are summarized in Figure 5.3.

The steps policy (blue longer dash line) leads to the largest percentage of adherent participants when on average less than 2.6 interventions were delivered to each participant. As we increase the number of interventions, the DiPS policy (red solid line) leads to the largest percentage of adhere participants. The random policy (green short dash line) and steps policy (blue longer dash line) have lower performance, and the random policy appears to perform slightly better than the step-based policy. However, the simple dynamic step model used in our simulation makes some assumptions that may not accurately capture participants' sensitivity to financial incentives. Empirical study is warranted to confirm these simulation findings.

Incorporation of DiPS into Physical Activity Promotion Programs

DiPS can be easily incorporated into mobile technology based physical activity promotion programs that collect objectively measured outcome data. With the development of motion sensors and wearable devices, DiPS is able to make predictions using real-time data and automatically trigger immediate interventions. Below, we discuss three potential intervention strategies where DiPS can be useful.

The first intervention strategy is just-in-time messages. Intervention messages can be delivered through push notifications for app-based programs or through text messages for more traditional mobile-based interventions. Such interventions can be triggered automatically when a low DiPS is predicted. The content of such interventions can provide an interactive dialog to identify the reason for relapse and provide personalized suggestions accordingly.

DiPS can also provide automated goal adjustments where future step goals are reduced for those who are experiencing exercise relapse. Goal setting is an important consideration for physical activity promotion programs and past studies have identified that personalized goal setting is more effective than standard goal setting [143, 107, 3, 2]. Setting unrealistically high step goals can discourage relapsed participants, reduce their self-efficacy, and prevent them from doing physical activity. In contrast, realistic goal setting based on their steps during exercise relapses can increase their motivation and encourage them to re-engage in physical activity.

Lastly, DiPS can be used as a signal for in-person coaching session scheduling. Rather than scheduling in-person coaching sessions using a pre-defined schedule (i.e., once every month), session times can be adjusted using DiPS. For instance, when consecutively low DiPS is observed and other intervention methods fail to work, then researchers can schedule an in-person coaching session to create a bigger incentive to motivate the participant. This personalized approach for in-person coaching session scheduling have the potential to improve the efficacy of coaching sessions and result in cost-effective physical activity promotion programs.

5.5 Conclusion

Early prediction of individuals who are likely to relapse can significantly improve adherence to physical activity interventions. The Discontinuation Prediction Score (DiPS), a machine learning technique which uses logistic regression or SVM on objectively measured step and goal data, is able to accurately predict exercise relapse with a sensitivity of 85% and a specificity of 67%.

Simulation results confirm that using DiPS as a score to allocate resources leads to more cost-effective intervention for increasing adherence. Multiple ways of incorporating the DiPS score in physical activity interventions are discussed and empirical study is warranted to confirm the impact. We also believe that our methodology for designing the DiPS score will be valuable for engineering other similar scores to predict adherence to other medical treatments like medication, cancer screening, and dental cleanings.

5.6 Future Work

In the future, we would like to apply DiPS to more physical activity intervention datasets and explore if prediction accuracies are consistent. Furthermore, we wish to investigate if a trained DiPS model on one dataset can provide high prediction accuracy on another similar dataset. If so, this will indicate that study subjects have similar behavioral processes regarding adherence in physical activity intervention programs.

Secondly, we would like to integrate DiPS in healthcare intervention programs (such as the CalFit app). Empirical studies are needed to assess the performance of DiPS in efficiently allocate resources and whether it could achieve the best adherence outcome at the end of the trial. In addition, we want to test the performance of the two different versions of DiPS (LR based vs. SVM based) to advise future adoption.

Chapter 6
Conclusion

This dissertation presents human behavior modeling algorithms and incentive mechanism designs for online platforms. Empirical results using these approaches indicate the efficacy of these methods in enhancing the performance of the underlying platforms.

Chapter 2 and Chapter 3 discuss two novel collective intelligence platforms, i.e., M-CAFE and DebateCAFE, which fills the need for ongoing course evaluation and online deliberation respectively. Collaborative filtering algorithms were adopted to overcome the scale issue resulting from growing population on the platforms. An incentive mechanism inspired by the agent-principal model was introduced to mitigate selective exposure in DebateCAFE.

Chapter 4 and Chapter 5 focus on healthcare intervention platforms, in particular, human behavior modeling in mobile-based physical activity interventions applications. The behavioral analytics algorithm (BAA), based on reinforcement learning, was incorporated in an iOS app CalFit to set personalized, adaptive step goals for individuals. Results from randomized controlled trials (RCT) confirmed the efficacy of using complicated machine learning approaches for human behavior modeling to promote physical activity. Moreover, the Discontinuation Prediction Score (DiPS), a machine learning technique which uses logistic regression or SVM on objectively measured step and goal data, was developed and was validated to accurately predict exercise relapse. Various ways of incorporating DiPS in physical activity interventions to improve adherence were discussed.

The technical implications and future work of these approaches were presented at the end of each chapter. Beyond the methodologies introduced in this dissertation, a bigger challenge is to apply those methods to commercialized platforms to truly benefit a larger population. Bigger scale empirical studies are needed to thoroughly evaluate the algorithms. Though the problems this dissertation attempts to solve are specific to the specific domains, I hope this piece of work can be used as an inspiration on what potential methods (may it be machine learning-based models or incentive mechanism designs) can be used to build better online platforms to suit their functional goals. Building online platforms is an iterating process and many more features should be considered. I hope this dissertation will also help drive investigations in other features (for instance, personalized notification for healthcare interventions) that can help build more effective platforms.

Appendix A

Appendix

Table A.1: Self-reported physical activity scores and barriers to being active of the two groups pre and post the Cal Fitness study 1.

