Essays on Gamified Online Platforms

Behnaz Ghahestani Bojd

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Yong Tan, Chair

Hema Yoganarasimhan

Ming Fan

Program Authorized to Offer Degree:
Foster School of Business

University of Washington

Abstract

Essays on Gamified Online Platforms

Behnaz Ghahestani Bojd

Chair of the Supervisory Committee:
Michael G. Foster Professor Yong Tan
Information Systems and Operations Management

Gamified online platforms use game design elements in non-game contexts to increase users' engagement and improve their performance outcomes. In this dissertation, using data from two gamified online platforms, I study the effect of disclosing individual's performance ranking and popularity rating on different outcomes. First, I focus on an online weight-loss community which uses gamified challenges to enable users to set short-term weight-loss goals, and incentivize them to pursue their goals by sharing individuals' progress and rankings with other challenge members via leaderboards. I study the effect of participation in gamified challenges on weight-loss progress. I utilize the system GMM and Inverse Probability Weighting (IPW) approach to address endogeneity issues. The results indicate that participation in gamified challenges has a positive and significant but short-term effect on weight-loss. Moreover, participation in gamified challenges are less effective when users focus only on dietary or physical activity instructions, and more effective when users add a numeric weight-loss target to their dietary or physical activity instructions. Second, I focus on a gamified dating platform where users play a game and rank-order members of the opposite sex and are then matched based on a Stable Matching Algorithm. A key piece of information shown to users during the game is a popularity rating, ranging from one to three stars. I examine the effect of a user's popularity on her demand i.e. I quantify the causal effect of a user's star-rating on the rankings that s/he receives during the game and the likelihood of receiving messages

after the game. Popularity can increase one's appeal. However, popular people are less likely to reciprocate, and hence users may strategically shade their revealed preferences for them to avoid rejection. To overcome the endogeneity between a user's star-rating and her unobserved attractiveness, I employ non-linear fixed-effects models. The results indicate that three-star users receive worse rankings during the game but receive more messages after. I link the heterogeneity across outcomes and user-level observables to the perceived severity of rejection concerns and establish strategic shading as the mechanism. Further, I show that users' rejection concerns are consistent with Step-1 bounded rationality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude to those who helped me to pursue my doctoral studies at the University of Washington. I wish to express my sincere appreciation to my Ph.D. advisor Professor Yong Tan who gave me the opportunity to start my doctoral studies in the field of Information Systems. His patience and continuous support empowered me to explore the research questions that interest me and encouraged me to think more independently. He has been my mentor and generously helped me to find great resources and datasets. I am grateful and inspired by his research enthusiasm, methodological rigor and at the same time his lighthearted attitude towards work.

Also, I would like to express my sincere gratitude to Professor Hema Yoganarasimhan for mentoring me in the second essay of this dissertation. She shaped my approach to research. She taught me a lot, yet always made me feel like an equal coauthor. Her professional and emotional support has greatly influenced me. I also would like to thank my other committee member, Professor Ming Fan, for his support and guidance.

I would like to express special thanks to Shawna Reimers, Jessica Aceves, Jaime Banaag, and Beau Kirkeby who not only helped me patiently in different administrative processes but also supported me like members of a caring family in tough times. I would also want to thank my friends in Mackenzie Hall: Elnaz Jalilipour, Shima Nassiri, Sareh Nabi, Amir Fazli, Aravinda Garimella, Jane Xue Tan, Jinyang Zheng, Mohsen Sharifani, Shahryar Doosti, Mohammad Arbabian, Omid Rafieian, Eugene Pavlov, and Maria Mitkina. I am grateful for all the things you taught me and for making the Ph.D. years unbelievably fun.

Finally, I would like to express my deepest gratitude to my parents, whose unconditional love, patience and support encouraged me to pursue my Ph.D. in long years far from them.

DEDICATION

to my beloved parents, Mahvash and Hossein.

&

to my dear brother, Amir.

Chapter 1
INTRODUCTION

Gamified online platforms use game design elements in non-game contexts to increase users' engagement and improve their performance outcomes. In §1.1, I focus on an online weight-loss community which uses gamified challenges to enable users to set short-term weight-loss goals, and incentivize them to pursue their goals by sharing individuals' progress and rankings with other challenge members via leaderboards. In §1.2, I focus on a gamified dating platform where users play a game and rank-order members of the opposite sex and are then matched based on a Stable Matching Algorithm. A key piece of information shown to users during the game is a popularity rating, ranging from one to three stars. In this dissertation, using data from these two gamified online platforms, I study the effect of disclosing individual's performance ranking and popularity rating on different outcomes.

1.1 Introduction: A Gamified Online Weight-loss Community

Obesity has been associated with many adverse physical and psychological health conditions [33]. It has also become highly prevalent and a significant economic crisis costing healthcare systems billions of dollars per year [14, 32]. With the rise of health 2.0 movement, online weight-loss communities have emerged as a new tool with a mission to support individuals to lose weight. These communities connect users with similar weight-loss goals, and give them access to many features including 1) self-monitoring tools, enabling users to report and track their weight, food intake and exercise, 2) forums and groups, where users with common interests can share their weight-loss questions and experiences, 3) challenges, which let users set short-term goals, and incentivize them to pursue their goals with gamification elements such as cash prize, badges, and leaderboards.

Weight-loss challenges are one of the most popular features of online weight-loss communities. This feature became popular after the hit TV reality show "The Biggest Loser", where contestants compete to lose weight and win a cash prize. Online weight-loss challenges provide a tool for users to set short-term weight-loss goals, and share them with other users. These weight-loss goals usually focus on a numeric weight-loss target e.g. "to lose 10 kg", and/or a set of physical activity or dietary instructions such as "walking 10,000 steps a day" or "restricting daily caloric intake by 1,250 calories". Beyond providing a goal-setting tool, weight-loss challenges can enhance users' motivation via gamification tools, based on their weight-loss progress. For example, a leaderboard ranks challenge participants based on their weight-loss progress, enabling users to compare their progress with others and keep them motivated during the challenge period.

Despite the popularity of online weight-loss challenges, there is no research that examines or quantifies the effect of participation in gamified challenges on user's weight-loss progress. A few papers in the literature have shown the positive effect of participation in gamified walking contests on users' physical activity [17, 72]. However, there is a main difference between physical activity contests and weight-loss challenges. In physical activity contests, participants get incentivized based on the physical activity measure (e.g. get ranked via leaderboards based on number of steps). However, in weight-loss challenges, participants get incentivized based on their weight-loss outcome, i.e. how many kilograms they have lost. Individuals may have a higher control on their number of steps; however, they may not be able to directly control their weight-loss outcome. Unlike physical activity, weight-loss is a complicated process. Weight-loss is determined not only by individuals' motivation, but also by a variety of other factors such as genetic predisposition, stress, and environmental factors such as family and friends [33]. Therefore, based on prior literature, it is not clear whether participation in gamified weight-loss challenges can affect users' weight-loss progress.

In chapter 3, I am interested to answer three key questions. First, I seek to estimate the causal effect of users' participation in online gamified challenges on their weight-loss progress. Second, I am interested to quantify the long-term effect of participation in online weight-loss

challenges, if I find any positive short-term effect. It has been shown that most people, who succeed in losing weight in a short period, regained a substantial amount of their lost weight after participation in the weight-loss competitions [34]. Thus, it is important to quantify the long-term effect of participation in online gamified challenges. Finally, I am interested to measure the heterogeneous effect of challenges across different goal types. Many weight-loss challenges focus on healthy lifestyle changes, i.e. the challenge goal includes a set of dietary or physical activity instructions. However, some other challenges include a numeric weight-loss target. Incorporating a numeric weight-loss target in a challenge can have both positive and negative effects on weight-loss progress. On the one hand, since gamified challenges incentivize users based on the magnitude of their weight-loss outcome, a numeric weight-loss target can motivate users to stick to their goals. On the other hand, a numeric weight-loss target may draw users' attention to the final outcome and impair their focus on healthy lifestyle changes, and discourage them at times of slow weight-loss progress.

I empirically examine these questions using the data from a leading online weight-loss community in the Unites States. With an emphasis on user-generated content, this online weight-loss community lets users track their weight, nutrition, and physical activity and engage with a supportive community via general forums and specific groups. In mid-2008, weight-loss challenges were introduced as a new feature on the platform. This feature enables users to create/join challenges, where they can set and pursue short-term goals. These challenges are gamified via leaderboards, which rank users based on their weight-loss progress. For the empirical analysis, I randomly chose users and tracked them for eight months from April to November 2008. In the panel data, for each user at each month, I can observe whether s/he participates in a particular challenge. Further, I have access to the average weight that each user reports during a month, and I can calculate their monthly weight-loss progress. I can also observe users' engagement with other platform's features, their tenure on the platform, and their personal weight-loss goals.

There are three estimation challenges regarding the estimation of the causal effect of users' participation in challenges on their weight-loss progress. First, an individual's weight

at each point in time is highly dependent on its previous value. Second, weight-loss/gain is determined by a variety of unobserved elements such as individuals' motivation, or genetic predisposition, gender, etc. These unobserved fixed or time-varying factors could be correlated with users' decision to participate in a challenge. To overcome these concerns, I employ a dynamic model using a system GMM estimator (Blundell-Bond). The system GMM estimation utilizes instruments within the model, and the validity of these instruments is verified using Hansen and Arellano-Bond tests. Moreover, the system GMM model is useful to accurately estimate the inertial effects of the lagged dependent variable in the dynamic model and is used to calculate the long-term effect of users' participation in challenges on their weight-loss progress. The third estimation challenge stems from the fact that some individuals never participate in any challenge during the panel time frame. These challenge non-adopters do not appear in the system GMM analysis. This can result in a biased estimation, called incidental sample truncation, when non-adopters' decision to participate is non-random. To address the incidental sample truncation, I combine the system GMM estimation with the inverse probability weighting (IPW) approach. The validity of the IPW approach relies on the ignorability assumption. Although, it is not possible to directly test this assumption, I provide an indirect verification test.

the main findings indicate that participation in gamified challenges has a positive and significant effect on weight-loss. Users can achieve a weight-loss of 0.945 kg by participating in at least one challenge a month. Further, based on the results from the system GMM estimation, I calculate the long-term effect of participation in challenges, and find that a challenge participant will gain the lost weight back in a few months, if s/he does not participate in a new challenge in future. Finally, I show heterogeneous effects of challenges across different goal types controlling for other challenge attributes. The results suggest that challenges with a numeric weight-loss target (e.g. to lose 5 Kg) are more effective than challenges without a target.

The main contributions of this study are as follows. First, I quantify and isolate the effect of users' participation in gamified challenges on weight-loss from their engagement

with other platform's features such as self-monitoring tools and weight-loss forums/groups. A large stream of research studies the effect of eHealth interventions on weight-loss using randomized trials [41]. However, in those studies, a combination of different online features is used and it is not possible to determine which features of eHealth weight-loss interventions are driving the effect on the weight-loss progress. Second, this study is the first study to examine the effect of both performance and process goals in gamified weight-loss interventions. A performance goal refers to a numeric weight-loss target (e.g. to lose 10 Kg), and a process goal refers to a set of dietary or physical activity instructions, which can help in achieving the performance goal (e.g. walk 10,000 steps a day). I showed that a gamified weight-loss challenge with a numeric performance goal and detailed process goals is most effective. This finding can serve as a guideline for designing gamified information systems in goal-setting environments. Finally, I discuss and clarify the methodological strategies required to analyze the dynamic nature of the weight-loss outcome and overcome endogeneity problems in non-experimental settings. I compare the system GMM results with a difference-in-difference model with time-varying treatment, coupled with propensity score matching. Although the effect of challenges on weight-loss remains positive and significant in a difference-in-difference model, the magnitude of the effect is smaller and close to the magnitude of a biased fixed-effect model. Moreover, I show that a difference GMM dynamic approach performs poorly in the setting due to the high inertial effects of the lagged dependent variable.

The results have implications for the design of online weight-loss communities. I show that gamified challenges are an effective feature of online weight-loss communities in improving weight-loss outcome; however, users will regain their lost weight when they cease participating in challenges. Thus, it is important for online weight-loss communities to have strategies to help users achieve and maintain higher weight-loss goals over time. Further, I show that effective gamified weight-loss challenges should incorporate a numeric weight-loss target, as well as detailed instructions for healthy lifestyle change. Although, gamified challenges are user generated, online weight-loss communities can suggest users to set more effective goals.

The details of this essay is organized as follows. In §2.1 I review the related literature. I introduce the setting and data in §3.1 and §3.2. In §3.3, I discuss the empirical model. In §3.4, I present the results. In §3.5, I provide the robustness checks. In §5.1, I review the main findings of this essay, and the limitations, and suggestions for future research.

1.2 Introduction: A Gamified Online Dating Platform

Throughout human history, people have relied on their extended families, social networks, and religious organizations to help them find romantic partners. However, they are now increasingly turning to online dating for this purpose. The most recent *Singles in America Survey* found that the number one meeting place for singles is now online [70]. According to a study from Pew Research Center, 15% of U.S. adults ($\approx$ 40 million adults) reported that they have used online dating services [73]. Indeed, industry revenues for online dating now exceed three billion dollars a year in the United States [43].

Early businesses in this industry were mostly websites that allowed users to create detailed profiles, browse/search other users' profiles, and then establish contact through email exchanges. However, over the years, mobile dating apps have replaced dating websites as the dominant form of online dating because they offer a much simpler way for users to find matches [54]. First, users are shown a set of potential partners and asked to state their preference for them on some scale (e.g., rank-order them, vote up or down, or swipe right or left) within a fixed period of time. These stated-preferences are then fed into a matching schema/algorithm, which matches users who have expressed some preference for each other. The first step eliminates the need for users to browse and search profiles, and the second step ensures that users are not spending effort in crafting and sending emails to potential partners who have no interest in them.

The way information is presented in mobile dating apps has also evolved to reflect the simpler search process. Because users are only given a short (and fixed) amount of time to decide how much they like someone, most dating apps have moved away from showing long detailed profiles. Instead, they show a small set of salient pieces of information that a user

can process easily (e.g., photo and age of the potential partner). Many of them also display a summary measure of the popularity of a potential partner (e.g., star-rating, likes) next to her/his profile. The benefits of showing users' popularity information are that – (a) it is easier to process one cumulative popularity measure instead of parsing through detailed profile data, and (b) popularity measures can provide information on a potential partner's appeal in the dating market, and thereby help users calibrate the likelihood of achieving a match with that person.

However, there is no research that examines or quantifies the effect of such popularity measures on users' demand in dating platforms. A large stream of literature on e-commerce and online marketplaces has shown that displaying popularity information about products/sellers can have a positive impact on their demand [74, 81]. But those settings did not involve interpersonal interactions. Moreover, the mechanisms at play in e-commerce markets are likely to be quite different from those in dating contexts. Hence, the extent to which these results will translate to an online dating context is not clear.

In this essay, I am interested in two key questions related to popularity information and demand in online dating.

- First, I seek to quantify the causal effect of a user's popularity information on her/his demand measures in the dating market.
- Second, I am interested in identifying the source of these effects (if any), i.e., pin down the mechanism behind them.

In dating contexts, popularity information can have both positive and negative impact on demand. On the one hand, revealing that a potential partner is popular can increase her/his appeal, which in turn can increase a user's revealed preference for that potential partner [37]. On the other hand, a very popular potential partner is also more likely to have other options (or interest from other users) and is therefore less likely to reciprocate any interest. Thus, a user who wants to avoid the psychological costs of rejection may reveal lower preference (or *strategically shade* down her/his preference) for a popular user. A priori, it is not clear which of these effects will dominate, and what would be the overall impact of popularity

information on demand.

I empirically examine these questions using data from a popular mobile dating app in the United States during the 2014-15 time-frame. Users in the app are matched based on games where they rank members of the opposite sex. Each game starts with the random assignment of four men and four women to a virtual room. Then, each player has ninety seconds to rank-order members of the opposite sex from one to four, with one indicating the most preferred partner and four the least. (Throughout this essay, I use the term preference-ranking, which is reverse of ranking, to indicate users' ordered preferences to simplify exposition.[1])

The platform then uses these preference-rankings as inputs into a Stable Match Algorithm and matches each player in the room with a member of the opposite sex. After the game ends, users can initiate contact with their matched players and chat with them (if their matched partner reciprocates).

A key piece of information shown to users during and after the game is a star-rating for each member of the opposite sex (ranging from one to three stars). A user's star-rating is a cumulative measure of all the preference-rankings that s/he received in the past. So users with higher past preference-rankings are shown with higher stars. Stars are thus a salient and visible indicator of a user's popularity on the platform. At the same time, they do not contain any extra information on the unobserved quality of the user since they are not based on contact/engagement between previous raters and the ratee. They are, thus, pure popularity measures and do not help resolve asymmetric information about the user's quality as a date (unlike star-ratings based on purchase/experience in e-commerce settings).

This study consists of two major components, which mirror the two broad research questions. To answer the first research question, I quantify the causal impact of a user's star-rating on three demand measures: (1) preference-rankings received during a game, (2) likelihood of receiving a first message from the matched partner after the game, (3) likelihood of receiving a reply to a message sent after the game. There are two main challenges in

[1]Rank of one denotes a preference-ranking of four, rank of two indicates a preference-ranking of three, and so on.

this task. First, a user's star-ratings and her/his unobserved attractiveness are confounded: attractive users who received high preference-rankings in the past (and hence have higher stars now) are also likely to receive higher preference-rankings now – not necessarily because of their star-rating, but due to their inherent attractiveness, which may be unobservable to the researcher (e.g., great bio descriptions, fun-loving pictures). This can give rise to an upward bias in the estimates of the effect of star-ratings if I use naive estimation strategies. To overcome this challenge, I leverage the fact that a user's star-rating is not static; rather it changes over the course of the observation period as a function of her/his rankings in the previous games. Thus, I can use the *within*-person variation in star-ratings to causally infer the effect of a user's star-rating on her demand in the marketplace.

The second estimation challenge stems from the non-linearity of the three demand measures: the first measure (preference-ranking) is an ordered discrete outcome, and the other two measures (first and reply messages) are binary outcomes. I model the first measure using a fixed-effects ordered logit model, and the latter two are modeled using fixed-effects binary logit models. In all these models, I allow user-specific unobservables (i.e., the fixed-effects) to be arbitrarily correlated with star-ratings. While fixed effects are needed to control for the endogeneity issues discussed earlier, estimation of ordered and binary logit models with fixed-effects is tricky since there is no easy way to subtract out the unobserved user fixed-effect in a non-linear setting. To address this issue, Chamberlain proposed a general class of Conditional Maximum Likelihood (CML) estimators for non-linear models that condition on a subset of outcomes, which in turn allows them to condition-out all the fixed-effects (or nuisance parameters) and estimate only the main parameters of interest [15]. Usually, in a K outcome ordered logit model, we can derive $K-1$ consistent CML estimates. However, these $K-1$ estimates are inefficient because each of them only uses only part of the variation in the data for identification. A Minimum Distance (MD) estimator is developed that combines all the CML estimators and generates both consistent and efficient estimates [21]. I use this estimator to derive the effect of star-ratings on preference-rankings in this setting. For the two message-related binary outcome models, the CML and MD estimators are equivalent,

so I simply use Chamberlain's CML for them. Note that all these estimators rely on the within-user variation in star-ratings to identify the effect of stars on outcomes, and thereby address the endogeneity issues discussed earlier.

I now discuss the main findings from the first part of this essay. Everything else being constant, three-star users receive lower preference-rankings compared to two-star users during the game, i.e., popularity has a *negative* effect on preference-rankings. I also find that ignoring endogeneity problems would lead to draw the exact opposite conclusion. Interestingly, the effect of star-rating is different in after-game outcomes. In particular, three-star users are more likely to receive both first messages and replies after the game. These results suggest that users in the platform respond differently to popularity information at different stages of the matching process.

Next, I focus on the second research question, regarding the source of the popularity effect. Here, I leverage the differences in the risk of rejection across the observed demand measures and show that the negative effect of star-ratings during the game can be attributed to strategic shading. When a user is ranking someone during a game, s/he has no information on the other person's preferences, thus the potential for being rejected (i.e., not being matched) is high. In contrast, in the reply message case, the user has already received a message from her/his match and is considering whether to reply or not. Here, rejection is not a concern at all since the other party has already expressed interest. Using the fact that the effect of star-rating in the reply case is strictly positive, I can show that the negative effect of star-rating during the game is due to strategic shading.

I also provide additional evidence for strategic shading based on the heterogeneity in the effect of star ratings across user-level attributes. In particular, I show that the negative effect of star-ratings on preference-rankings is mainly driven by less-attractive users when they are ranking attractive potential partners. Since less-attractive users are more likely to have rejection concerns (especially when they are ranking attractive users), this finding corroborates the strategic shading hypothesis.