Variable	Control (M±SD)	Intervention (M±SD)	p-value	Overall p-value
IPAQ Questionnaire				
Physical activity past week				0.69
Baseline	5.0±2.3	5.1±2.6	0.92	
10-week	5.8±2.7	5.3±2.6	0.45	
Regular exercise routine				0.17
Baseline	5.4±2.9	4.7±3.5	0.38	
10-week	5.9±3.3	4.5±3.4	0.10	
Satisfied with health status				0.97
Baseline	6.8±2.0	7.0±1.8	0.65	
10-week	6.9±2.2	6.6±2.4	0.63	
Days doing vigorous PA				0.88
Baseline	1.7±1.2	2.2±1.6	0.16	
10-week	2.0±1.8	1.7±1.7	0.48	
Minutes doing vigorous PA				0.52
Baseline	60.2±61.3	49.8±32.9	0.43	
10-week	43.8±48.8	37.2±36.4	0.55	
Days doing moderate PA				0.23
Baseline	3.2±2.1	3.1±2.1	0.75	
10-week	3.0±2.4	2.2±2.2	0.16	
Minutes doing moderate PA				0.18
Baseline	44.8±42.6	58.0±61.5	0.35	
10-week	42.9±39.4	58.2±81.9	0.36	
Days walking				0.52
Baseline	5.7±1.6	5.7±1.9	0.98	

10-week	6.1±1.33	5.7±2.0	0.28	
Minutes walking				0.41
Baseline	43.1±32.1	42.9±34.9	0.99	
10-week	54.6±49.4	57.1±61.2	0.86	
Hours per day sitting				0.32
Baseline	8.8±4.7	9.4±5.6	0.66	
10-week	7.7±2.5	9.4±9.5	0.34	
Barriers to Being Active				0.41
Total score				
Baseline	19.1±8.7	20.9±10.3	0.45	
10-week	18.1±9.5	20.3±10.6	0.39	
Lack of time				0.51
Baseline	4.1±2.2	4.3±2.7	0.71	
10-week	4.2±2.5	4.0±2.8	0.83	
Social influence				0.55
Baseline	2.4±1.6	2.9±2.2	0.34	
10-week	2.2±1.6	2.6±2.0	0.39	
Lack of energy				0.24
Baseline	4.5±2.1	4.7±2.4	0.73	
10-week	4.3±2.3	4.4±2.4	0.91	
Lack of willpower				0.99
Baseline	4.3±2.6	4.4±2.7	0.93	
10-week	3.7±2.3	4.8±2.9	0.11	
Fear of injury				0.74
Baseline	0.7±1.2	0.6±0.9	0.74	
10-week	0.7±1.3	0.6±1.1	0.85	
Lack of skill				0.40
Baseline	1.3±1.8	1.2±1.6	0.96	
10-week	0.9±1.2	1.1±1.4	0.49	
Lack of resources				0.35
Baseline	1.9±1.7	2.8±1.6	0.03	
10-week	2.2±1.8	2.8±1.6	0.17	

Bibliography

[1] Panagiotis Adamopoulos. “What makes a great MOOC? An interdisciplinary analysis of student retention in online courses”. In: (2013).

[2] Marc A Adams et al. “Adaptive goal setting and financial incentives: a 2× 2 factorial randomized controlled trial to increase adults? physical activity”. In: *BMC public health* 17.1 (2017), p. 286.

[3] Marc A Adams et al. “An adaptive physical activity intervention for overweight adults: a randomized controlled trial”. In: *PloS one* 8.12 (2013), e82901.

[4] Meredith JD Adams and Paul D Umbach. “Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments”. In: *Research in Higher Education* 53.5 (2012), pp. 576–591.

[5] Aino Ahtinen et al. “User experiences of mobile wellness applications in health promotion: User study of Wellness Diary, Mobile Coach and SelfRelax”. In: *Pervasive Computing Technologies for Healthcare, 2009. PervasiveHealth 2009. 3rd International Conference on*. IEEE. 2009, pp. 1–8.

[6] Steffen Albrecht. “Whose voice is heard in online deliberation?: A study of participation and representation in political debates on the internet”. In: *Information, Community and Society* 9.1 (2006), pp. 62–82.

[7] Tim Althoff et al. “Large-scale physical activity data reveal worldwide activity inequality”. In: *Nature* 547.7663 (2017), pp. 336–339.

[8] Fatema El-Amrawy and Mohamed Ismail Nounou. “Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?” In: *Healthcare informatics research* 21.4 (2015), pp. 315–320.

[9] Eileen S Anderson et al. “Social-cognitive determinants of physical activity: the influence of social support, self-efficacy, outcome expectations, and self-regulation among participants in a church-based health promotion study.” In: *Health Psychology* 25.4 (2006), p. 510.

[10] Thomas A Angelo and K Patricia Cross. “Minute paper”. In: *Classroom assessment techniques: A handbook for college teachers* (1993), pp. 148–153.

[11] James J Annesi. “Goal-setting protocol in adherence to exercise by Italian adults”. In: *Perceptual and motor skills* 94.2 (2002), pp. 453–458.

[12] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1.* 2016. URL: `http://docs.mosek.com/7.1/toolbox/index.html`.

[13] Matt Apuzzo, Joseph Goldstein, and Eric Lichtblau. "Apple's Line in the Sand Was Over a Year in the Making". In: *The New York Times* (2016).

[14] Paul Araiza et al. "Efficacy of a pedometer-based physical activity program on parameters of diabetes control in type 2 diabetes mellitus". In: *Metabolism-Clinical and Experimental* 55.10 (2006), pp. 1382–1387.

[15] Anil Aswani, Zuo-Jun Max Shen, and Auyon Siddiq. "Inverse optimization with noisy data". In: (2015). URL: `http://arxiv.org/abs/1507.03266`.

[16] Anil Aswani et al. "Behavioral modeling in weight loss interventions". In: *SSRN 2838443* (2016). DOI: `http://dx.doi.org/10.2139/ssrn.2838443`.

[17] Audie A Atienza et al. "Using hand-held computer technologies to improve dietary intake". In: *American journal of preventive medicine* 34.6 (2008), pp. 514–518.

[18] Kristen MJ Azar et al. "Mobile applications for weight management: theory-based content analysis". In: *American journal of preventive medicine* 45.5 (2013), pp. 583–589.

[19] Yang Bai et al. "Comparison of Consumer and Research Monitors under Semistructured Settings." In: *Medicine and science in sports and exercise* 48.1 (2016), pp. 151–158.

[20] Peter Baldwin and David Price. "Debate-Graph Details". In: (2016).

[21] Albert Bandura. "Self-efficacy: toward a unifying theory of behavioral change." In: *Psychological review* 84.2 (1977), p. 191.

[22] Albert Bandura. "Social cognitive theory of moral thought and action". In: *Handbook of moral behavior and development* 1 (1991), pp. 45–103.

[23] Albert Bandura and Daniel Cervone. "Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems." In: *Journal of personality and social psychology* 45.5 (1983), p. 1017.

[24] Marco Bardus et al. "A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management". In: *International Journal of Behavioral Nutrition and Physical Activity* 13.1 (2016), p. 35.

[25] Milena Barz et al. "Self-efficacy, planning, and preparatory behaviours as joint predictors of physical activity: A conditional process analysis". In: *Psychology & health* 31.1 (2016), pp. 65–78.

[26] Stephanie Bauer et al. "Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight". In: *Patient education and counseling* 79.3 (2010), pp. 315–319.

[27] Jeannette M Beasley et al. "Evaluation of a PDA-based dietary assessment and intervention program: a randomized controlled trial". In: *Journal of the American College of Nutrition* 27.2 (2008), pp. 280–286.

[28] Lucian Bebchuk and Jesse Fried. *Pay without performance*. Vol. 29. Cambridge, MA: Harvard University Press, 2004.

[29] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. "Human-computer interaction and collective intelligence". In: *Handbook of collective intelligence* 57 (2015).

[30] Charles H Bombardier et al. "The relations of cognitive, behavioral, and physical activity variables to depression severity in traumatic brain injury: reanalysis of data from a randomized controlled trial". In: *The Journal of head trauma rehabilitation* 32.5 (2017), pp. 343–353.

[31] Michael L Booth et al. "Social–cognitive and perceived environment influences associated with physical activity in older Australians". In: *Preventive medicine* 31.1 (2000), pp. 15–22.

[32] Dena M Bravata et al. "Using pedometers to increase physical activity and improve health: a systematic review". In: *Jama* 298.19 (2007), pp. 2296–2304.

[33] Lori Breslow et al. "Studying learning in the worldwide classroom: Research into edX's first MOOC". In: *Research & Practice in Assessment* 8 (2013).

[34] J Brug et al. "The internet and nutrition education: challenges and opportunities". In: *European Journal of Clinical Nutrition* 59.S1 (2005), S130.

[35] Andrea Bunt, Matthew Lount, and Catherine Lauzon. "Are explanations always important?: a study of deployed, low-cost intelligent interactive systems". In: *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM. 2012, pp. 169–178.

[36] Lora E Burke et al. "The effect of electronic self-monitoring on weight loss and dietary intake: a randomized behavioral weight loss trial". In: *Obesity* 19.2 (2011), pp. 338–344.

[37] Michelle Clare Carter et al. "Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial". In: *Journal of medical Internet research* 15.4 (2013).

[38] Meredith A Case et al. "Accuracy of smartphone applications and wearable devices for tracking physical activity data". In: *JAMA* 313.6 (2015), pp. 625–626.

[39] Centers for Disease Control and Prevention. *Exercise or Physical Activity*. Tech. rep. 2017.

[40] Catherine B Chan, Daniel AJ Ryan, and Catrine Tudor-Locke. "Health benefits of a pedometer-based physical activity intervention in sedentary workers". In: *Preventive medicine* 39.6 (2004), pp. 1215–1222.

[41] Yan Chen et al. "Social comparisons and contributions to online communities: A field experiment on movielens". In: *American Economic Review* 100.4 (2010), pp. 1358–98.

[42] Lydia B Chilton et al. "Frenzy: collaborative data organization for creating conference sessions". In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM. 2014, pp. 1255–1264.

[43] Kristine K Clarke et al. "Promotion of physical activity in low-income mothers using pedometers". In: *Journal of the American Dietetic Association* 107.6 (2007), pp. 962–967.

[44] Doug Clow. “MOOCs and the funnel of participation”. In: *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM. 2013, pp. 185–189.

[45] Derrick Coetzee et al. “Chatrooms in MOOCs: all talk and no action”. In: *Proceedings of the first ACM conference on Learning@ scale conference*. ACM. 2014, pp. 127–136.

[46] Derrick Coetzee et al. “Should your MOOC forum use a reputation system?” In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 1176–1187.

[47] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[48] Kathleen Cotton. *Monitoring student learning in the classroom*. Northwest Regional Educational Laboratory, 1988.

[49] Cora L Craig et al. “International physical activity questionnaire: 12-country reliability and validity”. In: *Medicine and science in sports and exercise* 35.8 (2003), pp. 1381–1395.

[50] Karen A Croteau. “A preliminary study on the impact of a pedometer-based intervention on daily steps”. In: *American Journal of Health Promotion* 18.3 (2004), pp. 217–220.

[51] Carol O Cummins et al. “Assessing stage of change and informed decision making for Internet participation in health promotion and disease management.” In: *Managed care interface* 17.8 (2004), pp. 27–32.

[52] Todd Davies and Seeta Peña Gangadharan. “Online deliberation: Design, research, and practice”. In: (2009).

[53] Peter Denning et al. “Wikipedia risks”. In: *Communications of the ACM* 48.12 (2005), pp. 152–152.

[54] Elizabeth G Eakin, Russell E Glasgow, and Kimberly M Riley. “Review of primary care-based physical activity intervention studies”. In: *Journal of Family Practice* 49.2 (2000), pp. 158–158.

[55] Kathleen M Eisenhardt. “Agency theory: An assessment and review”. In: *Academy of management review* 14.1 (1989), pp. 57–74.

[56] Sean B Eom, H Joseph Wen, and Nicholas Ashill. “The determinants of students’ perceived learning outcomes and satisfaction in university online education: An empirical investigation”. In: *Decision Sciences Journal of Innovative Education* 4.2 (2006), pp. 215–235.