Finally, while users behave strategically given their beliefs, I find that their beliefs re-

garding rejection concerns during the game are not fully rational. The results can be explained by cognitive hierarchy model of games, which argues that users reason in steps [12]. Specifically, the findings are consistent with Step 1 bounded-rationality. Step-1 users believe that others are Step-0 users, who will naively reveal their preferences without taking rejection concerns into account. Thus, the best response for Step-1 users is to reduce their own preference-ranking for popular users. These findings are in line with the literature in behavioral economics and bounded rationality [56, 75].

In sum, this essay makes three key contributions to the literature. First, this study documents negative returns to popularity information in online platforms. Past empirical research has mainly documented positive returns to the revelation of popularity information. Second, this study is the first to provide empirical evidence for strategic shading in dating markets and directly link it to rejection concerns. While strategic shading has been discussed in the literature, none of the earlier papers have been able to causally identify it. Third, this study demonstrates that users exhibit bounded rationality in real-world online settings. The previous literature on bounded rationality come mainly from lab experiments; the results suggest that such behavioral effects may indeed play a significant role in platforms that involve strategic multi-player interactions.

These results have implications for the design of online dating platforms. On the one hand, displaying popularity information can simplify users' search process and help them quickly evaluate potential partners. However, doing so can have unintended consequences on the demand for popular users. These findings thus suggest that managers of online dating platforms should take this dampening effect of popularity information into account when designing their user-interface. More broadly, these findings are relevant to other two-sided matching markets with inter-personal rejection concerns, e.g., online labor markets.

The details of this essay is organized as follows. In Chapter §2.2, I discuss the related literature. I introduce the setting and data in §4.1 and §4.2. In §4.3, I present descriptive analyses on the effect of popularity information on users' demand. Next, in §refsec:effect-star and §4.5, I present the empirical specification, estimation and identification approaches, and

establish the causal impact of star-ratings on preference-rankings and messages, respectively. In §4.6, I provide a discussion on the mechanisms driving users' behavior and examine the bounds of rationality observed in the data. In §5.2, I review the main findings of this essay and avenues for future research.

Chapter 2
LITERATURE REVIEW

2.1 Literature: A Gamified Online Weight-loss Community

2.1.1 eHealth Weight-Loss Interventions

The first essay of my thesis relates to a long-standing literature studying the effect of eHealth interventions on weight-loss. The eHealth interventions employ technologies such as websites, apps, emails, text messages, and digital games, using devices such as PC, personal digital assistants (PDA), tablets, mobile/smartphones, and smart wears [59, 36, 76, 57]. In the past decade, several studies have explored the effectiveness of eHealth weight-loss interventions via randomized trials, where the treatment group were given access to goal-setting tools, self-monitoring tools (e.g. reporting weight, food intake, and exercise), forums for social support, online educational materials, virtual lifestyle coach providing meal or exercise plans, and reminder emails or text messages. In these randomized trials, the control groups were given either no treatment, or paper-based educational material, self-monitoring diaries, telephone-based feedback, individual or group face-to-face meetings. A systematic review with meta-analysis evaluates the papers from 1995 to 2014 studying the effectiveness of eHealth weight-loss interventions, and demonstrate significant greater weight-loss in eHealth interventions [41]. However, it could not determine which features of eHealth weight-loss interventions are effective, because a combination of different features was provided concurrently in those web- and app-based interventions.

In this chapter, I focus on measuring the effectiveness of weight-loss challenges, a popular but understudied feature of online weight-loss communities. Weight-loss challenges let users set short-term goals, and incentivize them to pursue their goals with gamification elements such as cash prize, badges, and leaderboards. The effectiveness of weight-loss challenges

stems from two main components: goal-setting strategy and gamification mechanism. Next, I review the literature on each of these components.

2.1.2 Goal-Setting

Weight-loss challenges provide a tool for users to set short-term weight-loss goals. There are two types of weight-loss goals: 1) performance (or outcome) goal, referring to a numeric weight-loss target e.g. to lose 10 Kg, 2) process (or learning) goal, referring to a set of dietary or physical activity instructions one acquires to accomplish the performance goal, e.g. walk 10,000 steps a day, or restrict daily caloric intake by 1,250 calories. Although, goal-setting has been a part of many weight-loss interventions, specific strategies for setting goals rarely have been the focus of the intervention research [77, 62]. In contrast, in educational and organizational settings, careful analyses and meta-analyses on goal-setting have been conducted.

In educational settings, some studies have shown that the performance goals have a deleterious effect on a wide range of educationally relevant outcome measures [27]. However, some other studies found that for college students, performance goals improved grades but did not affect interest, whereas process goals enhanced interest in the class but did not affect grades [38]. In organizational settings, it is shown that when a specific difficult process goal was set for a complex task, consistent with goal-setting theory, specific difficult goals led to significantly higher performance than did the general goal of urging people do their best [86]. Thus when tasks are complex for people, process goals can be superior to performance goals. However, there have been almost no studies examining the use of both goals together [53]. The outcomes studied in educational and organizational settings may involve cognitive and behavioral elements similar to health-related outcomes; however, it is not clear how the findings can be applied in a weight-loss context.

In the weight-loss setting, one may believe that process weight-loss goals are more effective than outcome goals, because process goals focus on behavior change and are more under individuals' control, but performance goals are physiological outcomes, which are sub-

ject to many other factors such as genetic predisposition and stress [77, 33]. However, the effectiveness of these two types of weight-loss goals was never studied together in gamified weight-loss interventions. Incorporating a performance goal, i.e. a numeric weight-loss target in a gamified challenge can have both positive and negative effects on weight-loss progress. On the one hand, since gamified challenges incentivize users based on the magnitude of their weight-loss outcome, a numeric weight-loss target can motivate users to stick to their goals. On the other hand, a numeric weight-loss target may draw users' attention to the outcome goal and impair their focus on healthy lifestyle changes i.e. process goals, and discourage users at times of slow weight-loss progress. In this chapter, I investigate the effectiveness of both outcome and process weight-loss goals, and I provide the evidence that both types of goals are required in an optimal gamified weight-loss challenge.

2.1.3 Gamification

Beyond providing a goal-setting tool, weight-loss challenges can enhance users' motivation via gamification tools, based on their weight-loss progress. Gamification has been defined as "the use of game design elements in non-game contexts" [24]. Gamification elements such as leaderboards, points, badges, and cash prize can induce individuals' extrinsic incentives and thus foster desired behaviors [52]. For example, using a leaderboard in a weight-loss challenge, which ranks challenge participants based on their weight-loss progress, enables users to compare their progress with others and keep them motivated during the challenge period.

This chapter relates to prior literature studying the effect of gamification in eHealth interventions. These studies have shown positive effects of gamification on health behavior change and user engagement. For example, in a mHealth app for the self-management of adolescent type 1 diabetes, it has been shown that awarding users with points for blood glucose testing results in an increase of 50% in daily frequency of blood glucose measurement [10]. In another study, it is shown that incorporation of gamification elements such as points, badges, medals, and leaderboard in a rheumatoid arthritis website to reward users'

interaction with different features of the website not only increase their engagement with the website, but also increase their physical activity, and decreased their medication overuse and health care utilization [1].

Among the research papers studying the effect of gamification in eHealth interventions, the online weight-loss challenges are most similar to online physical activity contests. For example, it is shown that walking competitions using mobile apps increase physical activity when users earn badges based on their performance, or when they view each other's step data and make comparisons via leaderboards [17, 72]. However, there is a main difference between physical activity contests and weight-loss challenges. In physical activity contests, participants get incentivized based on the physical activity measure (e.g. get ranked via leaderboards based on number of steps). However, in weight-loss challenges, participants get incentivized based on their weight-loss outcome, i.e. how many kilograms they have lost. As discussed earlier in §2.1.2, number of steps is a process goal, which individuals may have a higher control on; in contrast, the weight-loss outcome is a performance goal, and users may not be able to directly control it. Unlike physical activity, weight-loss is a complicated process. Weight-loss is determined not only by individuals' motivation, but also by a variety of other factors such as genetic predisposition, stress, and environmental factors such as family and friends [33]. Therefore, based on prior literature, it is not clear whether participation in gamified weight-loss challenges can affect users' weight-loss progress.

Focusing on the effect of gamification on weight-loss interventions, prior studies have found that monetary incentives can encourage weight-loss [16, 84]. Moreover, early in the development of weight-loss programs, competitions were proposed as a mean to motivate people in work sites [9, 78]. For example, it is shown that the effectiveness of weight-loss competitions held between employees at two worksites with competition incentives such as a large board of 4 by 5 feet showing participants' weekly progress and a cash prize [9]. Beyond offline settings, surprisingly little is known about the effectiveness of weight-loss competitions in online settings, where competitors do not know each other. Moreover, note that there are no financial incentives available in this setting. Thus, one of the goals of this chapter is to

estimate the effect of participation in online weight-loss challenges on weight-loss progress, especially in the large online weight-loss communities where users do not know each other, and when financial incentives are not available.

2.2 Literature: A Gamified Online Dating Platform

This essay contributes to two broad streams of literature in marketing and economics.

First, it contributes to a large stream of literature that seeks to measure the effect of online popularity information on demand in e-commerce settings and online marketplaces. This research has consistently established the herding effect, i.e. shown that popularity information has a positive effect on demand/sales of products and services in a variety of contexts such as the music industry [71, 25], books [74], restaurants [11], software downloads [26], kidney transplant market [94], movies [55], digital cameras on Amazon [18], and the wedding services market [81].[1] These studies have identified three underlying mechanisms for this positive effect: (1) observational learning or quality inference based on others' actions (e.g. purchase statistics), (2) salience effect or awareness of alternative choices, and (3) network effect or increase in value of a product/service as its user base expands. In this essay, I provide the negative result on the effect of popularity information, and in a previously unstudied context – dating markets. I also present evidence for a new mechanism that can moderate the effect of popularity information – strategic shading due to rejection concerns.

Second, this essay relates to the literature on the empirical measurement of mate preferences in marriage and dating markets. Early work in this stream mostly used data on observed marriages to estimate population-level mate preferences under the assumption of no search frictions [87, 20]. More recently, researchers have been able to access data from speed-dating and online dating platforms. In these settings, search frictions are minimal

[1] A related stream of work examines the effect of WOM or online ratings on demand outcomes [19, 79, 90, 92]. However, in these papers, the ratings are given after the interactions between the buyer and seller. Hence, they play the role of Word-of-Mouth or reputation effects, i.e., they help resolve asymmetric information on the quality of the product/seller. In contrast, in this essay, ratings are purely measures of popularity and do not convey any information on the unobserved quality of the user.

and researchers have direct visibility into the search process employed by users and their preferences. This has led to a stream of literature that attempts to directly estimate users' preferences for mates along a variety of dimensions, e.g., age, income, race, physical attractiveness [48, 30, 31, 28, 39, 5, 51].

An important concern when measuring user preferences is the possibility of strategic behavior – users may shade down their revealed preference for appealing users (physically attractive, popular, etc.) to avoid the psychological cost of rejection [13]. If users shade their revealed preferences, and researchers do not explicitly account for this in their estimation, then the estimates of user preferences will be biased. The effect of users' beliefs on match probabilities on their revealed preference has been examined by a few papers in the literature. In one empirical study, it is shown that shading is not a concern in their setting [39]. Nevertheless, their tests rely on aggregate data patterns and exclusion restrictions. As such, their results may not hold if we had variables that directly affect the perception of match-probability (e.g., popularity information) without affecting the attractiveness of a user at the individual-level. More recently, it is shown that individuals become more strategically selective when they believe they have more potential matches, and less selective when they believe they have more competition [93]. However, neither of these papers examine how revelation of popularity information affects users' demand in a dating platform, and connect it to strategic shading based on the differences in perceived rejection probabilities.

Chapter 3

A GAMIFIED ONLINE WEIGHT-LOSS COMMUNITY

3.1 Setting

In this chapter, I focus on one of the largest non-commercial online weight-loss communities launched in 2006. The platform has over 45 million users, who either want to lose or maintain their weight. The platform has international websites in more than 30 countries. This data comes from the platforms website in the Unites States. With an emphasize on user-generated content, the platform let users track their weight, nutrition, and physical activity and engage with a supportive community through its free website and mobile app.

To join and use this online weight-loss community, users should create a profile. They are required to provide the information about their current and goal weight on their profile. Users' personal goal weight on their profile is usually a fixed long-term goal. However, it can be updated anytime. Users in their personal profile can journal about their calorie intake, report exercise activities, and write a narrative about their daily weight-loss journey and post pictures. There are "general forums" on this online weight-loss community, where users can ask/provide answers to general questions such as "how often do you weigh yourself?". Further, users with common interests or concerns can create/join "group" forums, share their experiences with weight-management, and exchange diet or exercise tips. An example of a group forum is a group for "low carb lovers".

In mid-2008, weight-loss "challenges" were introduced as a new feature on the platform. This feature enables users to set a "short-term" weight-loss goal. Any user can create her/his own challenge, or join the challenges created by the other users on the platform. The creator of the challenge sets the objective of the challenge by choosing a name and defining a set of instructions. Table 3.1 shows some examples of the challenge names and instructions on this

platform. The name and the instructions of a challenge may include a numeric weight-loss target, such as "losing 5 Kg", or it can be specified by some weight-loss instructions such as "exercising 30 minutes a day", or "quitting soda". The numeric weight-loss target is a performance goal, and the weight-loss instructions are the process goals, which can help one achieve the performance goal. As shown in Table 3.1, there are some challenges without a weight-loss target. Further, the creator of the challenge sets a fixed start date and duration for the challenge, and other users can join the challenge only before its starting time. Users can join any number of challenges.

The weight-loss challenges are integrated with leaderboards. Every time a challenge participant reports his weight, he will be moved to a higher or lower rank on the leaderboard in comparison with the weight-loss progress of the other challenge participants. Weight-loss progress is normalized by users' initial weight at the beginning of the challenge. Users can see the ordered name of the challenge participants and their weight-loss progress on the leaderboard. The platform utilizes leaderboards as a gamification object to incentivize and engage challenge participants. Other gamification objects such as status points, badges, or financial prizes are not used on this platform. Further, each challenge is integrated with a challenge forum, where challenge participants can interact with each other, ask questions and provide answers, and encourage each other. Finally, users can join any number of challenges available on the platform.

The platform's mobile app enables users to log in their weight, count their intake calories, and upload pictures of their food. The general and group forums and challenges are not accessible via the mobile app.

3.2 Data

3.2.1 User Level Data

Online weight-loss challenges were introduced on the platform in August 2008. Therefore, I randomly chose 4,208 users who had joined the platform at least four months before the

Challenge Name	Instructions
Losing 5 Kg	1. Try to go without sugar
	2. Zero calorie sweetener only
	3. Exercise 30 minutes a day
Avoid the major fast food chains	1. Order a salad instead of a sandwich in a fast food restaurant
	2. Substitute the major unhealthy chains with a relatively healthy chain
Walk-walk-walk	1. No elevators, use the stairs
	2. Park further away and walk

Table 3.1: Examples of the challenge names and instructions.

introduction of challenges, and tracked their online activity for eight months from April 2008 to the end of November 2008. Users can choose to report their weight any number of times on the platform. In the data, I have access to the average weight that each user reports during a month, and I could calculate the monthly weight-loss progress for only 1,045 out of 4,208 users who have reported their weight in at least two consecutive months during the study period. I define $weight_{it}$ as the average weight that user i reports during month t (in kilograms), and I define the main dependent variable as $weightLoss_{it}$, which equals $weight_{it-1} - weight_{it}$. As summarized in Table 3.2, there are 4,719 observations for $weight_{it}$ and 3,159 observations for $weightLoss_{it}$, which shows that panel data set is unbalanced, i.e. the weight-loss progress is missing for some users at some periods. Further, the statistics show that the median user is around 80.75 kg and has lost 0.7 kg per month. Figure 3.1 shows the distribution of $weightLoss_{it}$ is close to normal, and users report weight-loss as well as weight-gain.

3.2.2 Challenge Level Data

In this data, users participated in 96 different challenges. I can observe the challenge creator, start date, duration, number of participants, number of instructions and whether the challenge has a weight-loss target. However, I do not observe the rankings of the participants

Variable	Mean	Std. Dev	50th	(Min, Max)	Sample size
$weight_{it}$	84.452	21.296	80.75	(41.1, 199.6)	4719
$weightLoss_{it}$	0.801	1.798	0.700	(-7, 16.425)	3159
$challenge_{it}$	0.062	0.24	1 0	(0, 1)	8360
$numReport_{it}$	1.637	2.781	1	(0, 31)	8360
$journal_{it}$	0.578	2.292	0	(0, 32)	8360
$generalForum_{it}$	0.357	2.704	0	(0, 75)	8360
$groupForum_{it}$	0.043	0.631	0	(0, 24)	8360
$challengeForum_{it}$	0.084	1.309	0	(0, 93)	8360
$tenure_{i1}$	5.041	4.254	4	(0, 19)	1045
$goal_i$	68.587	12.595	66.7	(43, 124.7)	1045

Table 3.2: The summary statistics of the user level data for 1,045 users.

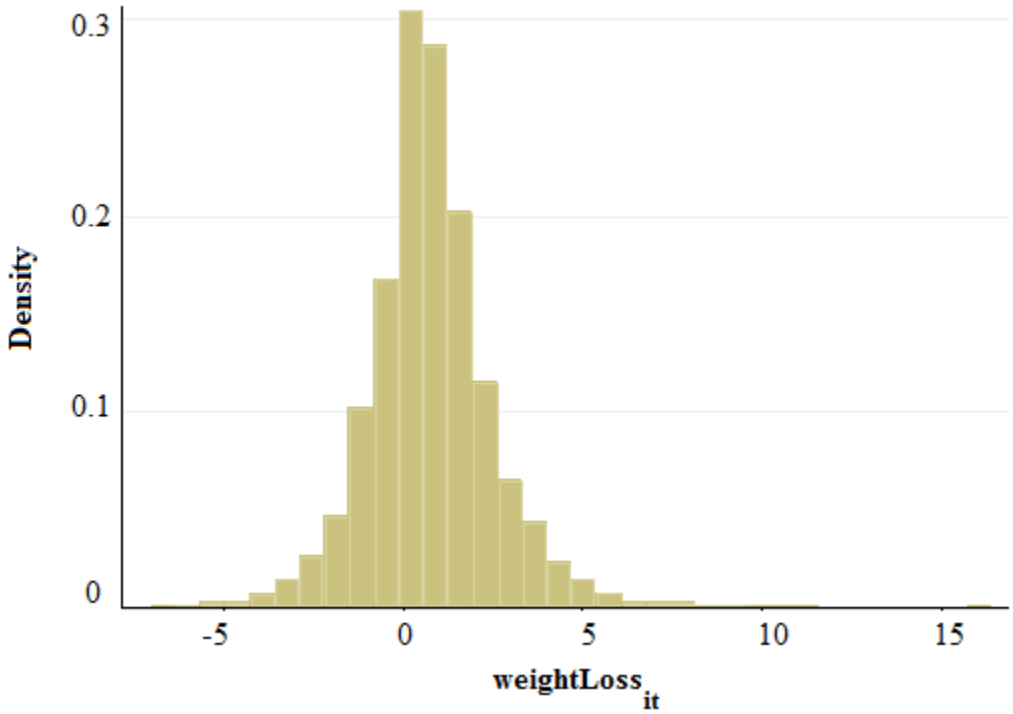

Figure 3.1: The average monthly weight-loss distribution.

Variable	Mean	Std. Dev	50th	(Min, Max)	Sample size
$duration_c$	43.05	23.36	42	(7, 84)	96
$member_c$	30.51	29.88	24	(1, 169)	96
$instruction_c$	2.50	2.51	2	(0, 13)	96

Table 3.3: The summary statistics of the challenge level data for 96 challenges.

on the challenge leaderboard. I denote the duration of a challenge (days) by $duration_c$. As summarized in Table 3.3, the duration of a median challenge is 42 days.[1]

Further, I denote the number of participants in a challenge by $member_c$. As summarized in Table 3.3, the number of challenge participants varies between one to 169 members, with a median challenge having 24 members. Based on the median challenge, I split challenges into two types: challenges with less than or equal to 24 members ($lowMember_c$), and challenges with more than 24 members ($highMember_c$).[2]

Regarding the objective of a challenge, I denote the number of instructions of a challenge by $instruction_c$. As shown in Table 3.3, the median challenge has two instructions. Based on the median challenge, I split challenges into two types: challenges with less than or equal to two instructions ($lowInstruction_c$), and challenges with more than two instructions ($highInstruction_c$).

I also observe whether a weight target (e.g. losing 5 Kg) is specified in a challenge. There are 15 challenges with a weight target, which I denote them by $target_c$, and there are 81 challenges without a target, which I denote them by $nonTarget_c$.