[57] Jason Fanning, Sean P Mullen, and Edward McAuley. “Increasing physical activity with mobile devices: a meta-analysis”. In: *Journal of medical Internet research* 14.6 (2012).

[58] Eric A Finkelstein et al. “A randomized study of financial incentives to increase physical activity among sedentary older adults”. In: *Preventive medicine* 47.2 (2008), pp. 182–187.

[59] Brianna S Fjeldsoe, Yvette D Miller, and Alison L Marshall. “MobileMums: a randomized controlled trial of an SMS-based physical activity intervention”. In: *Annals of Behavioral Medicine* 39.2 (2010), pp. 101–111.

[60] C Flaherty. “Zero correlation between evaluations and learning”. In: *Inside Higher Education* (2016).

[61] Rebecca Freeman and Kerry Dobbins. “Are we serious about enhancing courses? Using the principles of assessment for learning to enhance course evaluation”. In: *Assessment & Evaluation in Higher Education* 38.2 (2013), pp. 142–151.

[62] David P French et al. “Which behaviour change techniques are most effective at increasing older adults’ self-efficacy and physical activity behaviour? A systematic review”. In: *Annals of Behavioral Medicine* 48.2 (2014), pp. 225–234.

[63] Thomas Fritz et al. “Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 487–496.

[64] Yuichi Fujiki. “iPhone as a physical activity measurement platform”. In: *CHI’10 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2010, pp. 4315–4320.

[65] Yoshimi Fukuoka, William Haskell, and Eric Vittinghoff. “New insights into discrepancies between self-reported and accelerometer-measured moderate to vigorous physical activity among women–the mPED trial”. In: *BMC public health* 16.1 (2016), p. 761.

[66] Yoshimi Fukuoka, Eric Vittinghoff, and J Hooper. “A weight loss intervention using a commercial mobile application in Latino Americans - Adelgaza Trial”. In: *Translational Behavioral Medicine* (), In print.

[67] Yoshimi Fukuoka et al. “A novel diabetes prevention intervention using a mobile app: a randomized controlled trial with overweight adults at risk”. In: *American journal of preventive medicine* 49.2 (2015), pp. 223–237.

[68] Yoshimi Fukuoka et al. “Innovation to motivation – pilot study of a mobile phone intervention to increase physical activity among sedentary women”. In: *Preventive medicine* 51.3 (2010), pp. 287–289.

[69] Yoshimi Fukuoka et al. “The mPED randomized controlled clinical trial: applying mobile persuasive technologies to increase physical activity in sedentary women protocol”. In: *BMC Public Health* 11.1 (2011), p. 933.

[70] Y. Fukuoka et al. “Cluster Analysis of Objectively Measured Baseline Physical Activity Patterns in Women in the mPED Trial”. In: *JMIR Public Health and Surveillance* (2018). DOI: `10.2196/publichealth.9138`. URL: `http://dx.doi.org/10.2196/publichealth.9138`.

[71] Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. “An Intelligent Interface for Organizing Online Opinions on Controversial Topics”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM. 2017, pp. 119–123.

[72] Ben U Gelman et al. “Online Urbanism: Interest-based Subcultures as Drivers of Informal Learning in an Online Community”. In: *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM. 2016, pp. 21–30.

[73] Eric Gilbert. “Widespread underprovision on Reddit”. In: *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM. 2013, pp. 803–808.

[74] David Goldberg et al. “Using collaborative filtering to weave an information tapestry”. In: *Communications of the ACM* 35.12 (1992), pp. 61–70.

[75] Ken Goldberg et al. “Eigentaste: A constant time collaborative filtering algorithm”. In: *information retrieval* 4.2 (2001), pp. 133–151.

[76] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. “Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles”. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM. 2016, pp. 228–240.

[77] Charles Graham et al. “Seven principles of effective teaching: A practical lens for evaluating online courses”. In: *The technology source* 30.5 (2001), p. 50.

[78] Jeffrey A Greene, Christopher A Oswald, and Jeffrey Pomerantz. “Predictors of retention and achievement in a massive open online course”. In: *American Educational Research Journal* 52.5 (2015), pp. 925–955.

[79] Swathi Gujral et al. “The Role of Brain Structure in Predicting Adherence to A Physical Activity Regimen.” In: *Psychosomatic Medicine* (2017).

[80] Gurobi Optimization, Inc. *Gurobi Optimizer Reference Manual*. 2016. URL: `http://www.gurobi.com`.

[81] Dylan Hadfield-Menell et al. “Cooperative inverse reinforcement learning”. In: *Advances in neural information processing systems*. 2016, pp. 3909–3917.

[82] Tess J Harris et al. “What factors are associated with physical activity in older people, assessed objectively by accelerometry?” In: *British Journal of Sports Medicine* (2008).

[83] Health.gov. *2008 Physical activity guidelines for Americans*. `https://health.gov/paguidelines/pdf/paguide.pdf`. Accessed: 2017-09-30.

[84] Heart.gov. *Physical inactivity statistical fact sheet*. `https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_319589.pdf`. Accessed: 2017-09-30.

[85] Eric B Hekler et al. “Validation of physical activity tracking via android smartphones compared to ActiGraph accelerometer: laboratory-based and free-living validation studies”. In: *JMIR mHealth and uHealth* 3.2 (2015).

[86] Enamul Hoque and Giuseppe Carenini. “Convisit: Interactive topic modeling for exploring asynchronous online conversations”. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM. 2015, pp. 169–180.

[87] Cherilyn N Hultquist, Carolyn Albright, and Dixie L Thompson. “Comparison of walking recommendations in previously inactive women”. In: *Medicine & Science in Sports & Exercise* 37.4 (2005), pp. 676–683.

[88] Robert Hurling et al. “Using internet and mobile phone technology to deliver an automated physical activity program: randomized controlled trial”. In: *Journal of medical Internet research* 9.2 (2007).

[89] IBM. *IBM ILOG CPLEX Optimization Studio*. 2016.

[90] Yoshie Iwasaki et al. “Exercise Self-Efficacy as a Mediator between Goal-Setting and Physical Activity: Developing the Workplace as a Setting for Promoting Physical Activity”. In: *Safety and health at work* 8.1 (2017), pp. 94–98.

[91] John M Jakicic et al. “Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: the IDEA randomized clinical trial”. In: *Jama* 316.11 (2016), pp. 1161–1171.

[92] Barbara J Jefferis et al. “Adherence to physical activity guidelines in older adults, using objectively measured physical activity in a population-based study”. In: *BMC Public Health* 14.1 (2014), p. 382.

[93] Michael C Jensen and William H Meckling. “Theory of the firm: Managerial behavior, agency costs and ownership structure”. In: *Journal of financial economics* 3.4 (1976), pp. 305–360.

[94] Nam-Seok Joo and Bom-Taeck Kim. “Mobile phone short message service messaging for behaviour modification in a community-based weight control programme in Korea”. In: *Journal of Telemedicine and Telecare* 13.8 (2007), pp. 416–420.

[95] Rodney P Joseph et al. “Print versus a culturally-relevant Facebook and text message delivered intervention to promote physical activity in African American women: a randomized pilot trial”. In: *BMC women’s health* 15.1 (2015), p. 30.

[96] Julie J Keysor. “Does late-life physical activity or exercise prevent or minimize disablement?: a critical review of the scientific evidence”. In: *American journal of preventive medicine* 25.3 (2003), pp. 129–136.

[97] Arjun Kharpal. “Apple vs FBI: All you need to know”. In: *CNBC. Accessed September* 16 (2016), p. 2016.

[98] TL Khong. “The Validity and Reliability of the Student Evaluation of Teaching: A case in a Private Higher Educational Institution in Malaysia”. In: *International Journal for Innovation Education and Research* 2.9 (2016), pp. 57–63.

[99] Youngdeok Kim, Ilhyeok Park, and Minsoo Kang. “Convergent validity of the international physical activity questionnaire (IPAQ): meta-analysis”. In: *Public health nutrition* 16.3 (2013), pp. 440–452.

[100] Abby C King et al. “Promoting physical activity through hand-held computer technology”. In: *American journal of preventive medicine* 34.2 (2008), pp. 138–142.

[101] René F Kizilcec, Chris Piech, and Emily Schneider. “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”. In: *Proceedings of the third international conference on learning analytics and knowledge*. ACM. 2013, pp. 170–179.

[102] Mark Klein. “How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium”. In: *Center for Collective Intelligence working paper* (2011).

[103] Joseph A Knight. “Physical inactivity: associated diseases and disorders”. In: *Annals of Clinical & Laboratory Science* 42.3 (2012), pp. 320–337.

[104] Silvia Knobloch-Westerwick and Jingbo Meng. “Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information”. In: *Communication Research* 36.3 (2009), pp. 426–448.

[105] Tetsuro Kobayashi and Ken’ichi Ikeda. “Selective exposure in political web browsing: Empirical verification of ’cyber-balkanization’ in Japan and the USA”. In: *Information, Communication & Society* 12.6 (2009), pp. 929–953.

[106] Daphne Koller et al. “Retention and intention in massive open online courses: In depth”. In: *Educause review* 48.3 (2013), pp. 62–63.

[107] Artie Konrad et al. “Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. 2015, pp. 3829–3838.

[108] Joseph A Konstan et al. “GroupLens: applying collaborative filtering to Usenet news”. In: *Communications of the ACM* 40.3 (1997), pp. 77–87.

[109] Kenneth L Kraemer and John Leslie King. “Computer-based systems for cooperative work and group decision making”. In: *ACM Computing Surveys (CSUR)* 20.2 (1988), pp. 115–146.

[110] Travis Kriplean et al. “Supporting reflective public thought with considerit”. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM. 2012, pp. 265–274.

[111] Sanjay Krishnan et al. “Hirl: Hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards”. In: (2016). URL: `http://arxiv.org/abs/1604.06508`.

[112] Sanjay Krishnan et al. “Social influence bias in recommender systems: a methodology for learning, analyzing, and mitigating bias in ratings”. In: *Proceedings of the 8th ACM Conference on Recommender systems*. 2014, pp. 137–144.

[113] Sanjay Krishnan et al. “Using a social media platform to explore how social media can enhance primary and secondary learning”. In: *Learning International Networks Consortium (LINC) 2013 Conference*. 2013.

[114] Charlene Krueger and Lili Tian. “A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points”. In: *Biological research for nursing* 6.2 (2004), pp. 151–157.

[115] Kimberly Kulavic, Cherilyn N Hultquist, and John R McLester. “A comparison of motivational factors and barriers to physical activity among traditional versus nontraditional college students”. In: *Journal of American College Health* 61.2 (2013), pp. 60–66.

[116] Una Kyriacos, J Jelsma, and S Jordan. “Monitoring vital signs using early warning scoring systems: a review of the literature”. In: *Journal of nursing management* 19.3 (2011), pp. 311–330.

[117] Jean-Jacques Laffont and David Martimort. *The theory of incentives: the principal-agent model*. Princeton university press, 2009.

[118] Hyunho Lee and Youngseok Lee. “A Look at Wearable Abandonment”. In: *Mobile Data Management (MDM), 2017 18th IEEE International Conference on*. IEEE. 2017, pp. 392–393.

[119] Q Vera Liao and Wai-Tat Fu. “Can you hear me now?: mitigating the echo chamber effect by source position indicators”. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 184–196.

[120] Q Vera Liao and Wai-Tat Fu. “Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 2745–2754.

[121] Jennifer H Lin, Shumin M Zhang, and JoAnn E Manson. “Predicting adherence to tamoxifen for breast cancer adjuvant therapy and prevention”. In: *Cancer Prevention Research* 4.9 (2011), pp. 1360–1365.

[122] Rebecca Lindberg. “Active living: On the Road with the 10,000 Stepssm Program”. In: *Journal of the American Dietetic Association* 100.8 (2000), pp. 878–879.