[1]The challenge duration is indirectly captured by $challenge_{it}$ which shows whether user i is participating in a challenge at month t, either new challenges started in that month, or ongoing challenges started from previous months but not finished at month t. If a challenge is finished within the first week of month t, I do not consider it in $challenge_it$, but I consider it in $challenge_{it-1}$. Similarly, if a challenge is started within the last week of month t, I do not consider it in $challenge_{it}$, but I account for it in $challenge_{it+1}$. In this data, that there is no challenge with a starting date on the last week of month t and a finishing date on the first week of the next month.

[2]Note that users can participate in both type of low- and high-member challenges simultaneously.

3.3 Empirical Methodology

3.3.1 System GMM (Blundell-Bond) Estimator

To shed light on the effect of participation in challenges on the user's weight-loss, I employ a dynamic model in which the dependent variable, user's weight at any given time ($weight_{it}$), is modeled as a linear function of:

$$weight_{it} = c + \alpha weight_{it-1} + \beta challenge_{it} + X_{it}\Phi + \gamma z_i + m_t + \eta_i + e_{it} \quad t = 2, , 8 \quad (3.1)$$

where,

- $challenge_{it}$ is the main variable of interest, a binary variable indicating whether user i is participating in any challenge at month t or not.
- X_{it} is a vector of time-varying variables capturing the user's engagement on the platform during month t, including:
  - $numReport_{it}$ - the log of number of times that user i has reported her weight during month t.
  - $journal_{it}$ - the log of number of posts that user i has written in his journal during month t.
  - $groupForum_{it}$ - the log of number of posts that user i has written in group forums during month t.
  - $generalForum_{it}$ -the log of number of posts that user i has written in general forums during month t.
  - $challengeForum_{it}$ - the log of number of times that user i has written in challenge forums during month t.
  - $tenure_{it}$ - the number of months since user i first joined the online weight-loss community at time t.
  - $tenure_{it}^2$ - the squared term of $tenure_{it}$.
- z_i is is user i's fixed weight goal.
- m_t is a month dummy.

- η_i is user-specific fixed effect.
- e_{it} is a mean-zero error term.

In estimation of equation (3.1), several econometric problems may arise:

1. the main variable of interest, $challenge_{it}$, and other explanatory variables are potentially correlated with individual's unobserved time-invariant characteristics (η_i), such as age, gender, height.

2. the lagged dependent variable ($weight_{it-1}$), although not correlated with the current error term (e_{it}), is a pre-determined variable and correlated with previous shocks.[3]

3. the main variable of interest, $challenge_{it}$, and the explanatory variables capturing the users' activity on the platform, X_{it}, are potentially correlated with time-varying shocks (e_{it}). For example, a random shock at user's motivation level may result in user's self-selecting herself into participating in a challenge.

Problem (1) results in a biased OLS estimate.[4] An initial remedy to problem (1) is to use the fixed-effects model. However, due to problem (2), and because I have a short panel (T=8) with many users (N=253), the fixed-effects model will be biased [61].[5] To resolve this issue, I can use a "system GMM" estimation employing valid instrumental variables (IVs) within the model [7]. The system GMM starts by using transformed regressors as valid instruments in the level equation (3.1) e.g. $\Delta weight_{it-1}$ (or deeper lags) as an IV for $weight_{it-1}$, and

[3]I also consider $tenure_{it}$ to be a predetermined variable, i.e. independent of e_{it} but correlated with e_{i1}. Note that $tenure_{it}$ can be written as $tenure_{it} = tenure_{i1} + (t-1)$; thus, $tenure_{it}$ is a function of $tenure_{i1}$ and a deterministic time increment t. Since, most users join the platform when they have higher levels of motivation, $tenure_{i1}$ can be correlated with e_{i1}. Hence, $tenure_{it}$ can be correlated with e_{i1}.

[4]In an OLS model, the estimate of α is biased upward (downward) if I assume positive (negative) correlation between $weight_{it-1}$ and η_i. Other OLS coefficient estimates are also biased [80].

[5]The within estimator is inconsistent because the mean-differencing makes the $y_{it-1} - \bar{y}_i$ correlated with the error $e_{it} - \bar{e}_i$, because y_{it-1} is correlated with e_{it-1} and hence with $\bar{e}_i$. Because of the negative correlation of y_{it-1} with the error term $e_{it} - \bar{e}_i$, the estimation of in fixed effect model will be biased downward.

$\Delta challenge_{it-1}$ (or deeper lags) as an IV for $challenge_{it}$.[6] In addition to level equation 3.1, the system GMM uses a difference equation by transforming all regressors, usually by "first-differences" to eliminate fixed effects (η_i) as follows:[7]

$$\Delta weight_{it} = \alpha \Delta weight_{it-1} + \beta \Delta challenge_{it} + \Delta X_{it}\Phi + \Delta m_t + \Delta e_{it} \quad t = 2, , 8 \qquad (3.2)$$

In the differenced equation (3.2), I can use the lags of regressor as valid instrument, i.e. I can use lag 1 and deeper for predetermined variables (e.g. $weight_{it-2}$ as an IV for $\Delta weight_{it-1}$), and lag 2 or deeper for endogenous variables (e.g. $challenge_{it-2}$ as an IV for $\Delta challenge_{it}$). Finally, using equation (3.1) and (3.2), I can apply the Generalized Method of Moments (GMM) to calculate a consistent and efficient estimator. Thus, using valid IVs, I can mitigate both endogeneity concerns in problem (2) and (3).

Utilizing instruments within the model for the GMM estimator is useful when exogenous IVs are not available. However, compared with the traditional instrumental variable approach, we should proceed with caution when making a causal interpretation. The validity of the system GMM estimates relies on the validity of its IVs. For example, in the level equation, $\Delta challenge_{it-1}$ is a valid IV for $challenge_{it}$, if we assume that it is correlated with $challenge_{it}$, but not correlated with η_i and e_{it}; or in the difference equation, $challenge_{it-2}$ is a valid IV for $\Delta challenge_{it}$, if we assume that it is correlated with $\Delta challenge_{it}$, but not correlated with Δe_{it}. I use Hansen test to examine the validity of the group of IVs. Moreover, I assume that e_{it}s are iid across i and across t (no serial correlation), i.e. $e_{it} \sim iid(0, {}^{2}_{i})$ $\forall i, t$. The validity of assuming no serial correlation is tested and verified by using Arellano-Bond test [3].[8] Also, e_{it}s can be heteroskedastic across individuals (σ_i^2). I can ensure the

[6] I consider time dummies and challenge fixed effects as strictly exogenous variables in this setting, and I use them as their own IVs in the level equations.

[7] "First-differences" subtracts the previous observation from the contemporaneous one. Note that in equation (3.2), "forward orthogonal deviations" is used for month dummies. Forward orthogonal deviation subtracts the average of all future "available" observations of a variable [4]. In this analysis, in order to minimize data loss, I employ the orthogonal deviations for all variables.

[8] This test has a null hypothesis of no autocorrelation and is applied to differenced residuals. AR(1) tests the autocorrelation between Δe_{it} and Δe_{it-1}. Usually the test for AR(1) rejects the null because Δe_{it} and

robustness of the results to this heteroscedasticity using a two-step system GMM. The full set of assumptions of system GMM is summarized in Appendix A.2.

In this analysis, it is easier to interpret the dependent variable $weightLoss_{it}$ instead of $weight_{it}$, which is defined as:

$$weightLoss_{it-1} = weight_{it-1} - weight_{it}. \tag{3.3}$$

Therefore, I can easily transform Equation (3.1) to:

$$weightLoss_{it} = -c + (1-\alpha)weight_{it-1} - \beta challenge_{it} - X_{it}\Phi - \gamma z_i - m_t - \eta_i - e_{it}. \tag{3.4}$$

This transformation does not affect the use of System GMM approach explained above. To run the system GMM model, I use xtabond2 a Stata command written by [65]. With this command, I use a two-step option to make analysis robust to heteroscedasticity. Further, I use the robust option to apply the Windmeijer finite-sample correction to fix the downward bias of the system GMM standard errors [85].

3.3.2 Inverse Probability Weighting (IPW)

In this sample, there are 792 (near 76%) of the users who never participated in any challenge during the data collection period (non-adopters). Note that $challenge_{it}$ equals zero for all non-adopters at all time periods. Thus, when I estimate the effect of participation in challenges, the non-adopters data gets dropped from the system GMM analysis. Using only the adopters' data can result in a biased estimation; because an important kind of nonrandom selection, called incidental sample truncation, arises when certain individuals do not appear in a random sample due to individual choices or behaviors [88]. An approach to consistent estimation in the presence of incidental sample truncation is based on inverse probability weighting (IPW).[9] In the IPW approach, by considering ignorability assump-

Δe_{it-1} have e_{it-1} in common. AR(2) tests the autocorrelation between Δe_{it} and Δe_{it-2}. AR(2) p-value test is more important than AR(1) because rejected H0 in AR(2) reveals serial correlation between errors and indicates that the validity of IVs is violated.

[9]Unlike Heckman's approach, the IPW approach does not require identifying exogenous variables to satisfy the exclusion restrictions.

tion, we assume that conditioned on the observed variables, no unobserved variable exists that can affect both challenge adoption and weight-loss outcome, i.e. adoption is exogenous. Although this assumption is not directly testable, I discuss about a way to assess it indirectly in §3.5.4. In the IPW approach, I use observed variables that are not affected by the introduction of challenges (pre-treatment variables), including user's goal, tenure on the platform, and the average engagement level with other features of the platform during the four months before the introduction of challenges, including the average frequency of using the platform to: report weight ($avgNumReoprt_i$), write a journal ($avgJournal_i$), or post a comment or respond to a comment on general forums ($avgGeneralForum_i$), and group forums ($avgGroupForum_i$).

As illustrated in Table 3.4, I compare adopters and non-adopters' (pre-treatment) observed variables. Significant differences between adopters and non-adopters in their online activity and tenure on the platform show that active and high tenured users on the platform are more likely to adopt challenges. In order to employ IPW, I use a probit model to calculate the probability of adoption based on users' observed variables prior to the introduction of challenges. The results of this probit model are summarized in Appendix A.3. Next, I use the inverse of the probability of adoption to weight the observations. Thus, this approach gives less (more) weights to adopters who are likely (less likely) to adopt the challenges, by using the inverse probability of adoption as observation weights. IPW ensures that the sample of adopters used in the model and non-adopters dropped from the model are similar (see Table 3.4). Next, I incorporate these weights in the system GMM model utilizing the Stata xtabond2 command. [10]

[10]These weights are incorporated in the system GMM model by using the equation below, where for N observations, Y is the outcome matrix, and X is the matrix of regressors, and Z is the matrix of instruments, and W is the matrix holding the weights [65].

$$\hat{\beta} = (X'WZAZ'WX)^{-1}X'ZAZ'WY$$

	Mean			Mean (weighted sample)		
	adopter	non-adopter	$p > \|t\|$	adopter	non-adopter	$p > \|t\|$
Variables	(n=253)	(n=792)		(n=253)	(n=792)	
$avgNumReoprt_i$	0.883	0.755	0.001	0.786	0.784	0.954
$avgJournal_i$	0.431	0.176	0.000	0.246	0.243	0.948
$avgGeneralForum_i$	0.287	0.099	0.000	0.145	0.143	0.937
$avgGroupForum_i$	0.016	0.001	0.000	0.005	0.003	0.539
$tenure_{i1}$	5.549	4.879	0.029	5.182	5.101	0.782
$goal_i$	68.399	68.648	0.785	68.760	69.683	0.934

Table 3.4: The comparison of adopters and non-adopters before the introduction of challenges.

3.4 Results

The results from this estimation exercise are presented in Table 3.5. I start with an OLS model, and then modify it step by step to address all concerns discussed in §3.3.1, ending with the estimator of interest. The weighted observations are used in all these models. Applying OLS in model M1, and a fixed effect estimation in model M2, I find a positive and significant effect of $challenge_{it}$, suggesting that participation in challenges has a positive and significant effect on weight-loss. However, as explained in §3.3.1, both OLS and fixed effects models are biased due to Nickell bias and endogeneity concerns. The popular solution is to apply difference GMM, which starts by transforming the data by first differencing, and then instrumenting differenced regressors with their lags [3].[11]

Applying difference GMM in model M3, I find that the coefficient of $challenge_{it}$ remains positive and significant. Although, difference GMM is shown to be suitable for "small T, large N" panels, in dynamic panel models where the autoregressive parameter (α) is moderately large and the number of time series observations is moderately small, the difference GMM

[11]Lag 2 or deeper for endogenous variables (e.g. $challenge_{it-2}$ for $\Delta challenge_{it}$, and lag 1 and deeper for predetermined variables (e.g. $weight_{it-2}$ for $\Delta weight_{it-1}$). Here, I use only one lag to limit the number of instruments.

IVs (past levels of the regressors) convey little information about the transformed regressors (future changes) and therefore the weak instruments make the difference GMM estimator perform poorly [7]. The poor performance of the difference GMM in this setting is reflected in the estimated coefficient of $weight_{it-1}$ $(1-\alpha)$, which equals 0.434 in model M3. The correct estimated coefficient of $weight_{it-1}$ is expected to fall within the range of OLS and fixed effects model, i.e. (0.017, 0.339) [65]. It is shown that when α is big and T is small, system GMM estimator performs better than difference GMM [7]. In system GMM, I use level equations in addition to difference equations, and use transformed regressors as instruments in the level equation.

Applying system GMM in model M4, I find that the estimated coefficient of $weight_{it-1}$ falls within the range of OLS and fixed effects model. This provides an evidence showing the good performance of the system GMM model. Also, the Hansen test p-value (0.268) confirms that all the instruments as a group are exogenous and valid.[12] Also, the autocorrelation AR(2) p-value (0.243) supports the assumption of no serial correlation.

In model M4, the estimated effect of $challenge_{it}$ is positive and significant. I can interpret it as users can achieve a weight-loss of 0.945 kg a month, by participating in at least one challenge. Note that the estimated coefficient of $challenge_{it}$ shows the short-term effect of participation in challenges on $weightLoss_{it}$. Since, I include $weight_{it-1}$ in this dynamic model, I can calculate the long-term effect of participation in challenges as well. Based on equation (3.1), while the effect of $challenge_{it}$ on $weight_{it}$ is β (i.e. -0.945), its effect on $weight_{it+1}$ will be $\alpha\beta$ (i.e. -0.808), and on $weight_{it+2}$ will be $\alpha^2\beta$ (i.e. -0.691), and so on.[13] For example, as shown in Figure 3.2, if I consider a user who is around 75.945 kg, if he participates in a challenge at time $t = 5$, his weight will be approximately 75 kg in the next month. However, as shown in Figure 3.2, if he does not participate in another challenge in future, he will gain his lost weight back over the next few months. The long-term effect of

[12]Hansen test is used instead of the Sargan test due to robust standard errors.

[13]Similarly, based on equation 3.4, while the short-term effect of $challenge_{it}$ on $weightLoss_{it}$ is $-\beta$ (i.e. 0.945 kg), the effect of $challenge_{it}$ on $weightLoss_{it+1}$ is $-(1-\alpha)\beta$ (i.e. 0.137 kg), and on $weightLoss_{it+2}$ is $-\alpha(1-\alpha)\beta$ (i.e. 0.117 kg), and on $weightLoss_{it+3}$ is $\alpha^2(1-\alpha)\beta$ (i.e. 0.100 kg), and so on.

	(M1) OLS	(M2) FE	(M3) Diff-GMM	(M4) System-GMM
$challenge_{it}$	0.52906***	0.33649**	0.83332**	0.94558**
	(0.14747)	(0.15416)	(0.32586)	(0.36706)
$weight_{it-1}$	0.01727***	0.33860***	0.43412**	0.14544***
	(0.00312)	(0.03185)	(0.18282)	(0.04317)
Month dummies	✓	✓	✓	✓
Constant	-1.10952***	-27.90783***	-	-11.26733***
	(0.29719)	(2.70741)		(3.64033)
Observations	984	984	731	984
Individuals	253	253	197	253
IVs	-	-	14	24
AR(2) pvalue	-	-	0.303	0.243
Hansen pvalue	-	-	0.486	0.268
R-Squared	0.06066	0.30591	-	-

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Reported coefficient of $weight_{it-1}$ is $1 - \alpha$.

Table 3.5: The estimated effect of participation in challenges on weight-loss.

challenges on weight emphasizes the fact that online weight-loss communities should have strategies to help users maintain their weight-loss goals over time.

Next, in model M5, Table 3.6, I add individual control variables including user's weight goal, different activities on the platform, and tenure on the platform. One of the advantages of the system GMM is that we can include time-invariant regressors in the model, such as $goal_i$, which would disappear in difference GMM. It is important to note that adding control variables requires more instruments, and too many instruments in system GMM models can result in overfitting the endogenous variables [66]. One solution to limit number of

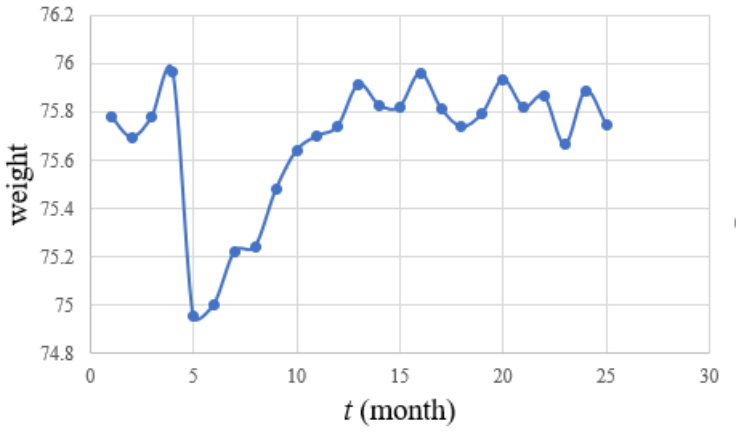

Figure 3.2: The short-term and long-term effect of participation in challenges.

instruments is to collapse instrument sets by creating one instrument for each variable and lag distance, rather than one for each time period, variable, and lag distance. This approach in small samples can avoid the bias due to the rising number of instruments. In model M5, Table 3.6, I reduce the number of instruments from 94 to 25 by collapsing them. As shown in model M5, I find that the effect of $challenge_{it}$ remains positive and significant. However, there is a severe multicollinearity between $weight_{it-1}$ and $goal_i$. Therefore, in model M6, I exclude $goal_i$. Removing this time-invariant variable does not change the results, and it is not expected to do so, because this model controls for the fixed heterogeneities very well.

In this setting, there are 96 different challenges, with different instructions, weight-targets, and number of participants. Participation in all challenges may not have the same effect. Participants in some challenges may benefit from well-designed instructions, weight-targets or within member dynamics. Further, individuals who participate in similar challenges may have similar motivation, i.e. the errors could be correlated across individuals. Including challenge dummies removes any challenge-related shocks from the error and can control correlations across individuals who chose to participate in a specific challenge. However, introducing 96 different challenge dummies in the system GMM model results in too many instruments, and makes the findings less reliable [66]. Therefore, as an alternative, I group

	(M5)	(M6)
$challenge_{it}$	1.28487**	1.18302**
	(0.51923)	(0.51670)
$weight_{it-1}$	0.12157**	0.09878***
	(0.05205)	(0.03592)
$goal_i$	-0.12982**	-
	(0.05644)	
$numReport_{it}$	2.25104**	2.29390**
	(1.12226)	(1.01705)
$journal_{it}$	-1.21830*	-1.13674*
	(0.69649)	(0.60648)
$generalForum_{it}$	0.43603	0.46655
	(0.50370)	(0.52444)
$groupForum_{it}$	0.96474	0.82055
	(0.75794)	(0.82491)
$challengeForum_{it}$	-0.42528	-0.17194
	(0.66155)	(0.84921)
$tenure_{it}$	0.05817	0.02865
	(0.04406)	(0.03957)
$tenure_{it}^2$	-0.00160	-0.00088
	(0.00581)	(0.00515)
Month dummies	✓	✓
Constant	-2.80255**	-10.03100***
	(1.08475)	(2.71564)
Observations	984	984
Individuals	253	253
IVs	25	24
AR(2) pvalue	0.390	0.381
Hansen pvalue	0.219	0.543

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.6: The estimated effect of participation in challenges on weight-loss.