[123] Greg Linden, Brent Smith, and Jeremy York. “Amazon. com recommendations: Item-to-item collaborative filtering”. In: *IEEE Internet computing* 7.1 (2003), pp. 76–80.

[124] LIFE Study Investigators*[* See Appendix for List of LIFE Study Investigators]. “Effects of a physical activity intervention on measures of physical performance: Results of the lifestyle interventions and independence for Elders Pilot (LIFE-P) study”. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 61.11 (2006), pp. 1157–1165.

[125] Paul Little et al. “A randomised controlled trial of three pragmatic approaches to initiate increased physical activity in sedentary patients with risk factors for cardiovascular disease.” In: *Br J Gen Pract* 54.500 (2004), pp. 189–195.

[126] Edwin A Locke and Gary P Latham. “Building a practically useful theory of goal setting and task motivation: A 35-year odyssey.” In: *American psychologist* 57.9 (2002), p. 705.

[127] Sui Mak, Roy Williams, and Jenny Mackness. “Blogs and forums as communication and learning tools in a MOOC”. In: *Proceedings of the 7th International Conference on Networked Learning 2010*. University of Lancaster. 2010.

[128] Bess H Marcus et al. “Physical activity behavior change: issues in adoption and maintenance.” In: *Health Psychology* 19.1S (2000), p. 32.

[129] Bess H Marcus et al. “Physical activity intervention studies”. In: *Circulation* 114.24 (2006), pp. 2739–2752.

[130] G Alan Marlatt and Judith R Gordon. *Relapse prevention: Maintenance strategies in addictive behavior change*. 1985.

[131] Herbert W Marsh and Lawrence A Roche. "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility." In: *American psychologist* 52.11 (1997), p. 1187.

[132] Jennifer E Mason et al. "Optimizing statin treatment decisions for diabetes patients in the presence of uncertain future adherence". In: *Medical Decision Making* 32.1 (2012), pp. 154–166.

[133] Mathworks. *The Mathworks Inc; 2016. MATLAB and ttatistics toolbox release 2016a*. `https://in.mathworks.com/products/new_products/release2016a.html`. Accessed: 2017-09-30.

[134] Edward McAuley and Bryan Blissmer. "Self-efficacy determinants and consequences of physical activity". In: *Exerc Sport Sci Rev* 28.2 (2000), pp. 85–88.

[135] Kevin D McCaul, Russell E Glasgow, and Lorraine C Schafer. "Diabetes regimen behaviors: Predicting adherence". In: *Medical Care* (1987), pp. 868–881.

[136] BD McCullough and Darrell Radson. "Analysing student evaluations of teaching: Comparing means and proportions". In: *Evaluation & Research in Education* 24.3 (2011), pp. 183–202.

[137] Ann McGinley and Rupert M Pearse. *A national early warning score for acutely ill patients*. 2012.

[138] Kathryn Mercer et al. "Behavior change techniques present in wearable activity trackers: a critical analysis". In: *JMIR mHealth and uHealth* 4.2 (2016).

[139] Dafna Merom et al. "Promoting walking with pedometers in the community: the step-by-step trial". In: *American Journal of Preventive Medicine* 32.4 (2007), pp. 290–297.

[140] Yonatan Mintz et al. "Behavioral Analytics for Myopic Agents". In: (2017). URL: `http://arxiv.org/abs/1702.05496`.

[141] Marc S Mitchell et al. "Financial incentives for exercise adherence in adults: systematic review and meta-analysis". In: *American journal of preventive medicine* 45.5 (2013), pp. 658–667.

[142] Donald Morley. "Assessing the reliability of student evaluations of teaching: choosing the right coefficient". In: *Assessment & Evaluation in Higher Education* 39.2 (2014), pp. 127–139.

[143] Sean A Munson and Sunny Consolvo. "Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity". In: *Pervasive computing technologies for healthcare (PervasiveHealth), 2012 6th international conference on*. IEEE. 2012, pp. 25–32.

[144] Elena N Naumova, Aviva Must, and Nan M Laird. "Tutorial in biostatistics: evaluating the impact of 'critical periods' in longitudinal studies of growth using piecewise mixed effects models". In: *International journal of epidemiology* 30.6 (2001), pp. 1332–1341.

[145] Mark EJ Newman. “Clustering and preferential attachment in growing networks”. In: *Physical review E* 64.2 (2001), p. 025102.

[146] Andrew Y Ng, Stuart J Russell, et al. “Algorithms for inverse reinforcement learning.” In: *Icml*. 2000, pp. 663–670.

[147] Ellinor K Olander et al. “What are the most effective techniques in changing obese individuals’ physical activity self-efficacy and behaviour: a systematic review and meta-analysis”. In: *International Journal of Behavioral Nutrition and Physical Activity* 10.1 (2013), p. 29.

[148] Gillian A O’Reilly and Donna Spruijt-Metz. “Current mHealth technologies for physical activity assessment and promotion”. In: *American journal of preventive medicine* 45.4 (2013), pp. 501–507.

[149] Bo Pang, Lillian Lee, et al. “Opinion mining and sentiment analysis”. In: *Foundations and Trends in Information Retrieval* 2.1–2 (2008), pp. 1–135.

[150] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[151] R Paterson et al. “Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit”. In: *Clinical Medicine* 6.3 (2006), pp. 281–284.

[152] Kevin Patrick et al. “A text message–based intervention for weight loss: randomized controlled trial”. In: *Journal of medical Internet research* 11.1 (2009).

[153] Virginia Paul-Ebhohimhen and Alison Avenell. “Systematic review of the use of financial incentives in treatments for obesity and overweight”. In: *Obesity Reviews* 9.4 (2008), pp. 355–367.

[154] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[155] Christian Pentzold. “Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their ’community’?” In: *New Media & Society* 13.5 (2011), pp. 704–721.

[156] Pew Research Center. *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*. Tech. rep. 2016.

[157] Sarah M Phillips et al. “Energy-dense snack food intake in adolescence: longitudinal relationship to weight and fatness”. In: *Obesity* 12.3 (2004), pp. 461–472.