Variables	Definition
d_{HHT}	$I(highMember_c \times highInstruction_c \times target_c)$
d_{HHN}	$I(highMember_c \times highInstruction_c \times nonTarget_c)$
d_{HLT}	$I(highMember_c \times lowInstruction_c \times target_c)$
d_{HLN}	$I(highMember_c \times lowInstruction_c \times nonTarget_c)$
d_{LHT}	$I(lowMember_c \times highInstruction_c \times target_c)$
d_{LHN}	$I(lowMember_c \times highInstruction_c \times nonTarget_c)$
d_{LLT}	$I(lowMember_c \times lowInstruction_c \times target_c)$
d_{LLN}	$I(lowMember_c \times lowInstruction_c \times nonTarget_c)$

Table 3.7: Definition of challenge-category dummies.

the challenges with similar attributes into similar categories and add the interaction of these challenge-category dummies with $challenge_{it}$ to this analysis. As explained earlier in §3.2.2, I divide challenges based on three main characteristics of challenges: number of the challenge participants (high vs. low), number of instructions (high vs. low), and the existence of a weight target (one vs. none). Next, I define eight challenge-category dummies based on these attributes. The denotation of these challenge-categories is shown in Table 3.7.

In model M7, Table 3.8, adding the challenge-categories, I find the effect of $challenge_{it}$ to remain positive and significant. In order to interpret the magnitude of the estimated coefficient of $challenge_{it}$, we need to consider the interaction terms with respect to the category of the challenge. The interesting pattern among the interaction effects shows that keeping the number of participants and instructions the same, participation in challenges without a target has a smaller effect. For example, the interaction effect of participating in challenges with high number of participants, and high number of instructions, with a target ($challenge_{it} \times d_{HHT}$) has an insignificant interaction effect; however, participating in similar challenges with high number of participants, and high number of instructions, without a target ($challenge_{it} \times d_{HHN}$) has a negative and significant interaction effect. I can see the same pattern among the challenges with high number of participants and low instructions.

Similarly, I see this pattern among the challenges with low number of participants and high number of instructions. The only exception is the negative and significant interaction effect of challenges with low number of participants and low number of instructions, regardless of the presence of a target. This exception may look intuitive because challenges with low number of participants and low number of instructions may be poorly defined challenges overall.

In order to interpret the causal effect of the challenge category on participant's weight-loss, we need to proceed with caution. The negative and significant interaction effect of challenges without a target might be due to lower levels of motivation among participants who chose these kinds of challenges. However, as explained in §3.2, one important point to note in this setting is that on average only 25 challenges are created in a month in random days, and on average only four of them have a numeric target. Therefore, not that many challenges are available for users to join at each day. Thus, if a user is motivated to participate in a challenge, s/he does not have a big set of challenges to chose from, and self-selecting into a specific type of challenge versus another does not play a role. Therefore, although self-selection into a challenge can be strong, self-selection into a certain type of challenge versus another type is much weaker.

3.5 Robustness Checks

3.5.1 The Challenge Creators

The challenge creators (or admins) might be more motivated than others to obtain a higher performance in their own challenge. Thus, the effect of challenge participation for challenge creators is less likely to be causal. There are 14 users who created all the 96 challenges in this data. In model M8, Table 3.9, as a robustness check, I removed these challenge creators from this analysis. As shown in model M8, the results remain qualitatively similar to model M6.

	(M7)
$challenge_{it}$	5.71988*
	(3.13685)
$weight_{it-1}$	0.07954**
	(0.03632)
$challenge_{it} \times d_{HHT}$	-3.25304
	(2.03217)
$challenge_{it} \times d_{HHN}$	-3.59215*
	(2.06065)
$challenge_{it} \times d_{HLT}$	-1.05491
	(1.35435)
$challenge_{it} \times d_{HLN}$	-3.14035*
	(1.86277)
$challenge_{it} \times d_{LHT}$	-4.42172
	(3.07459)
$challenge_{it} \times d_{LHN}$	-3.68953**
	(1.74144)
$challenge_{it} \times d_{LLT}$	-1.14363**
	(0.52728)
$challenge_{it} \times d_{LLN}$	-2.94407*
	(1.74724)
Controls	✓
Constant	-10.19023***
	(3.12065)
Observations	984
Individuals	253
IVs	32
AR(2) pvalue	0.433
Hansen pvalue	0.787

Standard errors in parentheses * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.8: The estimated effect of participation in challenges on weight-loss.

Variables	(M8) System-GMM	(M9) System-GMM	(M10) Fixed Effects	(M11) System-GMM	(M12) OLS
$challenge_{it}$	1.348**	1.033***	0.694**		
	(0.593)	(0.394)	(0.273)		
$adopt_i$				-0.614	-0.119
				(0.982)	(0.108)
Challenge FE	-	-	✓	-	-
Controls	✓	✓	✓	✓	✓
Observations	922	535	984	1633	1633
Individuals	239	86	253	841	841
IVs	24	24	-	15	-
AR(2) pvalue	0.264	0.658	-	-	-
Hansen pvalue	0.476	0.833	-	-	-

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Controls include $weight_{it-1}$, $numReport_{it}$, $journal_{it}$, $generalForum_{it}$, $groupForum_{it}$, $challengeForum_{it}$, $tenure_{it}$, $tenure_{it}^2$, the month dummies and the constant term.

Table 3.9: Robustness checks for the estimated effect of participation in challenges on weight-loss.

3.5.2 Self-Reported Weight Outcomes

One potential source of endogeneity rises with the fact that users self-report their weight, and each user can make two decisions: 1) choose to report/hide her weight, and 2) choose to misreport a lower weight. Such decisions if not random and correlated with users' challenge participation, can bias the estimated effect of challenge participation; the first decision can create an unbalanced panel due to non-random selection, and the second decision can create a non-random measurement error. For example, if those who experience a weight-gain during participating in a challenge choose to not report their weight or decide to misreport a weight-loss, the estimated effect of challenge participation will be biased.

To address the first concern about the users' decision to report/hide weight, I consider

a conceptual robustness test by analyzing the sensitivity of the results to subsamples of users who report their weight at least once every month, during the four months after the introduction of challenges, regardless of participation in a challenge or not. If the results were largely driven by decisions to report/hide weight during challenge participation, then these results should disappear if we focus on users who report at a regular frequency. However, as shown in model M9, Table 3.9, the results are robust to such self-report selection.

To address the second concern about the users' decision to misreport a lower weight, I consider two cases: a) If an individual has a motivation to misreport his weight, and his behavior is consistent throughout the time, this error can be captured by his unobserved fixed effect η_i in equation (3.1). Since, this estimation method controls for η_i, the estimation of β will be unbiased; b) If an individual's motivation to misreport varies over time, I can model this behavior by breaking the error e_{it} to Δ_{it} and e'_{it}:

$$weight_{it} = c + \alpha weight_{it-1} + \beta challenge_{it} + X_{it}\Phi + \gamma z_i + m_t + \eta_i + e'_{it} - \Delta_{it} \quad t = 2, , 8 \quad (3.5)$$

where, Δ_{it} captures the amount of weight that the individual reports lower than her true weight. In this case, if an individual misreports regardless of him participating in a challenge, i.e. if Δ_{it} is independent of $challenge_{it}$, then this error does not create a bias in the estimation of β. However, if an individual systematically misreports higher weight-loss when he is participating in a challenge, i.e. if Δ_{it} is correlated with $challenge_{it}$, the effect of participating in a challenge will be biased and over-estimated. However, for two reasons I believe this correlation is unlikely. First, individuals who participate in a challenge may misreport their weight to show that they have achieved the challenge target, and/or to stand in a better ranking position on the challenge leaderboard. With these two incentives, untruthful challenge participants may never report a weight-gain. Although, I cannot directly measure whether the challenge participants are truthfully reporting their weight-amount, I can examine how often they report a weight-gain. Among 253 challenge adopters, 76 users ($\sim 30\%$) have reported a weight-gain at least once, during the challenge participation. The fact that a high number of challenge participants report weight-gain at least once, increases

	Before Introduction of Challenges (Four Months)		After Introduction of Challenges (During Participation)	
	Frequency	Percentage	Frequency	Percentage
Reporting a Weight-gain	116	27.75%	264	26.49%
Reporting a Weight-loss	289	69.14%	98	71.35%
No Reporting	13	3.11%	8	2.16%
Total	418	100	370	100

Table 3.10: The adopters' weight-reporting frequency.

the reliability of their self-reporting behavior.

Second, I compare the frequency with which challenge participants report a weight-gain before and after the introduction of challenges. Here, by frequency I mean the number of monthly observations where the average weight is higher than the average in previous month; thus, I infer that the user reported a weight-gain at least once. As shown in Table 3.10, during the first four months before introduction of challenges, the adopters have reported a weight-gain 27.75% of the times (116 out of 418 weight observations). After the introduction of challenges, I observed that challenge adopters, when participating in a challenge, have reported a weight-gain 26.49% of the times (98 out of 370 weight observations). This shows that the rate with which users report a weight-gain before and after the introduction of challenges has remained the same. Thus, I can infer that users are less likely to misreport during challenge participation. One potential reason might be the absence of any actual prize for the winners of these challenges.

3.5.3 Fixed Effects Model

As explained earlier, including challenge dummies can control correlations across individuals who chose to participate in a specific challenge. Controlling for 96 different challenge dummies in a system GMM model results in too many instruments and less reliable results. However, I can include these many dummies in a fixed effects model. As shown in model

M10, Table 3.9, the effect of $challenge_{it}$ remains positive and significant after controlling for the challenge dummies in a fixed effects model. Thus, I can infer that the correlation across participants in a challenge is not a concern in previous models. However, as explained earlier, the fixed effects model in this setting cannot control for the endogenous and predetermined variables well enough, and the results in model M10 are biased. Thus, using challenge-categories in model M7 is a better solution.

3.5.4 Ignorability Assumption

I use the IPW approach to address the incidental sample truncation bias. The main assumption of the IPW approach is the ignorability assumption. By considering ignorability, I assume that conditioned on the observed variables, no unobserved variable exists that can affect both challenge adoption ($adopt_i$) and weight-loss outcome.[14] For example, in this setting, one major unobserved factor that may affect both challenge adoption and weight-loss outcome is the users' unobserved motivation level.

The ignorability assumption is not directly testable. However, there are ways to assess it indirectly. One test relies on estimating the causal effect of the treatment (i.e. adoption) on a variable known to be unaffected by it, typically because its value is determined prior to the treatment itself. Lagged (pretreatment) outcome variables are best for this test, because they are closely related to the outcome of interest [45]. If the estimated effect of treatment on the lagged outcome variable is close to zero, it is more plausible that the unconfoundedness assumption holds. Therefore, I estimate the effect of $adopt_i$ on $weightLoss_{it}$ for periods before the introduction of challenges on the platform ($t \leq 4$), using the weighted sample. As shown in model M11, Table 3.9, $adopt_i$ does not have a significant effect on the weight-loss outcome before the introduction of challenges. Thus, I can conclude that the ignorability assumption is plausible. The intuition behind this test is that if ignorability does not hold, if being an adopter is positively (or negatively) correlated with having high (or low) motivation

[14] In other words, adoption is exogenous, and I can write: $adopt_i | X_i \perp e_{it}$.

to lose weight, we should see a positive (or negative) effect of on the weight-loss outcomes before the introduction of challenges. However, model M11 does not show such relationship. Note that in model M11, due to small number of periods, the Hansen test p-value and the autocorrelation AR(2) p-value is not calculated. As a robustness check, in model M12, I run the same analysis using OLS, and I found qualitatively similar results.

3.5.5 Difference-in-Difference

I compare the system GMM results with a difference-in-difference model with time-varying treatment, coupled with propensity score matching. The difference-in-difference approach and results are explained in detail in Appendix A.1. The results show that although the effect of challenges on weight-loss remains positive and significant in a difference-in-difference model, the magnitude of the effect is smaller and close to the magnitude of a biased fixed-effect model.

Chapter 4

A GAMIFIED ONLINE DATING PLATFORM

4.1 Setting

4.1.1 Mobile Dating App

The data in this study come from a popular online dating iOS mobile application in the United States. The app (or platform) is targeted at a younger demographic, and those using it are often looking for fun and flirtation rather than long-term dating/marriage partners. To join and use the app, users need a Facebook ID. When the user first logs in to the app (using his/her Facebook ID), the user's name, gender, age, education and employment information, and Facebook profile picture are automatically imported from his/her Facebook account into the user's dating profile in the app. Users cannot change this information in their dating profile directly. However, they can upload up to five more pictures, and add a short bio to their profile. Further, the app has access to a user's real-time geographic location (based on the GPS in the mobile device) when the user is actively using the app.

The app requires users to participate in a structured matching game, which is described in detail below. A user, in fact, cannot directly access or browse other users' profiles through the app; the only way to use the app is to play the ranking game described next.

4.1.2 Description of the Game: Game Assignment

Initiation and completion of a game requires the live participation of four men and four women. When a user logs in to the app and decides to play a game, s/he is assigned to a game-room by the platform. Among the available players, only two criteria are used by the platform to assign players to games – proximity in geographic location and age. The

exact algorithm is as follows: the geographic location of the first player assigned to a game-room is set as the initial center point of that game; the next player is then assigned to that game if he/she is within 500 miles of this center point. The center point is then updated as the average location of the first two players. The third player assigned to the game has to be within 500 miles of the new center point and after s/he is assigned to the game, the geographic center is again updated. This continues until four men and four women have been added to the game. Similarly, the platform ensures that the age gap between any two members in a game is no more than six years (older or younger). In this data, I find that this constraint is trivially satisfied because a vast majority of players belong to a small age bandwidth. Therefore, conditional on geography and age, the assignment of users to games is random.

4.1.3 Description of the Game: Game Activity

When a game starts, participants can see a list of four short profiles of the members of the opposite sex. As shown in the left panel of Figure 4.1, these short profiles display a thumbnail version of users' profile picture, name, age, location and their star-rating (see §4.2.3 for a detailed description on star-ratings). Tapping on a short profile leads to the full profile of the user. As shown in the right panel of Figure 4.1, full profiles typically contain a larger version of the profile picture (and possibly additional photos) and other information, such as bio, education or employment information.

Each user then indicates his/her rank-ordered preference for the four members of the opposite sex. All users have exactly 90 seconds from the start of the game to finalize their rank-orderings.[1]

Two points are worth noting here. First, players do not know the identities and attributes of the other members of their own sex in the game, i.e., men (women) do not know which other men (women) are in the same game. Thus, players do not have any visibility into their

[1]If one or more users leave the game or do not complete their rank-ordering, the game is deemed incomplete and no matches are assigned. In this data, I see a very high rate (over 97%) of game completions.

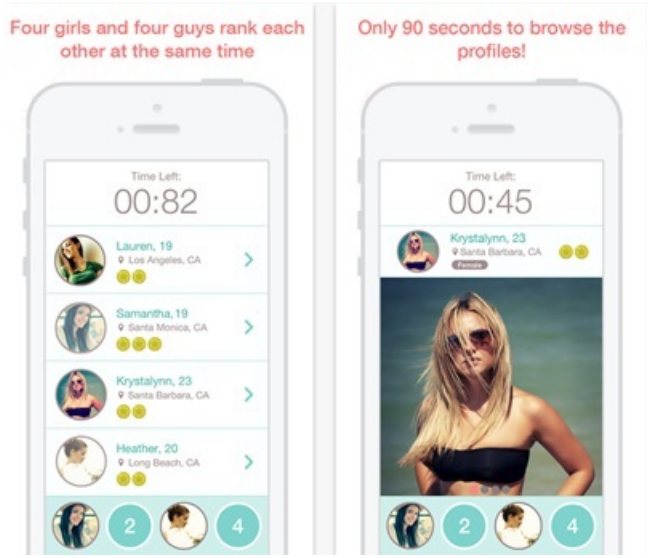

Figure 4.1: Screen shot of the app during a game (from the perspective of a male user). Players indicate their rank-ordered preference for the players from the opposite sex by dragging their profile pictures into the circles labeled one through four at the bottom of the app. In this example, the focal player has picked his first and third choices, and is yet to decide his second and fourth choices.

competition within each game, though they may have a sense of the general distribution of players of their own sex. Second, players' actions are simultaneous and private, i.e., each user only has visibility into his/her own actions and at no point is the rank-ordering of the other players revealed to them (though they may be able to make some inferences after the game based on their match assignments). Hence, while choosing their rank-orderings, they cannot use information on other players' preferences to make their own choices.

4.1.4 Description of the Game: Match Allocation

The platform uses the rank-ordered preferences of all players in a game to derive a set of "stable matches", where the concept of stability is based on the canonical Stable Marriage Problem (SMP): "Given n men and n women, where each person has ranked all members of the opposite sex in order of preference, match the men and women such that there are no

Figure 4.2: Screen shots of the application before and after a game

two people of opposite sex who would *both* prefer each other over their current partners," [35].

There are a few noteworthy points about the SMP. First, for any combination of preferences, there always exists at least one stable match, i.e., there is at least one solution to a SMP. Second, the SMP can have more than one solution even for a relatively small number of players and the optimality of these solutions can depend on the algorithm used. For instance, it is shown that a "Men-proposing Gale-Shapley Deferred Acceptance algorithm" is men-optimal, i.e., none of the men can do better under a different algorithm [35].[2,3]

In this study, the platform first calculates all possible solutions for a game by considering

[2]Similarly, a women-proposing Gale-Shapley Deferred Acceptance algorithm is women-optimal, i.e., none of the women can do better using a different algorithm.

[3]I briefly describe the Men-proposing Gale-Shapley Deferred Acceptance algorithm here: In the first iteration, each man proposes to the woman he prefers most. Then, each woman accepts the offer she prefers most. In each subsequent iteration, each unmatched man proposes to the most-preferred woman to whom he has not yet proposed regardless of whether the woman is already matched or not. Then, each woman chooses among the set of all the men who propose in this iteration as well as the one whom she is currently matched. This process is repeated until all men are matched. It can be shown that this algorithm always reaches a stable solution [35].

all combinations of matches and checking for stability. If a game has only one unique solution, then the platform allocates matches based on this solution. If there are two or more solutions, the solution that offers the highest average match is chosen. The average match of a solution is calculated as follows: take the ranking that each player gave the person s/he is paired with in a stable match and sum this number over all players. The intuition here is to pick the solution that, on average, gives each player her highest preference (or lowest numerical rank). Thus, the platform does not optimize for either men or women, but instead tries to pick the best globally optimal solution.

The entire matching process takes less than a second and users can see the match assigned to them as well as all the other matches allocated in the room (see the right panel of Figure 4.2).

4.1.5 Description of the Game: Post-Game Actions

After they have been assigned a match, users have the option to send a message to their match. Each matched pair can communicate via text and/or picture and video messages, as shown in Figure 4.2 on the right panel. Users also have the choice to not initiate a conversation with their assigned partner and instead play another game, go to the home page or close the app. However, if they choose any of the latter actions without first sending a message to their matched partner, they lose the option to communicate with them in the future (unless the matched person sends them a message, in which case they can respond to it and continue the conversation). Once users initiate or receive a message, the message stays in their Inbox, and they can continue to communicate with that person in the future, if they choose to. Finally, note that users cannot start or receive any communication from other players in the game with whom they have not been matched.

4.2 Data

The data in this study comprises of 94,386 games played by 24,653 unique users during the ten month period from September 15$^{\text{th}}$ 2014 to July 15$^{\text{th}}$ 2015. The data can be categorized

into three groups: 1) User-level data, 2) User-User level data, and 3) User-Game level data. I now describe the variables in each of these categories and present some summary statistics on them.

4.2.1 User-level Data

I start by describing the variables that characterize the time-invariant attributes associated with a user. These remain fixed for the duration of the observation period.[4]

For each user i in this data, I have information on:

- $gender_i$: A dummy variable indicating user i's gender; is 1 for men and 0 for women.
- age_i: User i's age.
- bio_i: The length of user i's bio in his/her profile (i.e., number of words).
- $education_i$: Categorical variable that denotes the user i's highest education level (either earned or working towards), where 1 = High-school, 2 = College, and 3 = Graduate school.
- $employment_i$: Number of positions/companies mentioned in user i's profile.
- $initial\_game_i$: Total number of games played by user i before the data collection period.
- $total\_game_i$: Total number of games played by user i during the data collection period.
- num_pic_i: Number of uploaded pictures in the dating profile.

In addition, I also have access to the profile picture of user i. To obtain a measure of the physical attractiveness of a user's profile picture, I conducted a survey. I asked 384 heterosexual subjects in a research lab to rate the profile pictures of the opposite sex (men rated women and vice-versa), on a scale of 1 to 7, with 1 being "not at all attractive" and 7 being "very attractive". The subjects were undergraduate students at the University of Washington, with an equal fraction of male and female, and their ages ranged between 18-25 (with a median age of 21). This demographic distribution closely mimics the age and gender

[4]In principle, some of these attributes may change over time. However, I do not observe many such changes during the period, and therefore treat them as time-invariant attributes.

Variables	Mean	Std. Dev	25^{th}	50^{th}	75^{th}	(Min, Max)	Size
age_i	21.53	5.41	19	21	22	(13, 109)	22024
bio_i	67.04	275.58	0	0	63	(0, 29519)	22948
$employment_i$	2.05	1.59	1	2	3	(1, 68)	15579
$initial\_game_i$	59.50	64.32	0	48	90	(0, 2146)	24653
$total\_game_i$	31.27	37.90	6	18	45	(1,1069)	24653
num_pic_i	4.26	1.01	4	4	4	(0, 6)	22669
pic_score_i	0.00	0.68	-0.52	-0.09	0.43	(-2.88, 3.29)	17739
$gender_i$	(0) female: 42.45%		(1) male: 57.55%				24653
$education_i$	(1) high-school: 19.24%		(2) college: 78.12%		(3) graduate: 2.64%		21604

Table 4.1: Summary statistics of user-level data.

distribution of the app users.

During the lab study, each subject rated 100 pictures in approximately 20 minutes. In order to minimize biases due to boredom or fatigue, subjects were shown the profile pictures in a random order. On average, each profile picture was rated by five subjects to ensure that the ratings captured average appeal rather than idiosyncratic preferences of a specific subject. It is possible that some subjects give consistently higher or lower ratings than other subjects. I therefore standardized each rating by subtracting the mean rating given by the subject and dividing by the standard deviation of the subjects ratings [6]. I then take the average of all the standardized ratings that user i's picture received in the study and denote it as:

- pic_score_i: The average physical attractiveness score of user i's profile picture.

Finally, because of constraints in subject-pool time, I could only obtain the picture-scores for a random sub-sample of users instead of the full pool of users; thus I have picture-score information for 17,753 of the 24,653 unique users.[5]

[5]The lack of *pic_score*s for 6,900 users does not affect the main analysis since I use a fixed-effects specification, which conditions out all user-specific variables.

The summary statistics of all the user-level variables are shown in Table 4.1. Of the 24,653 users, 14,189 (57.55%) are male and 10,464 (42.45%) are female. The median user is 21 years old, has no bio written on her/his profile, has/is working towards a college degree, and two employment-related information listed on her/his profile. In terms of activity, the median user had played 48 games before the data collection period and plays 18 games during the observation period. However, there is quite a bit of variation across users in the extent of activity, with some users playing over 1000 games during the observation period.

4.2.2 User-User level data

Each game consists of eight unique users – four men and four women. For each man-woman pair in a game, I have data on the preference-ranking that they gave each other, their match outcome, and their post-game interactions. I describe these variables in detail below.

- $pref_{ijt}$: An integer variable that denotes the preference-ranking that user i *receives* from user j in game t; it can take values from one to four, with four indicating the highest preference and one the lowest.

Users rank members of the opposite sex in a game from one through four (as shown in Figure 4.1), with a rank of one indicating their highest preference and four indicating the lowest preference. I convert these rank orderings to preference-rankings, such that rank of one denotes a preference-ranking of four, rank of two indicates a preference-ranking of three, and so on. The transformed variable *pref* is easier to interpret and more intuitive because higher values of this variable correspond to more preference (unlike rank, where lower rank indicates higher preference, which complicates exposition).

- $match_{ijt}$: A dummy variable indicating whether user i is matched with player j in game t. In each game, all players are uniquely matched with one other player from the opposite sex. So for woman (man) i in a game, this variable is set to one for only man (woman).
- $first_{ijt}$: A dummy variable indicating whether user i receives the first message from the player he/she is matched with (denoted by j here) after game t. Note that users are not

given the option to communicate with players they have not been matched with, i.e., they can only communicate with the person they have been matched with by the platform. So, by default, this variable is zero if $match_{ijt} = 0$.

- $reply_{ijt}$: A dummy variable indicating whether user i receives a reply message from the player j after game t, conditioned on user i initiating the first message. By default, this variable is zero if $match_{ijt} = 0$ or $first_{jit} = 0$.

The summary statistics of these variables are shown in Table 4.2. The sample sizes of *pref* and *match* reflects the fact that there are 32 observations per game.[6] The distributions of *pref* and *match* are determined by the game structure, and their summary statistics are as expected. The sample size of $first_{ijt}$ reflects the fact that there are eight users matched with each other, and each of them can potentially initiate the first message. It is worth noting that the mean of $first_{ijt}$ is around 0.05 (of the 713,014 matches, only 39,377 messages were initiated). The observed number of first messages (39,377) defines the sample size of $reply_{ijt}$. The mean of $reply_{ijt}$ is around 0.08 (among 39,377 initiated message only 3380 of them receive a reply). Interestingly, 76% of the conversations are initiated by men, which indicates that women are less likely to approach men after being matched. Further, men receive a reply to their messages 5% of the times, and women receive a reply 20% of the times. These statistics are consistent with previous research on online dating, which find that men are more likely to initiate contact and respond to emails/messages, compared to women [48, 30, 39].

[6]Eight users participate in each game and each user receives four preference-rankings from players of the opposite sex. So I have a total of $8 \times 4 = 32$ preference-rankings per game. Also, since each user can get matched with only one user among the four potential mates, $match_{ijt}$ becomes one once, and becomes zero three times. Thus, for each game I have $8 \times 1 + 8 \times 3 = 32$ data points for $match_{ijt}$. Therefore, the size of $pref_{ijt}$ and $match_{ijt}$ should be the number of games (94,386) $\times$ 32 = 3,020,652. However, some of these data points are related to users whose gender changes in the data set over the data collection period time (42 users), and I exclude them from our analysis.

Variables	Mean	Std. Dev	25^{th}	50^{th}	75^{th}	(Min, Max)	Size
$pref_{ijt}$	2.5	1.12	2	3	4	(1, 4)	3008560
$match_{ijt}$	0.25	0.43	0	0	0.5	(0, 1)	3008560
$first_{ijt}$	0.05	0.23	0	0	0	(0, 1)	713014
$reply_{ijt}$	0.08	0.28	0	0	0	(0, 1)	39377

Table 4.2: Summary statistics of user-user level data.

4.2.3 User-Game level data

I now describe user-game level variables, i.e., user-specific data that varies with each game.

- $match\_level_{it}$: An integer variable that denotes how much user i prefers his match in game t.

$$match\_level_{it} = pref_{jit} \quad \text{if } match_{ijt} = 1 \tag{4.1}$$

- $total\_game_{it}$: Total number of games that user i has played before game t. This is updated by one after each game played by user i.
- $star_{it}$: Indicates the star-rating that the user is shown with in game t; see Figure 4.1 for an example. It is updated in real time after each game and is calculated as follows:

$$star_{it} = \begin{cases} 1, & \text{if } 1 \leq popularity_{it} < 2 \\ 2, & \text{if } 2 \leq popularity_{it} < 3 \\ 3, & \text{if } 3 \leq popularity_{it} \leq 4, \end{cases} \tag{4.2}$$

  where popularity is defined as the average of the preference-rankings that user i has received *before* the t^{th} game, as shown below:

$$popularity_{it} = \frac{\sum_{q=1}^{total\_game_{it}} \sum_{j=1}^{4} pref_{ijq}}{4 \times total\_game_{it}}. \tag{4.3}$$

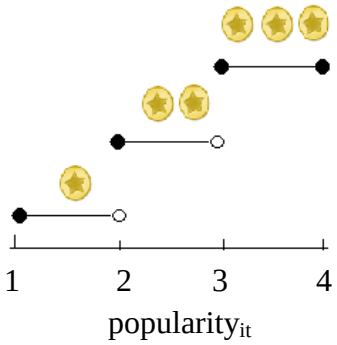

Figure 4.3: Pictorial representation of the star-rating rule (as a function of average preference-ranking in past games).

While users know their own star-rating before each game, and members of the opposite sex in the game room can observe a user's star rating, the platform does not reveal a user's popularity scores to her/him or to anyone else in the platform.

Figure 4.3 illustrates the relationship defined in Equation (4.2). Intuitively, an individual's star-rating captures how popular or sought after s/he was in her/his past games. Three star users, on average, are those who were among the top two choices of other players. Two star players are those who, on average, were the second or third choice of players in the past . Finally, one star players, on average, are those who were the third or fourth choice of others in the past. Thus, there is a clear monotonic relationship between past popularity and current star-rating.

The summary statistics of all the user-game level variables are shown in Table 4.3. There are a few interesting points of note. First, the average *match_level* is 3.19, which implies most users get matched with their first or second top choices, on average. I also find that the median of $total\_game_{it}$ is 59, which suggests that most users have played a good number of games before a median game in the observation period. Finally, I also see that users are shown with a two-star rating on average.

Finally, I examine the extent of variation in star-ratings within an individual. Of the

Variables	Mean	Std. Dev	25^{th}	50^{th}	75^{th}	(Min, Max)	Size
$match\_level_{it}$	3.19	0.95	3	3	4	(1, 4)	752140
$total\_game_{it}$	74.75	74.25	29	59	97	(0, 2194)	752140
$star_{it}$	2.00	0.10	2	2	2	(1,3)	745037

Table 4.3: Summary statistics of user-game level variables.

24,653 users in this data, 85.83% (21,159 users) are shown with two stars in all their games, i.e., they never experience a star change. However, 3,494 users experience a star change. Of these, 1,287 users were shown with a minimum of one star and a maximum of two stars, and 2,185 users were shown with a minimum of two stars and a maximum of three stars. Very few users (22) experienced a minimum of one star and a maximum of three stars. In sum, while a majority of users never experience a star change, there is a sufficiently large portion that goes through at least one star change.

4.3 Descriptive Analysis

I now examine the relationship between a user's star-rating and three measures of her/his demand – preference-rankings received during the game, and whether s/he receives a first messages, or reply message after the game – using simple model-free analyses. In this section, I focus on users who experienced at least one change in their star-rating during the observation period.

The relationship between a user's star-rating in a given game and the average preference-ranking that s/he receives in that game is illustrated in Figure 4.4. The solid increasing line shows the relationship between the average preference-rankings received for all user-game observations calculated for each star-rating.[7] I see that in observations where users have higher star-ratings, they also receive higher preference-rankings. However, there is an

[7]For example, the average preference-ranking for the data point at $star1$ on the solid line is $\frac{\sum_i \sum_t \sum_j (pref_{ijt} star_{it}=1)}{4 \times \sum_i \sum_t I(star_{it}=1)}$.

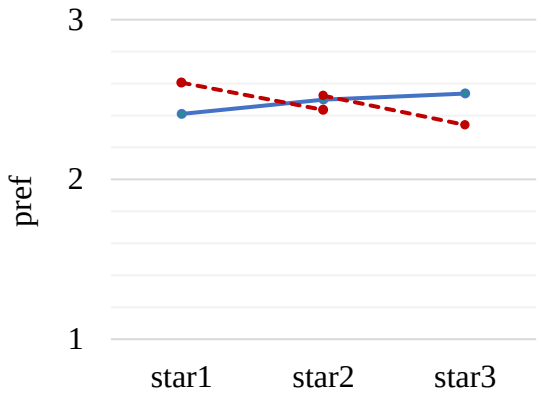

Figure 4.4: The relationship between star-ratings and average preference-rankings received. The solid line is for all user-game data points and the dashed lines are for within-individual data points.

obvious issue of correlated unobservables here, i.e., users with higher star-ratings are likely to be more attractive on other unobserved dimensions (e.g., physical attractiveness) as well. To examine if this conjecture is true, I plot the average of users' *pic_score* for each star-rating. As shown in Figure 4.5, users with higher star-ratings also have higher physical attractiveness score, on average. Thus, the effects shown by the solid line in Figure 4.4 cannot be interpreted as causal.

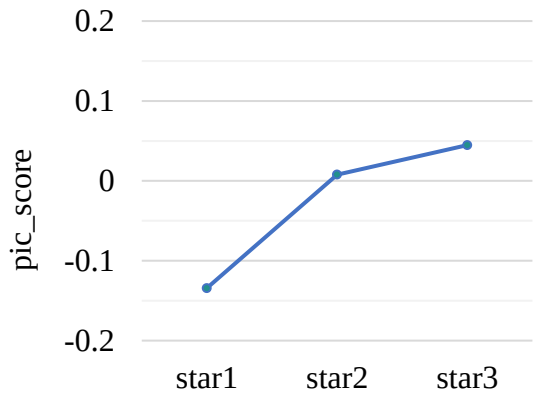

Figure 4.5: The relationship between star-ratings and average physical attractiveness score.

One possible way to cleanly capture the effect of star-ratings is to look at the effect of star-ratings *within* an individual, i.e., if preference-rankings received for the same individual are compared when s/he is shown with different star-ratings, then these comparisons are less likely to be subject to endogeneity concerns. I will expand on this theme in the next two sections, but for now, I present some graphical model-free evidence using this intuition.

First, I consider individuals who were shown with a minimum of one star and a maximum of two stars. For each of these individuals, I calculate two averages: (1) the average of preference-rankings received in games where s/he is shown with one star, and (2) the average of preference-rankings received in games where s/he is shown with two stars. I then perform an analogous exercise for users who were shown with a minimum of two stars and a maximum of three stars. The results of these comparisons are presented using dashed lines in Figure 4.4. As it is shown, on average, the same set of users receive higher preference-rankings when they are shown with one star compared to two stars. Moreover, on average, the same set of users receive higher preference-rankings when they are shown with two stars compared to three stars. In sum, the dashed lines in Figure 4.4 suggest that higher star-ratings leads to lower preference-rankings, i.e., users avoid those with higher stars! Note that the direction of the effect of star-rating on preference-rankings in solid line and dashed lines in Figure 4.4 are exactly opposite. This discrepancy implies that controlling for the endogeneity between star-ratings and unobserved factors that affect user attractiveness is essential to deriving the causal impact of star-ratings in this setting.

Similarly, Figures 4.6 and 4.7 show the relationship between a users star-rating and the likelihood of her receiving the first message and receiving a reply if she initiates a message, respectively. The solid lines show the relationship between the likelihood of receiving a message (first or reply) for all user-game observations calculated for each star-rating.[8] I see that observations where users have higher star-ratings are more likely to receive the first

[8]For example, in Figure 4.6, the data point on the solid line for $star1$ is given by $\frac{\sum_i \sum_t (first_{ijt} star_{it}=1, match_{ijt}=1)}{\sum_i \sum_t I(star_{it}=1)}$, and in Figure 4.7, the data point on the solid line for $star1$ is given by $\frac{\sum_i \sum_t (reply_{ijt} star_{it}=1, match_{ijt}=1, first_{jit}=1)}{\sum_i \sum_t I(star_{it}=1)}$.

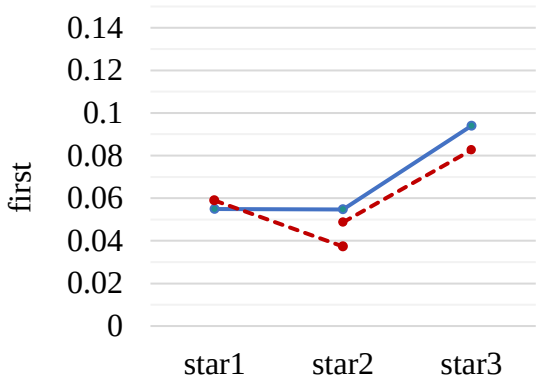

Figure 4.6: The relationship between star-ratings and the average likelihood of receiving the first message. Solid lines are for all user-game observations and dashed lines are for within-individual observations.

messages and replies.

Next, I perform a within individual analysis on users' messaging behavior. As shown by the dashed lines in Figure 4.6, on average, the same set of users are more likely to receive first messages when they are shown with one star compared to two stars. However, this effect does not carryover when two and three stars are compared. In the case of *reply*, the same set of users are more likely to get a reply when shown with higher star-ratings (see dashed lines in Figure 4.7).

In sum, the simple correlation between star-ratings and revealed preferences shows a positive effect. However, the within-individual comparisons shows quite different results. Interestingly, the effect of star-ratings seems to be negative for preference-rankings during the game, partially-negative for initiating communication after the game (first message), and positive when it comes to replying to messages after the game. In the rest of the paper, I focus on deriving the unbiased causal effects of star-ratings on these three revealed preference measures using econometric methods, and exploring the mechanisms driving these effects.

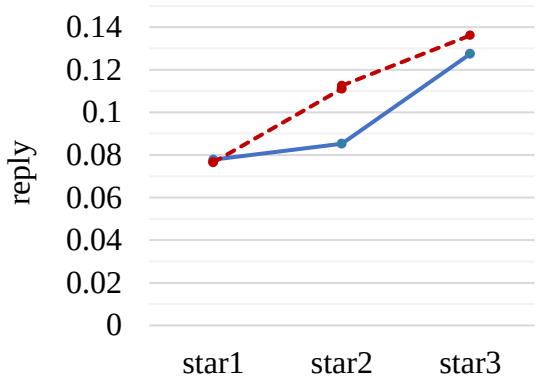

Figure 4.7: The relationship between star-ratings and the average likelihood of receiving a reply message. Solid lines are for all user-game observations and dashed lines are for within-individual observations.

4.4 Effect of Star-ratings on Preference-Rankings

In this section, I formalize the causal impact of a user's star-rating during a game on the preference-rankings that s/he receives during the game. Since preference-rankings are ordinal, I use an ordered logit model to estimate this effect. In §4.4.1 and §4.4.2, I present the model specification and estimation. I discuss the findings in §4.4.6. I present some tests and robustness checks to validate the model and results in §4.4.7.

4.4.1 Model Specification

The outcome variable of interest here is $pref_{ijt}$, which denotes the preference-ranking that user i *receives* from j during game t. Note that *pref* is an ordinal integer value going from 4 to 1, with four indicating the highest preference-ranking and one representing the lowest preference-ranking. Therefore, I use an ordered logit model relates the observed outcome variable $pref_{ijt}$ to a *latent* variable $pref^*_{ijt}$ where:

$$pref^*_{ijt} = \beta_1 star1_{it} + \beta_2 star3_{it} + \gamma z_i + \eta_i + \epsilon_{ijt}, \tag{4.4}$$

The latent variable $pref^*_{ijt}$ is thus modeled as a linear function of:

- $star1_{it}$, $star3_{it}$ – indicator variables for the star-rating of user i in game t, where $star2$ is considered the base.
- z_i – set of user-specific observables that can affect j's ranking of i, e.g., age of i.
- η_i – set of unobservable (to the researcher) characteristics of user i that is visible to j and affects j's ranking of i. These could include the aspects of user i's physical attractiveness not captured in the lab study (e.g., other photos of the user), details in her/his bio description, employment details, her geographic location, etc.
- ϵ_{ijt} – These are factors uncorrelated to the star-rating of user i that can affect the preference-ranking s/he receives from j in game t. Three key sets of variables are subsumed here.
  - First, it includes j's attributes (both observable z_j and unobservable η_j) since there is no correlation between j and i's attributes.
  - Second, it also includes all the attributes of the other three players of i's gender who i is being compared with, in game t.

    The reason neither of the above two sets of variables affect the inference on star-ratings is because the app adds users into a game randomly. Thus, there is no correlation between the attributes of users within a game.[9]
  - Third, ϵ_{ijt} may include idiosyncratic factors that affect j's ranking of i within the game, e.g., j's mood for going on a date with someone of i's type etc.

  I assume that ϵ_{ijt}s have a logistic cumulative distribution. Although, the second point above can create correlation between ϵ_{ijt}s in one game, in §4.4.7, I show that the results are robust to such correlations.

The endogeneity concerns in this model mainly stem from the potential correlation between

[9]In principle, because the app only considers adding new users who are within a 500 mile radius of users already in a game, the geographic locations of users in a game are correlated. However, conditional on being in the same room, there is no correlation between the location of two users, and the distance between the users is random. In other words, if I denote the geographic location of users by g, then I can write the location of j as: g_j, where $g_j = g_i + \delta$, where g_i, g_j, δ are two dimensional vectors (latitude, longitude) such that $||g_j - g_i|| \leq 500$. Since I already control for user i's location (g_i) through η_i, the remaining δ is random noise.

η_i and $star_{it}$, i.e., it is expected that $E[star_{it} \cdot \eta_i] \neq 0$. I will come back to this issue when discussing estimation approaches.