[158] Eva Pila et al. “Self-conscious emotions in response to physical activity success and failure: Findings from a global 112-day pedometer intervention”. In: *Journal of Exercise, Movement, and Sport (SCAPPS refereed abstracts repository)* 48.1 (2016).

[159] Josée Poirier et al. “Effectiveness of an activity tracker-and internet-based adaptive walking program for adults: a randomized controlled trial”. In: *Journal of medical Internet research* 18.2 (2016).

[160] Vincent Price and Joseph N Cappella. “Online deliberation and its influence: The electronic dialogue project in campaign 2000”. In: *IT & Society* 1.1 (2002), pp. 303–329.

[161] Barton K Pursel et al. “Understanding MOOC students: motivations and behaviours indicative of MOOC completion”. In: *Journal of Computer Assisted Learning* 32.3 (2016), pp. 202–217.

[162] I Quillen. “Why do students enroll in (but don’t complete) MOOC courses”. In: *Mind/Shift* (2013).

[163] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[164] Lynda B Ransdell et al. “Generations exercising together to improve fitness (GET FIT): a pilot study designed to increase physical activity and improve health-related fitness in three generations of women”. In: *Women & health* 40.3 (2005), pp. 77–94.

[165] Eric Rasmusen and Basil Blackwell. “Games and information”. In: *Cambridge, MA* 15 (1994).

[166] Justin Reich et al. “Computer-assisted reading and discovery for student generated text in massive open online courses”. In: (2014).

[167] William T Riley et al. “Health behavior models in the age of mobile interventions: are our theories up to the task?” In: *Translational behavioral medicine* 1.1 (2011), pp. 53–71.

[168] Jillian Ryan, Sarah Edney, and Carol Maher. “Engagement, compliance and retention with a gamified online social networking physical activity intervention”. In: *Translational behavioral medicine* 7.4 (2017), pp. 702–708.

[169] Mehdi SM Sajjadi, Morteza Alamgir, and Ulrike von Luxburg. “Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines”. In: *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM. 2016, pp. 369–378.

[170] Niloufar Salehi et al. “We are dynamo: Overcoming stalling and friction in collective action for crowd workers”. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM. 2015, pp. 1621–1630.

[171] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. “Experimental study of inequality and unpredictability in an artificial cultural market”. In: *science* 311.5762 (2006), pp. 854–856.

[172] Brian M Sandroff et al. “Randomized controlled trial of physical activity, cognition, and walking in multiple sclerosis”. In: *Journal of neurology* 261.2 (2014), pp. 363–372.

[173] Cristina Sarasua, Elena Simperl, and Natalya F Noy. “Crowdmap: Crowdsourcing ontology alignment with microtasks”. In: *International Semantic Web Conference*. Springer. 2012, pp. 525–541.

[174] Badrul Sarwar et al. “Item-based collaborative filtering recommendation algorithms”. In: *Proceedings of the 10th international conference on World Wide Web*. ACM. 2001, pp. 285–295.

[175] Craig N Sawchuk et al. “Peer Reviewed: Barriers and Facilitators to Walking and Physical Activity Among American Indian Elders”. In: *Preventing chronic disease* 8.3 (2011).

[176] Patrick L Schneider et al. “Effects of a 10,000 steps per day goal in overweight adults”. In: *American Journal of Health Promotion* 21.2 (2006), pp. 85–89.

[177] Douglas Schuler. “Online civic deliberation with e-Liberate”. In: *Online deliberation: Design, research, and practice* (2009), pp. 293–302.

[178] Mical Kay Shilts, Marcel Horowitz, and Marilyn S Townsend. “Goal setting as a strategy for dietary and physical activity behavior change: a review of the literature”. In: *American Journal of Health Promotion* 19.2 (2004), pp. 81–93.

[179] Cara L Sidman, Charles B Corbin, and Guy Le Masurier. “Promoting physical activity among sedentary women using pedometers”. In: *Research Quarterly for Exercise and Sport* 75.2 (2004), pp. 122–129.

[180] Joanna E Siegel, Jessica Waddell Heeringa, and Kristin L Carman. “Public deliberation in decisions about health research”. In: *Virtual Mentor* 15.1 (2013), p. 51.

[181] Ronald J Sigal et al. “Physical activity/exercise and type 2 diabetes: a consensus statement from the American Diabetes Association”. In: *Diabetes care* 29.6 (2006), pp. 1433–1438.

[182] Laurey R Simkin and Alan M Gross. “Assessment of coping with high-risk situations for exercise relapse among healthy women.” In: *Health Psychology* 13.3 (1994), p. 274.

[183] Stephanie Solomon and Julia Abelson. “Why and when should we use public deliberation?” In: *Hastings Center Report* 42.2 (2012), pp. 17–20.

[184] Pieter Spooren, Bert Brockx, and Dimitri Mortelmans. “On the validity of student evaluation of teaching: The state of the art”. In: *Review of Educational Research* 83.4 (2013), pp. 598–642.

[185] Philip B Stark and Richard Freishtat. “An evaluation of course evaluations”. In: *Science Open Research* 9 (2014).

[186] State of Obesity. *Physical inactivity in the United States*. `https://stateofobesity.org/physical-inactivity/`. Accessed: 2017-09-30.

[187] Janna Stephens and Jerilyn Allen. “Mobile phone interventions to increase physical activity and reduce weight: a systematic review”. In: *The Journal of cardiovascular nursing* 28.4 (2013), p. 320.

[188] Kristin Stephens-Martinez, Marti A Hearst, and Armando Fox. “Monitoring moocs: which information sources do instructors value?” In: *Proceedings of the first ACM conference on Learning@ scale conference*. ACM. 2014, pp. 79–88.

[189] Barbara A Stetson et al. "Prospective evaluation of the effects of stress on exercise adherence in community-residing women". In: *Health Psychology* 16 (1997), pp. 515–520.

[190] Valerie E Stone et al. "Perspectives on adherence and simplicity for HIV-infected patients on antiretroviral therapy: self-report of the relative importance of multiple attributes of highly active antiretroviral therapy (HAART) regimens in predicting adherence". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 36.3 (2004), pp. 808–816.