I then model the relationship between $pref_{ijt}$ and $pref^*_{ijt}$ as follows:

$$pref_{ijt} = k \quad \text{if } \mu_k < pref^*_{ijt} \leqslant \mu_{k+1} \quad \forall \quad k = 1, 2, 3, 4, \tag{4.5}$$

where the thresholds μ_k are strictly increasing. Further, I assume that $\mu_1 = -\infty$ and $\mu_5 = \infty$. This specification is simply the ordinal choice analog of a binary logit model. Thus, $pref_{ijt}$ can take four possible values, denoted by k. Because the error terms are drawn from a logistic distribution, I can write the cumulative probability function of ϵ_{ijt} as

$$F\left(\epsilon_{ijt} X_{it}, \beta_1, \beta_2, \gamma, \eta_i, \mu_k, \mu_{k+1}\right) = \frac{1}{1 + exp(-\epsilon_{ijt})} \equiv \Lambda(\epsilon_{ijt}), \tag{4.6}$$

where $X_{it} = \{star1_{it}, star3_{it}, z_i\}$. Therefore, the probability of observing outcome k in game t for a pair of users (where user i receives a rank k from user j) can be written as:

$$\begin{aligned} Pr\left(pref_{ijt} = k X_{it}, \beta_1, \beta_2, \gamma, \eta_i, \mu_k, \mu_{k+1}\right) &= \Lambda\left(\mu_{k+1} - \beta_1 star1_{it} - \beta_2 star3_{it} - \gamma z_i - \eta_i\right) \\ &- \Lambda\left(\mu_k - \beta_1 star1_{it} - \beta_2 star3_{it} - \gamma z_i - \eta_i\right) \end{aligned} \tag{4.7}$$

Using this model formulation, I can then write the log-likelihood of the preference-rankings observed in the data as:

$$LL(\beta_1, \beta_2, \gamma, \eta_i, \mu_k, \mu_{k+1}) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{j=1}^{4} \sum_{k=1}^{4} \ln \left[Pr\left(pref_{ijt} = k X_{it}, \beta_1, \beta_2, \gamma, \eta_i, \mu_k, \mu_{k+1}\right)^{I(pref_{ijt} = k)} \right], \tag{4.8}$$

where N is the total number of users observed and T_i is the total number of games played by user i. Notice that the unknown parameters in Equation (4.8) are $\beta_1, \beta_2, \gamma, \eta_i s, \mu_2, \mu_3, \mu_4$. I discuss their estimation in the next section.

4.4.2 Estimation

I am interested in estimating the effect of star-ratings (coefficients β_1 and β_2). There are two possible estimation strategies for this: (1) A pooled estimation strategy, where the user-specific unobservables η_i are ignored, and (2) a fixed-effects approach, where the user-specific

unobservables η_i are allowed to be arbitrarily correlated with the star-ratings. I discuss both these approaches below.

The first approach is straightforward. It simply involves pooling all the user-game data, ignoring the user-specific unobservable η_i, and then maximizing the log-likelihood in Equation (4.8). However, it is important to recognize that the estimates from this approach will be biased in the presence of correlated unobservables. Therefore, in the rest of this section, I focus on estimating β_1 and β_2 after controlling for η_i.

A naive approach to estimation with fixed-effects is to treat the η_i's as parameters and maximize the log-likelihood in Equation (4.8) directly. However, such a Maximum Likelihood Estimator (MLE) is inconsistent with large N and finite T due to the well-known incidental parameters problem [60]. As a result, the estimates of β_1 and β_2 from this approach will be inconsistent too. Chamberlain provides an elegant solution to the incidental parameters problem for the case of binary variable by dichotomizing the ordered outcome variable [15]. In §4.4.3, I describe how to apply the Chamberlain estimator to this setting, in §4.4.4 I clarify the conditions necessary for identification, and in §4.4.5 I describe how the Chamberlain estimators can be combined to form an efficient Minimum Distance estimator.

4.4.3 Estimation: Chamberlain's Conditional Maximum Likelihood Estimator

The ordered outcome variable $pref_{ijt}$ can take $K = 4$ possible integer values, {1,2,3,4}. Therefore, I can transform the random variable $pref_{ijt}$ into $K-1 = 3$ possible binary variables $pref^k_{ijt}$ where:

$$pref^k_{ijt} = I(pref_{ijt} \geq k), \quad \text{where} \quad k = 2, 3, 4. \tag{4.9}$$

For example, the binary variable $pref^4_{ijt}$ indicates whether user i received a preference-ranking of 4 from user j in game t, or not. Similarly, the binary variable $pref^3_{ijt}$ indicates whether user i receives a preference-ranking of 3 or higher (i.e., 3 or 4) from user j in game t, or not. I can specify Chamberlain's Conditional Maximum Likelihood (CML) estimator on each of

these transformed binary variables. For each k, $pref_{ijt}^k$ is a binary logit variable such that:

$$Pr(pref_{ijt}^k = 1 X_{it}, \beta_1, \beta_2, \gamma, \eta_i, \mu_k) = 1 - \Lambda(\mu_k - \beta_1 star1_{it} - \beta_2 star3_{it} - \gamma z_i - \eta_i) \quad (4.10)$$

Next, I denote $pref_i^k$ as the entire history of preference-rankings at level k received by user i over time, i.e. $pref_i^k = \{pref_{i11}^k, pref_{i21}^k, pref_{i31}^k, pref_{i41}^k, ..., pref_{i1T_i}^k, pref_{i2T_i}^k, pref_{i3T_i}^k, pref_{i4T_i}^k\}$. Further, I denote s_i^k as the sum of all the binary transformed preference-rankings at level k received by user i over time:

$$s_i^k = \sum_{t=1}^{T_i} \sum_{j=1}^{4} pref_{ijt}^k$$

In other words, s_i^k shows the count of ones in the set of $pref_i^k$. Further, I denoted B_i^k as the set of all possible vectors of length $4 \times T_i$ with s_i^k elements equal to 1, and $4 \times T_i - s_i^k$ elements equal to 0. That is:

$$B_i^k = \{d \in \{0,1\}^{4 \times T_i} \sum_{t=1}^{T_i} \sum_{j=1}^{4} d_{jt} = s_i^k\} \quad (4.11)$$

Note that the size of $B_i^k = \binom{4 \times T_i}{s_i^k}$.[10]

Now, I can write the conditional probability of $pref_i^k$ given s_i^k as:

$$Pr\left(pref_i^k star1_{it}, star3_{it}, s_i^k, \beta_1, \beta_2\right) = \frac{exp\left(pref_i^k \cdot (\beta_1 star1_{it} + \beta_2 star3_{it})\right)}{\sum_{d \in B_i^k} exp\left(d \cdot (\beta_1 star1_{it} + \beta_2 star3_{it})\right)} \quad (4.12)$$

A key observation is that this conditional probability does not depend on η_i's or the thresholds μ_k's, i.e., s_i^k is a sufficient statistic for η_i. Thus, I can now specify a Conditional Log-Likelihood that is independent of η_is and μ_ks as shown below:

$$CLL(\beta_1^k, \beta_2^k) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \ln\left[Pr(pref_i^k star1_{it}, star3_{it}, s_i^k, \beta_1^k, \beta_2^k)\right] \quad (4.13)$$

---

[10]For example, consider user i who plays only two games ($T_i = 2$). For $k = 4$, I have $pref_{ijt}^4 \in \{0,1\}$ that denotes whether user i has received a preference-ranking of 4 from user j or not. Now, let's consider a scenario where user i receives a preference-ranking of four only in her first game and from j_1, i.e., $pref_i^4 = \{1,0,0,0,0,0,0,0\}$. Thus, $s_i^4 = 1$. Next, I can write B_i^4 or the set of all possible ways that user i can get only one preference-ranking of 4 in her games by $B_i^4 = \{(1,0,0,0,0,0,0,0), (0,1,0,0,0,0,0,0), ..., (0,0,0,0,0,0,1,0), (0,0,0,0,0,0,0,1)\}$. Note that each element of B_i^4 is itself a vector with eight elements, because user i has played two games and in each game s/he receives four preference-rankings ($4 \times 2 = 8$). I denote each element of set B_i^4 with vector d. Also, notice that the size of B_i^4 is eight, because $\binom{4 \times 2}{1} = 8$.

Since I can dichotomize $pref_{ijt}$ into three binary variables at each of the three cutoffs ($pref^4_{ijt}$, $pref^3_{ijt}$, and $pref^2_{ijt}$), the above CLL can be specified for each $pref^k_{ijt}$, where $k \in \{2, 3, 4\}$. Maximizing each of these CLLs gives us three separate but consistent estimates of β_1, β_2, which I denote as $\{\beta^k_1, \beta^k_2\}$, where $k \in \{2, 3, 4\}$. These are referred to as Chamberlain CML estimators.

4.4.4 Estimation: Identification

Two necessary conditions need to be satisfied for the identification of $\{\beta^k_1, \beta^k_2\}$. First, within-user variation is needed in $star1_{it}$ and $star3_{it}$. Intuitively, this estimator takes advantage of the variation in star-ratings "within" a user for identifying the effect of star-ratings. This can circumvent the problem of user-specific correlated unobservables since they remain constant for the user across time. If the same user i receives lower preference-rankings when s/he is shown with three stars as opposed to two stars, that difference can be directly attributed to the change in star-rating since it is the only variable that has changed across time (assuming that the inherent attractiveness of the user remains constant over the duration of observation).

Second, within-user variation is needed in the outcome variable $pref^k_{ijt}$ because users with constant $pref^k_{ijt}$ do not contribute to the CLL for cut-off k (and hence identification).[11] I now illustrate this condition using an example. For $k = 4$, consider a user i who has either received a preference-ranking of 4 in all her games, or never ever received a preference-ranking of 4 in any of her games. This user does not contribute to the CLL because her outcome ($pref^4_{ijt}$) is constant over time even if her/his star-rating varies over time. Thus, only users for whom I have across-time variation in both the outcome variable ($pref^k_{ijt}$) and the independent variables ($star1_{it}$, $star3_{it}$) contribute to the identification of $\{\beta^k_1, \beta^k_2\}$.

Intuitively, at any cut-off k, only the variation around k is used for identification because of dichotomization; for example, the CLL for $k = 4$ only considers whether $pref_{ijt}$ is greater

[11] Constant $pref^k_{ijt}$ means that all elements of B^k_i are either zero or one.

than or equal to 4 and ignores the variation in $pref_{ijt}$ when it is less than 4. Thus, while Chamberlain's CML estimator at each k is consistent, it is not efficient because it does not exploit all the variation in data.[12]

4.4.5 Estimation: Minimum Distance Estimator

To address the efficiency issue in Chamberlain's CML, a Minimum Distance (MD) estimator is developed that combines all the Chamberlain estimates [21]. I now describe the application of this method for this study below.

Recall that there are $K-1=3$ estimates for each of $\{\beta_1, \beta_2\}$: $\{\beta_1^1, \beta_2^1\}, \{\beta_1^2, \beta_2^2\}, \{\beta_1^3, \beta_2^3\}$. Since each of these three estimates are consistent, any weighted average of these estimates will be consistent too. The main idea in the MD estimator is to use the variance and covariances of $K-1$ estimators as weights and generate one efficient estimate. It thus involves solving the minimization problem:

$$\hat{\beta}^{MD} = \underset{b}{\operatorname{argmin}} (\tilde{\beta} - Mb)' var(\tilde{\beta})^{-1} (\tilde{\beta} - Mb), \tag{4.14}$$

where $\tilde{\beta}$ is the 6×1 matrix of Chamberlain estimators, M is the matrix of 3 stacked 2-dimensional identity matrices, and $var(\tilde{\beta})$ is the variance-covariance matrix of the stacked

[12]For individuals who have played a large number of games (large T_i) and have a large number of positive values of $pref_{ijt}^k$ (large s_i^k), calculating all combinations of outcomes can lead to numerical overflow and computational issues. For example, if user i plays 100 games ($T_i = 100$) and receives one preference-ranking of four in each game, then $s_i^4 = 100$ and $\binom{4\times100}{100} = 2.24e + 96$. Therefore, I limit the empirical analysis to users' first 100 games. Of the 3,494 users who experience a star change, only 352 (10%) users play more than 100 games. The consistency of the estimates is not affected if I choose a subset of games for players who have played a large number of games.

Chamberlain estimates. In other words, b_1 and b_2 should be such that:

$$\underset{b_1,b_2}{\text{argmin}} \begin{pmatrix} \tilde{\beta}_1^2 - b_1 \\ \tilde{\beta}_2^2 - b_2 \\ \tilde{\beta}_1^3 - b_1 \\ \tilde{\beta}_2^3 - b_2 \\ \tilde{\beta}_1^4 - b_1 \\ \tilde{\beta}_2^4 - b_2 \end{pmatrix}' \begin{pmatrix} var(\tilde{\beta}_1^2) & & & & & \\ cov(\tilde{\beta}_2^2, \tilde{\beta}_1^2) & var(\tilde{\beta}_2^2) & & & & \\ . & . & var(\tilde{\beta}_1^3) & & & \\ . & . & . & var(\tilde{\beta}_2^3) & & \\ . & . & . & . & var(\tilde{\beta}_1^4) & \\ cov(\tilde{\beta}_2^4, \tilde{\beta}_1^2) & . & . & . & cov(\tilde{\beta}_2^4, \tilde{\beta}_1^4) & var(\tilde{\beta}_2^4) \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\beta}_1^2 - b_1 \\ \tilde{\beta}_2^2 - b_2 \\ \tilde{\beta}_1^3 - b_1 \\ \tilde{\beta}_2^3 - b_2 \\ \tilde{\beta}_1^4 - b_1 \\ \tilde{\beta}_2^4 - b_2 \end{pmatrix}$$

The solution to the above minimization problem (b_1 and b_2) is a weighted average of the Chamberlain estimators and is equal to:

$$\hat{\beta}^{MD} = \{M' var(\tilde{\beta})^{-1} M\}^{-1} M' var(\tilde{\beta})^{-1} \tilde{\beta} \tag{4.15}$$

and its variance is given by $var(\hat{\beta}^{MD}) = \{M' var(\tilde{\beta})^{-1} M\}^{-1}$. I implement this MD estimator using Stata [40].

4.4.6 Results

The results from the estimation exercise are presented in Table 4.4. I start by considering two pooled ordered logit models: (1) model M1 – a simple model that only includes star-ratings as the independent variable, and (2) model M2 – a slightly more elaborate model that includes all the user-specific observables (z_i) as controls.

In the basic pooled ordered logit model (model M1), I see a positive and significant effect for higher star-ratings. That is, one-star users receive lower preference-rankings compared to two-star users, and two-star users receive lower preference-rankings compared to three-star users. Thus, it seems like higher star-ratings lead to higher preference-rankings. This result is consistent with Figure 4.4 (solid line). Since a user's current star-rating is likely to be positively correlated with her/his appeal in the dating market (through physical attractiveness, age, education, etc.), in model M2 I control for all the user-specific observables. However, the direction of the results remain unchanged. Nevertheless, without explicitly controlling

for the endogeneity concerns discussed earlier ($E[star_{it}.\eta_i] \neq 0$), the estimates are likely to be biased.

Therefore, now I focus on the results from the fixed-effects MD estimator (model M3). Interestingly, here I find that the effect of star-rating is *negative* – a user gets worse preference-ranking when s/he is shown with three stars as opposed to two stars. I do not find any significant effect of one star compared to two stars. In §4.4.7, I present a battery of robustness checks to confirm the validity of these empirical findings.

	(M1)	(M2)	(M3)
	(Ordered Logit)	(Ordered Logit)	(FE Ordered Logit)
$star1_{it}$	-0.14452***	-0.11431***	0.02852
	(0.02315)	(0.03021)	(0.01804)
$star3_{it}$	0.06063***	0.05578**	-0.05101***
	(0.01560)	(0.02328)	(0.01464)
Controls		✓	
μ_2	-1.09924***	-1.11220***	
	(0.00203)	(0.01533)	
μ_3	-0.00053	-0.00841	
	(0.00188)	(0.01527)	
μ_4	1.09828***	1.09372***	
	(0.00205)	(0.01529)	
Individuals	24393	11639	3494
Observations	2980148	1580848	630160

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Controls in Model M2 include: age_i, $college_i$, $graduate_i$, pic_score_i, num_pic_i, $employment_i$, and bio_i.

Table 4.4: Ordered logit estimates of the effect of star-rating on preference-rankings received.

The main takeaway from these findings is that popularity information has a negative effect on users' demand during the game. There results document negative returns to popularity in online platforms. As discussed in §2.2, past empirical research has mainly documented positive gains to popularity information or herding effects. In this setting, there could be multiple reasons for the deviation from the standard positive results. It could be because users may dislike the popular users. Or, they may like popular users but avoid them due to rejection concerns: raters (rank-givers) may be concerned that popular users are harder to achieve matches with, and therefore shade their preferences for them in order to avoid rejection costs. In §4.6, I formalize the discussion of the mechanism behind the negative effect of popularity information, tease out these two explanations, and rule out other alternative mechanisms.

In sum, these findings suggest that researchers and managers need to understand the behavioral underpinnings of the mechanism through which popularity information operates within a given market instead of assuming positive effects based on prior work.

4.4.7 Robustness Checks

First, I examine whether the substantive results from §4.4.6 hold if I directly model the outcome as a linear function of star-ratings and other relevant variables. Therefore, I consider three linear specifications – (1) a simple model that only includes star-rating variables as the independent variable, (2) a slightly more elaborate model that includes all the user-specific observables (z_i), and (3) a linear fixed-effects model. These are the linear analogs of models M1, M2, and M3 in Table 4.4. The estimates from these models are substantively similar to those from the ordered logit models. Please see Appendix §B.1.1 for model details and the full table of results.

Next, I examine if the results are driven by the estimation sample used. Recall that the Minimum Distance estimator for the fixed-effects ordered logit model utilizes only a subset of the data for inference – data on users who went through at least one star change during the observation period. In principle, this sub-population can be different from the full

population, and the fixed-effects estimates could simply reflect that difference. In that case, the findings would only apply to local sub-population that saw at least one star change.

I consider two validation checks to confirm that the results are not driven by the sample. First, the results from a pooled ordered logit model on the estimation sample used in the Minimum Distance estimates are similar to those obtained from the full sample. Second, I find no systematic user-level differences between users who go through at least one star change in the data compared to those who do not go through any star change. Please see Appendix §B.1.2 for details.

Second, recall that the effect of star-ratings on preference-ranking and replies were quite different. One explanation of this difference was based on the differences in perceived probabilities of rejection. However, this might be due to the differences in the estimation samples used in models M3 and M6. In model M3, it includes all users who experienced a star-change, and in model M6, it includes users who experienced a star-change and those who initiated a message. As a robustness check, I therefore re-estimate model M3 with the sample used in model M6. I find that the results from this exercise are the same as those presented in M3 (see Table B.4 in Appendix §B.1.2).

Recall that ϵ_{ijt}s can include all the attributes of the other three players of i's gender who i is being compared with in game t. Technically, this can create a correlation between the error ϵ_{ijt}s in one game, if the observation of all competitors in one game are included in the model. As discussed in §4.4.1, this correlation does not affect the consistency of the results, i.e., the estimates are unbiased. However, it can affect the efficiency of the results. To examine if this is an issue, I conduct another robustness check.

Note that a majority of users in the sample never experienced a star change, and recall that the observations of those competitors who never experienced a star change are dropped from the analysis. Therefore, to confirm that the results are not affected by the within game correlation between the errors, I re-estimate the fixed-effects ordered logit model with the games in which only one of the four competitors experienced a star change in the observation period. I find that the results remain similar to those presented in model M3. (See Table

B.5 in Appendix §B.1.3.)

It is possible that users self-select their entry time when they expect certain types of competitors and this may affect the star configuration of the games. So I examine the star configuration of competitors in the games in the data (see Table B.6 in Appendix B.1.4). I find that in 95.92% of the rooms (or games) all four competitors are shown with two stars. This reflects the fact that the majority of users on the platform never experience a star change. For those who do experience a star-change and are shown with three stars, all their three competitors have two stars in 4,735 games. Similarly, for those who experience a star-change and are shown with one stars, all their three competitors have two stars in 2,606 games. Other star configurations are pretty rare in the data. Therefore, regardless of when a three-star or one-star user decides to play a game, they are almost always being compared to other two star users. This ensures that the effect of star-ratings is not driven by users' self-selection into games at certain points in time etc.

4.5 Effect of Star-ratings on Messaging Behavior

In this section, I examine the causal impact of a user's star-rating on her likelihood of receiving messages. I focus on two variables: (1) first_{ijt}: a dummy variable indicating whether user i receives a first message from her match j after game t, and (2) reply_{ijt}: a dummy variable indicating whether user i receives a reply message from player j after game t, conditional on user i initiating the first message. I present the model and estimation in §4.5.1 and discuss the results in §4.5.2.

4.5.1 Model and Estimation

The outcome variables *first* and *reply* are binary. Hence, I consider logit formulations that relate them to latent variables first^*_{ijt} and reply^*_{ijt} as follows:

$$\mathit{first}_{ijt} = \begin{cases} 1, & \text{if } \mathit{first}^*_{ijt} > 0 \\ 0, & \text{else} \end{cases} \tag{4.16}$$

$$reply_{ijt} = \begin{cases} 1, & \text{if } \ reply^*_{ijt} > 0 \\ 0, & \text{else} \end{cases} \tag{4.17}$$

These latent variables are defined as:

$$first^*_{ijt} = \beta^f_1 star1_{it} + \beta^f_2 star3_{it} + \gamma^f z_i + \eta^f_i + \epsilon^f_{ijt}, \tag{4.18}$$

$$reply^*_{ijt} = \beta^r_1 star1_{it} + \beta^r_2 star3_{it} + \gamma^r z_i + \eta^r_i + \epsilon^r_{ijt}, \tag{4.19}$$

where the interpretations of $\{\beta^f_1, \beta^f_2, \gamma^f, \eta^f_i, \epsilon^f_{ijt}\}$ and $\{\beta^r_1, \beta^r_2, \gamma^r, \eta^r_i, \epsilon^r_{ijt}\}$ are similar to that in §4.4.1. Further, following the same arguments, I allow for η^f_i and η^r_i to be arbitarily correlated to $star1_{it}$ and $star3_{it}$. Assuming that ϵ_{ijt}s are IID and drawn from a logistic distribution, the probability that user i receives a first message from user j (conditional on i and j being matched in game t) is:

$$Pr(first_{ijt} = 1 match_{ijt} = 1, X_{it}, \eta^f_i) = \frac{exp(\beta^f_1 star1_{it} + \beta^f_2 star3_{it} + \gamma^f z_i + \eta^f_i)}{1 + exp(\beta^f_1 star1_{it} + \beta^f_2 star3_{it} + \gamma^f z_i + \eta^f_i)}$$

Similarly, the probability that user i receives a reply from user j (conditional on them being matched in game t *and* user i having initiated the first message) can be written as:

$$Pr(reply_{ijt} = 1 match_{ijt} = 1, first_{jit} = 1, X_{it}, \eta^r_i) = \frac{exp(\beta^r_1 star1_{it} + \beta^r_2 star3_{it} + \gamma^r z_i + \eta^r_i)}{1 + exp(\beta^r_1 star1_{it} + \beta^r_2 star3_{it} + \gamma^r z_i + \eta^r_i)}$$

As in the case of the ordered logit model, I can use these probabilities to specify Conditional Log-Likelihoods that are independent of ηs and then maximize the two CLLs to derive consistent estimates of $\{\beta^f_1, \beta^f_2\}$ and $\{\beta^r_1, \beta^r_2\}$. Since these steps are very similar to that described in §4.4.2, I relegate the details to Appendix §B.2.

4.5.2 Results

The results for both message outcomes are shown in Table 4.5. I start with a discussion of first messages (shown in models M4 and M5). Model M4 is a pooled logit model that controls only for the observable attributes of the (potential) receiver and M5 is a fixed-effects logit model estimated using CLL that accounts for the endogeneity between star-ratings and

	First Message		Reply Message	
	(M4)	(M5)	(M6)	(M7)
	(Logit)	(Logit FE)	(Logit)	(Logit FE)
$star1_{it}$	0.06327	0.43077***	-0.19146	-0.42609*
	(0.14248)	(0.09954)	(0.23242)	(0.25744)
$star3_{it}$	0.57401***	0.64912***	0.27474	0.28316*
	(0.08990)	(0.06280)	(0.17387)	(0.15125)
Controls	✓		✓	
Constant	-2.00281***		-1.01352***	
	(0.07918)		(0.22234)	
Individuals	11634	1972	3115	536
Observations	374727	118627	20485	8573

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Controls in Model M13 and M15 include: $gender_i$, age_i, $college_i$, $graduate_i$ pic_score_i, num_pic_i, $employment_i$, and bio_i.

Table 4.5: Effect of star-rating on messages received.

user-specific unobservables. In model M4, I find that three-star users are more likely to receive first messages compared to two-star users. I do not find any significant effect of one star compared to two stars. However, after controlling for the endogeneity issues in model M5, I find both three- and one-star users are more likely to receive first messages compared to two-star users. This is consistent with dashed-lines in Figure 4.6.[13]

In this case, the results are somewhat different from those in model M3 (that characterizes the effect of star-ratings on preference-rankings). On the one hand, the positive effect for

[13]Note that I have only 1,972 users in model (M5). Although, there are 3,494 users who experienced a star-change, some of them are dropped from a fixed-effects logit model because of no variation in their outcome $first_{ijt}$.

one-star users suggests that rejection concerns may be at play since players may expect one-star users to be more responsive to their message. On the other hand, the positive effect of three-star users suggests the possibility that players may value higher-star users more. Thus, these results can be explained by a combination of both higher utility for higher star users as well as lower rejection concerns, and it is hard to tease out the exact mechanism based on them.

So next, I present the results for *reply* behavior in models M6 and M7, which are analogous to M4 and M5. Interestingly, I find that compared to two-star users, one-star users are less likely and three-star users are more likely to receive a reply when they initiate contact. That is, the effect of star-ratings on preference-ranking and replies are quite different (compare models M3 and M7).[14]

The main takeaway here is that, in the case of replies, the effect of popularity information is positive and consistent with the earlier literature on herding. Intuitively, users in the reply condition are unlikely to be concerned about rejection and therefore rejection concerns may not play any role in their reply behavior. In the next section, I formalize and discuss the mechanism that can explain the difference in the effect of star-ratings on preference-ranking and reply behavior in greater detail.

4.6 Discussion of Mechanism

I now examine the mechanism behind the effects established in §4.4 and §4.5. First, I formalize the ranking strategy of players during the game and their messaging behavior after the game. Next, I define strategic shading and discuss how the empirical results can be explained by strategic shading. Finally, I examine the rationality of strategic shading in this setting.

---

[14]One possible reason for the difference in the results in models M3 and M7 could be the difference in the estimation samples used. M3 includes all users who experienced a star-change, whereas model M7 only includes users who experienced a star-change *and* those who initiated a message. As a robustness check, I therefore re-estimate model M3 with the sample used in model M7. I find that the results from this exercise are the same as those presented in M3 (see Table B.4 in Appendix §B.1.2). Thus, these differences are not driven by the estimation sample used.

4.6.1 Ranking Strategy During the Game

I assume that the preference-ranking that user j gives to user i is induced by j's underlying expected utilities. Let $EU(pref_{ijt})$ denote the expected utility that user j gets from giving user i preference-ranking $pref_{ijt}$, such that:

$$EU(pref_{ijt}) = U(star_{it}) \times \mathcal{P} - C \times (1 - \mathcal{P}). \tag{4.20}$$

Here, $U(star_{it})$ denotes the utility that user j expects to receive from a potential conversation/date with i. $U(\cdot)$ can also be a function of other observed i and j specific variables. However, I suppress them in the notation to keep the expressions simple. If j does not get matched with i, s/he incurs a psychological rejection cost of C. I assume this cost is incurred because j can infer that i did not rank him/her high enough. The cost of rejection can also be a function of i's attributes, i.e., C can be written as $C(star_{it}, z_i)$. For instance, j may suffer higher rejection costs if i is popular (three-star) or attractive. However, this does not affect any of the arguments used to demonstrate strategic shading in §4.6.4 and therefore I simply denote it as C to keep the notation simple.[15] Finally, $\mathcal{P}$ denotes j's perceived expected probability of being matched with i conditional on giving i a preference-ranking of $pref_{ijt}$. In §4.6.7, I present the full expansion of $\mathcal{P}$, and show that it is a function of $pref_{ijt}$; for now it is sufficient to simply define it.

4.6.2 Messaging Strategy after the Game

After the game, each user makes a decision on whether to initiate a message with her/his match and whether to reply to a message (if s/he receives one from her match). The decision to send a first message is not central to the discussion in this chapter, so I do not define it in the text.[16] However, the decision to reply to a received (first) message is important. So I

[15] In this context, C only refers to the psychological cost of knowing that other player did not rank you sufficiently high, and not opportunity costs. This is because of the following reason: in the stable matching algorithm that is used to match users, if a user does not get matched with her first choice, then it does not affect her chance of getting matched with her second choice, and so on.

[16] User j's decision to send the first message to user i is based on j's underlying expected utility, and is analogous to Equation (4.20). Let $EU(first_{ijt})$ denote the expected utility that user j gets from sending

now formally define it.

I assume that user j replies to the message sent by user i based on her underlying expected utility. Since i initiated the first message, j is unlikely to have any rejection concerns when replying to i. Thus, unlike Equation (4.20), there is no rejection probability or cost in the expected utility that user j gets from replying to i. Thus I can write:

$$EU(reply_{ijt}) = U(star_{it}). \tag{4.21}$$

4.6.3 Strategic Shading

I now formally define *strategic shading*.

Definition1. Strategic shading: User j's revealed preference for a potential partner i is *not* just based on the expected utility from being matched with her/him ($U(\cdot)$). Instead, user j's revealed preference also takes into account the perceived probability of rejection and rejection costs. This distortion of revealed preference away from $U(\cdot)$ is referred to as *strategic shading*.

Strategic shading can be easily understood in this setting as follows: suppose that users value more popular users, i.e., expect higher utility (U) from dating a popular partner. However, if there is a non-zero probability of rejection and rejection costs are positive, they may reveal lower preferences for popular users. That is, users would strategically shade down their preferences for popular users in order to avoid rejection.

---

a first message to user i:

$$EU(first_{ijt}) = U(star_{it}) \times \mathcal{P}_f - C_f \times (1 - \mathcal{P}_f).$$

The definition of $U(star_{it})$ is the same as before. If j does not receive a reply form i, conditional on initiating a conversation with her/him, s/he incurs a rejection cost of C_f, and $\mathcal{P}_f$ denotes j's perceived expected probability of receiving a reply from i conditional on initiating a conversation with her/him. Although, users may perceive a lower probability of rejection once they have been matched with a partner, the probability of rejection is unlikely to be zero.

4.6.4 Evidence for Strategic Shading

I can identify the presence of strategic shading in this setting based on the differences in the effect of popularity information (star-ratings) on two revealed preference measures that vary only in the severity of rejection concerns: preference-rankings during the game and reply choice after the game.

I start by invoking the empirical findings on the *reply* message from §4.5, which suggests that user j is more likely to send a reply message to a three-star match (who has initiated a first message) compared to two-star match. This implies that

$$EU(reply_{ijt} star_{it} = 3, first_{jit} = 1) > EU(reply_{ijt} star_{it} = 2, first_{jit} = 1). \tag{4.22}$$

Then, based on Inequality (4.22) and Equation (4.21), I can infer that:

$$U(star_{it} = 3) > U(star_{it} = 2). \tag{4.23}$$

This implies that users receive higher utility from a potential conversation/date with a three-star partner compared to a two-star partner.

Next, I characterize the empirical findings from §4.4 (on *pref*) in Inequality (4.24). Recall that user j is more likely to give a lower preference-ranking to i, when i is presented with three stars compared to two stars. Thus, I have:

$$EU(pref_{ijt} star_{it} = 3) < EU(pref_{ijt} star_{it} = 2). \tag{4.24}$$

The above inequality is based on the assumption that users' ranking behavior during the game reflects their true preferences, i.e., preference-rankings reflect users' underlying expected utilities. I refer readers to Appendix §B.3 for a formal statement and validation of this assumption.

Since I know from Inequality (4.23) that $U(star_{it} = 3) > U(star_{it} = 2)$, Inequality (4.24) can only be explained by rejection concerns, i.e., due to positive perceived expected probability of rejection $\mathcal{P}$ and non-zero rejection cost $\mathcal{C}$. Thus, the negative effect of star-ratings during the game can be directly attributed to rejection concerns.

4.6.5 Additional Evidence for Strategic Shading based on Heterogeneous Effects

In the previous section, I showed that the negative effect of three-star ratings on preference-rankings can be explained by rejection concerns. I now provide some additional evidence in support of this idea based on the heterogeneity in the effect of star-ratings on users' ranking behavior during the game. In particular, I examine the heterogeneity in the effect of star-ratings based on physical attractiveness.

I start by stratifying users (rank-givers) based on their physical attractiveness. As summarized in Table 4.1, the median user has a standardized *pic_score* of -0.09. Based on this value, I stratify the data into two groups: (i) Attractive rank-givers: data where the rank-giver's *pic_score* is greater than -0.09, and (ii) Unattractive rank-givers: data where the rank-giver's *pic_score* is less than or equal to -0.09. Then, I re-run the analysis on these two strata of data separately and report the results in the first two columns of Table 4.6. I find that attractive users are not likely to be influenced by the star-ratings of potential partners (model M8). On the other hand, for unattractive rank-givers, the results show a negative and significant effect of $star3_{it}$ (model M9).[17] This suggests that only unattractive users avoid popular users. This is consistent with the hypothesis of strategic shading due to rejection concerns since I expect unattractive users to be more concerned about rejection

---

[17]One important thing to keep in mind is that I cannot compare the magnitudes of the effects of $star1_{it}$ and $star3_{it}$ across models M8 and M9 because the variance of errors is not identified in (ϵ_{ijt}) in (ordered) logit models [2]. For example, suppose that the following models are for attractive and unattractive users:

$$\begin{aligned}
\text{Attractive:} \quad pref^*_{ijt} &= \beta^a_1 star1_{it} + \beta^a_2 star3_{it} + \gamma^a z_i + \eta_i + \epsilon^a_{ijt} \quad \text{where} \quad \epsilon^a_{ijt} \sim GEV1(0, \sigma^2_a) \\
\text{Unattractive:} \quad pref^*_{ijt} &= \beta^u_1 star1_{it} + \beta^u_2 star3_{it} + \gamma^u z_i + \eta_i + \epsilon^u_{ijt} \quad \text{where} \quad \epsilon^u_{ijt} \sim GEV1(0, \sigma^2_u)
\end{aligned}$$

Because the latent variable $pref^*$ is not observed, the variance of error terms are not identified and the above models are rescaled such that what is estimated in practice is:

$$\begin{aligned}
\text{Attractive:} \quad pref^*_{ijt}/\sigma^2_a &= (\beta^a_1/\sigma^2_a) star1_{it} + (\beta^a_2/\sigma^2_a) star3_{it} + \epsilon^a_{ijt}/\sigma^2_a \\
\text{Unattractive:} \quad pref^*_{ijt}/\sigma^2_u &= (\beta^u_1/\sigma^2_u) star1_{it} + (\beta^u_2/\sigma^2_u) star3_{it} + \epsilon^u_{ijt}/\sigma^2_u
\end{aligned}$$

The coefficients of $star1_{it}$ in Models M8 and M9 are not β^a_1 and β^u_1; rather they are β^a_1/σ^2_a and β^u_1/σ^2_u. I report coefficients calculated under the assumption that the variance of error is $\pi^2/3$. However, if residual variation differs between groups, comparing the magnitude of the coefficients across groups can lead to incorrect conclusions.

	(M8)	(M9)	(M10)	(M11)
			Unattractive Rank-Giver	
	Attractive Rank-Giver	Unattractive Rank-Giver	Attractive Rank-Receiver	Unattractive Rank-Receiver
$star1_{it}$	0.00064	0.02566	-0.02015	0.02220
	(0.02938)	(0.02792)	(0.05177)	(0.03998)
$star3_{it}$	-0.03107	-0.07558***	-0.07665**	-0.05318
	(0.02246)	(0.02164)	(0.03190)	(0.03605)
Individuals	3444	3453	1231	1354
Observations	258527	285554	116350	131440

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4.6: Heterogeneous effect of star-rating on received preference-rankings using ordered logit fixed-effects model.

than attractive users.

To further examine the source of the effect for unattractive users, I stratify the data for unattractive rank-givers based on the physical attractiveness of the rank-receivers. I re-run the analysis on these two strata separately. The results from this exercise are shown in models M10 and M11 in Table 4.6. I find that there is a negative and significant effect of $star3_{it}$, only when unattractive rank-givers are ranking attractive potential partners (model M10). Again, these results suggest that the findings are driven by rejection concerns, since unattractive rank-givers are more likely to expect the probability of being matched to attractive users to be lower.

In sum, I find that star-rating effects are mainly driven by users who are less-attractive than average, and especially in cases where they are considering attractive potential partners. These findings provide additional evidence in support of the hypothesis that preference shading is driven by fear of rejection concerns.

4.6.6 Alternative Mechanisms

I now consider and rule out a few other alternative explanations for the results in §4.4 and §4.5.

First, one possible explanation for the negative effect of three-stars during the game is the salience effect. Since most users are shown with two-stars (see Table B.6 in Appendix B.1.4 for the distribution of stars in a game), three-star users may appear more salient and people may therefore pay more attention to them. However, this is unlikely to be the case because of two reasons. First, salience effect should also come into play for one-star users, but I see no significant effect for one-star users during the game. Second, usually demand increases with increased salience; however I see a negative effect for three-star users. Thus, it is unlikely that these results can be explained by the salience effect.

A second alternative explanation for the negative effect of higher stars during the game could be that users dislike popular users. However, the results show that three-star users are more likely to receive a reply to their first messages after the game. This implies that users receive higher utility from a conversation with a three-star user (i.e., Inequality (4.23)). Thus, I can rule out the explanation that users give lower preference-rankings to three-star users during the game because they dislike popular users.

Finally, a third possible mechanism for the negative effect of higher stars during the game could be the reference-point effect: when a user (rank-giver) see a potential partner with a higher popularity rating, s/he may set a higher reference-point for the rank-receiver. As such, that person is held to a higher standard (for attractiveness/appeal) and if they do not match up to that reference point, a loss component may be added to them. In other words, the rank-giver may not dislike popular users, but perceive them to be less appealing conditional on their popularity rating. Similar, to the discussion above, I can rule out this explanation because such behavioral biases are not supported by the fact that three-star users receive higher number of replies after the game.

Moreover, none of these alternative explanations are consistent with the heterogeneous

effects I found in §4.6.5.

4.6.7 Bounded Rationality: Limits of Strategic Thinking

Thus far, I have shown that players act rationally given their beliefs, i.e., conditional on their beliefs that popular players are hard to get, players respond strategically by shading their preferences for them. However, their beliefs may be mistaken. So, I now examine whether users' rejection concerns are rational.

Intuitively, if all the women (men) in a game lower their preference-ranking for a three-star man (woman), then the likelihood of being matched with him (her) should not be adversely affected. In other words, if a user can rationally infer that other players in the game may also be suffering from rejection concerns (and hence give lower preference-ranking to popular users), then they should recognize that popularity does not necessarily lead to a lower likelihood of match.

I can formalize this argument by expanding the expected probability that user j will be matched with user i, $\mathcal{P}$, from Equation (4.20) as:

$$\mathcal{P} = \int P\left(match_{ijt} = 1 pref_{ijt}, pref_{-ijt}, pref_{-jt}(star_{it})\right) g(pref_{-jt}) dg, \tag{4.25}$$

where $P\left(match_{ijt} = 1 pref_{ijt}, pref_{-ijt}, pref_{-jt}(star_{it})\right)$ denotes the probability that j will match with i conditional on $pref_{ijt}$ (the preference-ranking that user j gives to i), $pref_{-ijt}$ (the preference-ranking that user j gives to other players), and $pref_{-jt}$ (the preference-ranking that other users in the game give to everyone else, including i).[18] Although, j does not observe other players' preference-rankings, s/he may have beliefs about their distributions, which I denote by $g(pref_{-jt})$. Thus, I can integrate over these beliefs to obtain the expected match probability $\mathcal{P}$.

A key point to note here is that the expected probability of j being matched with i also depends on how other people in the game $(-j)$ rank i (denoted by $pref_{-jt}(star_{it})$). I know

[18]Note that $pref_{-ijt}$ and $pref_{-jt}$ should be ideally expressed as a function of the other players' star ratings, but I suppress them to keep the notation simple.

from the empirical findings in §4.4 that $pref_{-j't}(star_{it} = 3) < pref_{-j't}(star_{it} = 2)$ for all $j' \in -j$. Therefore, conditional on the preference-ranking that j gives to i, the rational expected probabilities of being matched with i when she is shown with three-stars should be higher than when she is shown with two-stars. However, in a fully rational world, other players will also recognize this effect, and in turn increase their preference-ranking for i. Thus, there cannot exist a Bayesian Nash equilibrium of this game where the average effect of star-rating on match probabilities is negative.

These results thus suggest that users do not fully internalize the idea that other players in the game may also suffer from rejection concerns, and therefore express lower preference for popular users. This may be because users' beliefs about the likelihood of matches with popular users are based on their observations in the offline world and/or their bounded rationality.

Indeed, these results are consistent with experimental findings on users' behavior in guessing games, where it has been shown that most users can only reason one step ahead. The classic example here is that of Keynesian beauty contest games, where all participants are asked to simultaneously pick a number between 0 and 100, and the winner is the player(s) whose pick is closest to p times the average of all numbers submitted, where p is some fraction. The Nash equilibrium of this game is (when all players are fully rational) is zero [46]. However, the guesses are far from zero in nearly all experimental settings [56, 8].

To explain these results, it has been argued that individuals reason in steps as follows [12]:

- Step 0: These individuals naively state their preferences without considering others' response.
- Step 1: These individuals think one step ahead, i.e., they believe that others are Step-0 players and best respond to that.
- Step 2: And so on....