[191] Steven D Stovitz et al. "Pedometers as a means to increase ambulatory activity for patients seen at a family medicine clinic". In: *The Journal of the American Board of Family Practice* 18.5 (2005), pp. 335–343.

[192] Andreas Ströhle. "Physical activity, exercise, depression and anxiety disorders". In: *Journal of neural transmission* 116.6 (2009), p. 777.

[193] Abhay Sukumaran et al. "Normative influences on thoughtful online participation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2011, pp. 3401–3410.

[194] Paul WG Surgenor. "Obstacles and opportunities: addressing the growing pains of summative student evaluation of teaching". In: *Assessment & Evaluation in Higher Education* 38.3 (2013), pp. 363–376.

[195] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[196] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation". In: *Advances in neural information processing systems*. 2000, pp. 1057–1063.

[197] Karen Swan. "Virtual interaction: Design factors affecting student satisfaction and perceived learning in asynchronous online courses". In: *Distance education* 22.2 (2001), pp. 306–331.

[198] Adam G Tabák et al. "Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study". In: *The Lancet* 373.9682 (2009), pp. 2215–2221.

[199] Paul D Thompson et al. "Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease: a statement from the Council on Clinical Cardiology (Subcommittee on Exercise, Rehabilitation, and Prevention) and the Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity)". In: *Circulation* 107.24 (2003), pp. 3109–3116.

[200] W Ben Towne and James D Herbsleb. "Design considerations for online deliberation systems". In: *Journal of Information Technology & Politics* 9.1 (2012), pp. 97–115.

[201] Catrine Tudor-Locke. "Taking steps toward increased physical activity: Using pedometers to measure and motivate." In: *President's Council on Physical Fitness and Sports Research Digest* (2002).

[202] Gabrielle M Turner-McGrievy et al. “Comparison of traditional versus mobile app self-monitoring of physical activity and dietary intake among overweight adults participating in an mHealth weight loss program”. In: *Journal of the American Medical Informatics Association* 20.3 (2013), pp. 513–518.

[203] U.S. Department of Health and Human Services. *Physical Activity Guidelines for Americans*. 2008.

[204] Marleen H Van den Berg, Johannes W Schoones, and Theodora PM Vliet Vlieland. “Internet-based physical activity interventions: a systematic review of the literature”. In: *Journal of medical Internet research* 9.3 (2007).

[205] Elizabeth M Venditti et al. “Short and long-term lifestyle coaching approaches used to address diverse participant barriers to weight loss and physical activity adherence”. In: *International Journal of Behavioral Nutrition and Physical Activity* 11.1 (2014), p. 16.

[206] Geert Verbeke. “Linear mixed models for longitudinal data”. In: *Linear mixed models in practice*. Springer, 1997, pp. 63–153.

[207] Ted Vickey, John Breslin, and Antonio Williams. “Fitness–There’s an App for That: Review of Mobile Fitness Apps.” In: *International Journal of Sport & Society* 3.4 (2012).

[208] C Vorwerk. “MEWS: predicts hospital admission and mortality in emergency department patients”. In: *Emergency Medicine Journal* 26.6 (2009), pp. 466–466.

[209] Julie B Wang et al. “Wearable sensor/device (Fitbit One) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial”. In: *Telemedicine and e-Health* 21.10 (2015), pp. 782–792.

[210] Yuan Wang. “Exploring possible reasons behind low student retention rates of massive online open courses: A comparative case study from a social cognitive perspective”. In: *AIED 2013 Workshops Proceedings Volume*. Citeseer. 2013, p. 58.

[211] Dong Wen et al. “Evaluating the consistency of current mainstream wearable devices in health monitoring: a comparison under free-living conditions”. In: *Journal of medical Internet research* 19.3 (2017).

[212] Bernadette R Williams et al. “The effect of a behavioral contract on adherence to a walking program in postmenopausal African American women”. In: *Topics in Geriatric Rehabilitation* 21.4 (2005), pp. 332–342.

[213] B Williams et al. “National Early Warning Score (NEWS): standardising the assessment of acute-illness severity in the NHS”. In: *London: The Royal College of Physicians* (2012).

[214] Edwin B Wilson. “Probable inference, the law of succession, and statistical inference”. In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212.

[215] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[216] Anita Williams Woolley et al. “Evidence for a collective intelligence factor in the performance of human groups”. In: *science* 330.6004 (2010), pp. 686–688.

[217] World Health Organization. *Health topics: physical activity*. http://www.who.int/topics/physical_activity/en/. Accessed: 2017-09-30.

[218] World Health Organization. *Health topics: physical activity*. 2017. URL: http://www.who.int/topics/physical_activity/en/.

[219] World Health Organization. *Physical activity fact sheet*. 2017. URL: http://www.who.int/mediacentre/factsheets/fs385/en/.

[220] World Heart Federation. *Physical inactivity*. 2017. URL: http://www.world-heart-federation.org/cardiovascular-health/cardiovascular-disease-risk-factors/physical-inactivity/.

[221] Judith Wylie-Rosett et al. "Computerized weight loss intervention optimizes staff time: the clinical and cost results of a controlled clinical trial conducted in a managed care setting". In: *Journal of the Academy of Nutrition and Dietetics* 101.10 (2001), pp. 1155–1162.

[222] Yao Xiong et al. "Examining the relations among student motivation, engagement, and retention in a MOOC: A structural equation modeling approach". In: *Global Education Review* 2.3 (2015).

[223] Xiwang Yang et al. "A survey of collaborative filtering based social recommender systems". In: *Computer Communications* 41 (2014), pp. 1–10.

[224] Saijing Zheng et al. "Understanding student motivation, behaviors and perceptions in MOOCs". In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM. 2015, pp. 1882–1895.

[225] M. Zhou et al. "Evaluating Machine Learning Based Automated Personalized Daily Step Goals Delivered through a Mobile Phone App: a Randomized Controlled Trial". In: *JMIR Mhealth Uhealth* (2018). To appear.