The results in this study are consistent with Step 1 bounded-rationality, i.e., users believe that

others will naively reveal their preferences without taking rejection concerns into account, i.e., increase their ranking for popular users. Thus, the best response for Step-1 users is to reduce their own preference-ranking for popular users (given rejection concerns). The findings in this study are in line with other studies which also find that individuals in most experimental and real-life settings exhibit Step-1 bounded rationality [56, 75].

Next, I examine the extent of rationality in users' messaging behavior. I first quantify users' beliefs about the likelihood of receiving a response to a first message that she sends as a function of the star rating of the receiver. Intuitively, if a user is concerned about rejection, then s/he should be more willing to send first messages to those users who are more likely to respond. To examine if this is the case, I regress user j's likelihood of receiving a reply from i on i's star-rating (see Table B.8 in Appendix B.4). The results from this regression suggest that user j is more likely to receive a reply from one and three-star matches in response to her/his first messages. These response probabilities are consistent with j's likelihood of sending a message (model M5 in Table 4.5). Thus, when there is no game involved (i.e., users do not need to engage in multiple steps ahead reasoning), they seem to be able to form rational beliefs. Interestingly, this can also be interpreted as Step-1 thinking, and indeed, Step-1 thinking is fully-rational when no game is involved.

Chapter 5

CONCLUSION

5.1 Conclusion: A Gamified Online Weight-loss Community

In the first essay, I examine the causal effect of users' participation in online gamified challenges on their weight-loss progress. I empirically examine the effectiveness of gamified weight-loss challenges, using the data from a leading online weight-loss community in the Unites States. The findings indicate that participation in gamified challenges has a positive and significant effect on weight-loss. Users can achieve a weight-loss of 0.945 kg by participating in at least one challenge a month. Further, based on the results from the system GMM estimation, I calculate the long-term effect of participation in challenges, and find that a challenge participant will gain the lost weight back in a few months, if s/he does not participate in a new challenge in future. Further, I show that not all gamified weight-loss challenges are the same; challenges with a numeric weight-loss target are more effective than challenges without a target, controlling for the number of participants and number of instructions in the challenge.

The main contributions of this study are as follows. First, I quantify and isolate the effect of users' participation in gamified challenges on weight-loss from their engagement with other platform's features such as self-monitoring tools and weight-loss forums/groups. A large stream of research studies the effect of eHealth interventions on weight-loss using randomized trials [41]. However, in those studies, a combination of different online features is used and it is not possible to determine which features of eHealth weight-loss interventions are driving the effect on the weight-loss progress. Second, this is the first study to examine both performance and process goals in gamified weight-loss interventions. I showed that in a gamified weight-loss challenge where users get incentivized based on the performance

goal, a challenge with both performance and process goal is most effective. This finding can serve as a guideline for designing gamified information systems in goal-setting environments. Finally, I discuss and clarify the methodological strategies required to analyze the dynamic nature of the weight-loss outcome and overcome endogeneity problems in non-experimental settings. I compare the system GMM results with a difference-in-difference model with time-varying treatment, coupled with propensity score matching. Although the effect of challenges on weight-loss remains positive and significant in a difference-in-difference model, the magnitude of the effect is smaller and close to the magnitude of a biased fixed-effect model. Moreover, I show that a difference GMM dynamic approach performs poorly in the setting due to the high inertial effects of the lagged dependent variable.

This study raises interesting questions for future research. First, I show a positive short-term effect of challenges on weight-loss, and I calculate the long-term effect based on the results from the system GMM estimation. Prior studies have shown that adherence to short-term goals can breed long-term weight-loss achievement [83]. However, due to my short panel data set, I could not evaluate the effect of repeated participation in challenges on weight-loss progress over time. Investigating the effect of continuous engagement in gamified challenges on long-term weight-loss can be an interesting question for future research. Second, in my setting, I could not directly shed light on how leaderboards can incentivize challenge participants. Prior research shows that the information of best performers can affect individuals' weight-loss progress differently compared to the information of average or worst performers [82]. Future research can shed light on how leaderboards in online settings can encourage users by showing the ranking and progress information of specific users. Finally, in my data I cannot observe the network connection between users who participate in a challenge. The presence of online friends and their rankings in a weight-loss challenge might strengthen the effect of challenges. Further, participating in online challenges can provide an opportunity for users to get to know successful users, and follow them online even after the challenge. Having access to such online social network data, future research can provide design implications regarding recommendations or constraints on who can join a challenge.

5.2 Conclusion: A Gamified Online Dating Platform

In the second essay, I examine how users respond to the popularity information of potential partners in a mobile dating app. On the one hand, knowing that a potential partner is popular can increase her/his appeal. On the other hand, popular people are less likely to reciprocate, and hence users may strategically shade down their revealed preference for popular users to avoid the psychological costs of rejection. In this setting, users interact with each other by playing a ranking game, where they rank-order members of the opposite sex and are then matched based on a Stable Match Algorithm. A key piece of information shown to users during this process is a star-rating for each member of the opposite sex, which is a function of the past preference-rankings received. I quantify the causal impact of a user's star-rating on the preference-rankings that s/he receives during a game and her likelihood of receiving messages after a game. To overcome the endogeneity between a user's star-rating and her unobserved attractiveness, I employ non-linear fixed-effects models.

I find that, everything else being constant, compared to two-star users: (1) three-star users receive lower preference-rankings during the game, however, (2) three-star users are more likely to receive first and reply messages after the game. This heterogeneity across outcomes can be linked to the perceived severity of rejection concerns and can be interpreted as strategic shading. The perceived risk of rejection is highest during the game, since users have no information on the other person's preferences. In contrast, after the game, a focal user knows that the person who s/he has been matched with must have preferred her/him sufficiently high for them to be matched in the first place. This alleviates rejection concerns. I also show that the effect of star-ratings is heterogeneous across user-specific observables: users who are less-attractive than average, specifically when they are ranking attractive partners, shade their preferences for popular users. These findings are consistent with the hypothesis that shading is driven by fear of rejection since unattractive users are more likely to suffer from rejection concerns. Finally, I also find that users' beliefs regarding rejection during the game are not fully rational. Instead, the results suggest that users believe that

they are playing against naive users who reveal their preferences without considering others response.

This study makes three key contributions to the literature. It documents negative returns to higher star-ratings in online platforms. Second, it empirically establishes the presence of strategic shading in dating marketplaces. Third, it establishes that users exhibit bounded rationality in real-world online settings that involve strategic multi-player games.

These findings raise many interesting questions that can serve as avenues for future research. I study a platform where users are looking for short-term fun and flirtation rather than long-term partners or marriage. Moreover, in this setting, users only have a short amount of time to process information and make their decisions. It is not clear whether the results stem from the short-term nature of the relationship or from time pressure. Thus, I cannot comment on whether (and to what extent) these results would change if users were looking for long-term relationships/marriage and/or had more time to process information. Further research on these topics can have important implications for the design of dating and matching platforms.

BIBLIOGRAPHY

[1] Ahmed Allam, Zlatina Kostova, Kent Nakamoto, and Peter Johannes Schulz. The effect of social support features and gamification on a web-based intervention for rheumatoid arthritis patients: randomized controlled trial. *Journal of medical Internet research*, 17(1):e14, 2015.

[2] Paul D Allison. Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods & Research*, 28(2):186–208, 1999.

[3] Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

[4] Manuel Arellano and Olympia Bover. Another look at the instrumental variable estimation of error-components models. *Journal of econometrics*, 68(1):29–51, 1995.

[5] Ravi Bapna, Jui Ramaprasad, Galit Shmueli, and Akhmed Umyarov. One-Way Mirrors in Online Dating: A Randomized Field Experiment. *Management Science*, 62(11):3100–3122, 2016.

[6] Jeff E Biddle and Daniel S Hamermesh. Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre. *Journal of Labor Economics*, 16(1):172–201, 1998.

[7] Richard Blundell and Stephen Bond. Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, 87(1):115–143, 1998.

[8] Antoni Bosch-Domenech, Jose G Montalvo, Rosemarie Nagel, and Albert Satorra. One, Two,(Three), Infinity,...: Newspaper and Lab Beauty-Contest Experiments. *American Economic Review*, 92(5):1687–1701, 2002.

[9] Kelly D Brownell, Rita Yopp Cohen, Albert J Stunkard, MR Felix, and Nancy B Cooley. Weight loss competitions at the work site: impact on weight, morale and cost-effectiveness. *American Journal of Public Health*, 74(11):1283–1285, 1984.

[10] Joseph A Cafazzo, Mark Casselman, Nathaniel Hamming, Debra K Katzman, and Mark R Palmert. Design of an mhealth app for the self-management of adolescent type 1 diabetes: a pilot study. *Journal of medical Internet research*, 14(3):e70, 2012.

[11] Hongbin Cai, Yuyu Chen, and Hanming Fang. Observational Learning: Evidence from a Randomized Natural Field Experiment. *American Economic Review*, 99(3):864–82, 2009.

[12] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[13] Jessica J Cameron, Danu Anthony Stinson, and Joanne V Wood. The Bold and the Bashful: Self-Esteem, Gender, and Relationship Initiation. *Social Psychological and Personality Science*, 4(6):685–691, 2013.

[14] John Cawley and Chad Meyerhoefer. The medical care costs of obesity: an instrumental variables approach. *Journal of health economics*, 31(1):219–230, 2012.

[15] Gary Chamberlain. Analysis of Covariance with Qualitative Data. *The Review of Economic Studies*, 47(1):225–238, 1980.

[16] Gary Charness and Uri Gneezy. Incentives to exercise. *Econometrica*, 77(3):909–931, 2009.

[17] Yu Chen and Pearl Pu. Healthytogether: exploring social incentives for mobile fitness applications. In *Proceedings of the second international symposium of chinese chi*, pages 25–34. ACM, 2014.

[18] Yubo Chen, Qi Wang, and Jinhong Xie. Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning. *Journal of Marketing Research*, 48(2):238–254, 2011.

[19] Judith A Chevalier and Dina Mayzlin. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

[20] Eugene Choo and Aloysius Siow. Who Marries Whom and Why. *Journal of political Economy*, 114(1):175–201, 2006.

[21] Marcel Das and Arthur Van Soest. A Panel Data Model for Subjective Information on Household Income Growth. *Journal of Economic Behavior & Organization*, 40(4):409–426, 1999.

[22] Emely De Vet, Rob MA Nelissen, Marcel Zeelenberg, and Denise TD De Ridder. Aint no mountain high enough? setting high weight loss goals predict effort and short-term weight loss. *Journal of health psychology*, 18(5):638–647, 2013.

[23] Gabrielle Demange, David Gale, and Marilda Sotomayor. A Further Note on the Stable Matching Problem. *Discrete Applied Mathematics*, 16(3):217–222, 1987.

[24] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15. ACM, 2011.

[25] Sanjeev Dewan, Yi-Jen Ho, and Jui Ramaprasad. Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community. *Information Systems Research*, 28(1):117–136, 2017.

[26] Wenjing Duan, Bin Gu, and Andrew B Whinston. Informational Cascades and Software Adoption on the Internet: An Empirical Investigation. *MIS quarterly*, pages 23–48, 2009.

[27] Carol S Dweck and Ellen L Leggett. A social-cognitive approach to motivation and personality. *Psychological review*, 95(2):256, 1988.

[28] Paul W Eastwick and Eli J Finkel. Sex Differences in Mate Preferences Revisited: Do People Know What they Initially Desire in a Romantic Partner? *Journal of Personality and Social Psychology*, 94(2):245, 2008.

[29] Kristina Elfhag and Stephan Rössner. Who succeeds in maintaining weight loss? a conceptual review of factors associated with weight loss maintenance and weight regain. *Obesity reviews*, 6(1):67–85, 2005.

[30] Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. *The Quarterly Journal of Economics*, 121(2):673–697, 2006.

[31] Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. Racial Preferences in Dating. *The Review of Economic Studies*, 75(1):117–132, 2008.

[32] Katherine M Flegal, Margaret D Carroll, Cynthia L Ogden, and Lester R Curtin. Prevalence and trends in obesity among us adults, 1999-2008. *Jama*, 303(3):235–241, 2010.

[33] North American Association for the Study of Obesity, National Heart, Lung, Blood Institute, and NHLBI Obesity Education Initiative. *The practical guide: identification, evaluation, and treatment of overweight and obesity in adults*. National Institutes of Health, National Heart, Lung, and Blood Institute , 2000.

[34] Erin Fothergill, Juen Guo, Lilian Howard, Jennifer C Kerns, Nicolas D Knuth, Robert Brychta, Kong Y Chen, Monica C Skarulis, Mary Walter, Peter J Walter, et al. Persistent metabolic adaptation 6 years after the biggest loser competition. *Obesity*, 24(8):1612–1619, 2016.

[35] David Gale and Lloyd S Shapley. College Admissions and the Stability of Marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[36] Jessica Greene, Rebecca Sacks, Brigitte Piniewski, David Kil, and Jin S Hahn. The impact of an online social network with wireless monitoring devices on physical activity and weight loss. *Journal of primary care & community health*, 4(3):189–194, 2013.

[37] Sally L Hansen. Dating Choices of High school Students. *Family Coordinator*, pages 133–138, 1977.

[38] Judith M Harackiewicz, Kenneth E Barron, Suzanne M Carter, Alan T Lehto, and Andrew J Elliot. Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social psychology*, 73(6):1284, 1997.

[39] Günter J Hitsch, Ali Hortaçsu, and Dan Ariely. What Makes You Click?Mate Preferences in Online Dating. *Quantitative Marketing and Economics*, 8(4):393–427, 2010.

[40] Arne Risa Hole, Andy Dickerson, Luke Munford, et al. A Review of Estimators for the Fixed-Effects Ordered Logit Model. In *United Kingdom Stata Users' Group Meetings 2011*, number 05. Stata Users Group, 2011.

[41] MJ Hutchesson, ME Rollo, Rebecca Krukowski, Louisa Ells, J Harvey, PJ Morgan, R Callister, R Plotnikoff, and CE Collins. Ehealth interventions for the prevention and treatment of overweight and obesity in adults: A systematic review with meta-analysis. *Diabetes Technology and Therapeutics*, 18:S67, 2016.

[42] Kevin O Hwang, Allison J Ottenbacher, Angela P Green, M Roseann Cannon-Diehl, Oneka Richardson, Elmer V Bernstam, and Eric J Thomas. Social support in an internet weight loss community. *International journal of medical informatics*, 79(1):5–13, 2010.

[43] IBISWorld. Dating Services Industry in the US - Market Research Report, 2019. https://www.ibisworld.com/industry-trends/market-research-reports/other-services-except-public-administration/personal-laundry/dating-services.html.

[44] Kosuke Imai, In Song Kim, and Erik Wang. Matching methods for causal inference with time-series cross-section data. *Princeton University*, 1, 2018.

[45] Guido Imbens and Jeffrey Wooldridge. whats new in econometrics lecture 1. 2007.

[46] John Maynard Keynes. *The General Theory of Employment, Interest, and Money*. 1936.

[47] Anna Khaylis, Themis Yiaslas, Jessica Bergstrom, and Cheryl Gore-Felton. A review of efficacious technology-based weight-loss interventions: five key components. *Telemedicine and e-Health*, 16(9):931–938, 2010.

[48] Robert Kurzban and Jason Weeden. HurryDate: Mate Preferences in Action. *Evolution and Human Behavior*, 26(3):227–244, 2005.

[49] SangMok Lee. Incentive Compatibility of Large Centralized Matching Markets. *Working Paper*, 2016.

[50] SangMok Lee and Leeat Yariv. On the Efficiency of Stable Matchings in Large Markets. *Working Paper*, 2018.

[51] Soohyung Lee. Effect of Online dating on Assortative Mating: Evidence from South Korea. *Journal of Applied Econometrics*, 31(6):1120–1139, 2016.

[52] De Liu, Radhika Santhanam, and Jane Webster. Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS quarterly*, 41(4), 2017.

[53] Edwin A Locke and Gary P Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705, 2002.

[54] Jennifer Ludden. Do You Like Me?, 2016. https://www.npr.org/sections/thetwo-way/2016/02/11/466342716/do-you-like-me-swiping-leads-to-spike-in-online-dating-for-young-adults.

[55] Enrico Moretti. Social Learning and Peer Effects in Consumption: Evidence from Movie Sales. *The Review of Economic Studies*, 78(1):356–393, 2011.

[56] Rosemarie Nagel. Unraveling in Guessing Games: An Experimental Study. *The American Economic Review*, 85(5):1313–1326, 1995.

[57] Melissa A Napolitano, Sharon Hayes, Gary G Bennett, Allison K Ives, and Gary D Foster. Using facebook and text messaging to deliver a weight loss program to college students. *Obesity*, 21(1):25–31, 2013.

[58] Rob MA Nelissen, Emely de Vet, and Marcel Zeelenberg. Anticipated emotions and effort allocation in weight goal striving. *British journal of health psychology*, 16(1):201–212, 2011.

[59] M Neve, Philip J Morgan, PR Jones, and CE Collins. Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: a systematic review with meta-analysis. *Obesity reviews*, 11(4):306–321, 2010.

[60] Jerzy Neyman and Elizabeth L Scott. Consistent Estimates Based on Partially Consistent Observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[61] Stephen Nickell. Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426, 1981.

[62] Erin S Pearson. Goal setting as a health behavior change strategy in overweight and obese adults: a systematic literature review examining intervention components. *Patient education and counseling*, 87(1):32–42, 2012.

[63] Boris Pittel. The Average Number of Stable Matchings. *SIAM Journal on Discrete Mathematics*, 2(4):530–549, 1989.

[64] Rishika Rishika, Ashish Kumar, Ramkumar Janakiraman, and Ram Bezawada. The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information systems research*, 24(1):108–127, 2013.

[65] David Roodman. How to do xtabond2: An introduction to difference and system gmm in stata. *The stata journal*, 9(1):86–136, 2009.

[66] David Roodman. A note on the theme of too many instruments. *Oxford Bulletin of Economics and statistics*, 71(1):135–158, 2009.

[67] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[68] Alvin E Roth. Two-Sided Matching with Incomplete Information About Others' Preferences. *Games and Economic Behavior*, 1(2):191–209, 1989.

[69] Alvin E Roth and Marilda A Oliveira Sotomayor. *Two-sided matching*. Cambridge University Press, Cambridge, 1990.

[70] Valeriya Safronova. An Inside Look at Your Favorite Dating Sites, 2018. https://www.nytimes.com/2018/04/11/style/match-shaadi-league-farmersonly-dating-apps.html .

[71] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, 2006.

[72] Ali Shameli, Tim Althoff, Amin Saberi, and Jure Leskovec. How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 455–463. International World Wide Web Conferences Steering Committee, 2017.

[73] Aaraon Smith. 15% of American Adults Have Used Online Dating Sites or Mobile Dating Apps. *Pew Research Center*, pages 1–12, 2016. https://www.pewinternet.org/2016/02/11/15-percent-of-american-adults-have-used-online-dating-sites-or-mobile-dating-apps/fnref-15504-1.

[74] Alan T Sorensen. Bestseller Lists and Product Variety. *The Journal of Industrial Economics*, 55(4):715–738, 2007.

[75] Dale O Stahl and Paul W Wilson. On Players Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, 10(1):218–254, 1995.

[76] Dori M Steinberg, Deborah F Tate, Gary G Bennett, Susan Ennett, Carmen Samuel-Hodge, and Dianne S Ward. The efficacy of a daily self-weighing weight loss intervention using smart scales and e-mail. *Obesity*, 21(9):1789–1797, 2013.

[77] Victor J Strecher, Gerard H Seijts, Gerjo J Kok, Gary P Latham, Russell Glasgow, Brenda DeVellis, Ree M Meertens, and David W Bulger. Goal setting as a strategy for health behavior change. *Health education quarterly*, 22(2):190–200, 1995.

[78] Albert J Stunkard, Rita Yopp Cohen, and Michael RJ Felix. Weight loss competitions at the worksite: how they work and how well. *Preventive medicine*, 18(4):460–474, 1989.

[79] Monic Sun. How Does the Variance of Product Ratings Matter? *Management Science*, 58(4):696–707, 2012.

[80] Alain Trognon. Miscellaneous asymptotic properties of ordinary least squares and maximum likelihood estimators in dynamic error components models. In *Annales de l'INSEE*, pages 631–657. JSTOR, 1978.

[81] Catherine Tucker and Juanjuan Zhang. How Does Popularity Information Affect Choices? A Field Experiment. *Management Science*, 57(5):828–842, 2011.

[82] Kosuke Uetake and Nathan Yang. Inspiration from the biggest loser: Social interactions in a weight loss program. *Marketing Science*, 2019.

[83] Kosuke Uetake, Nathan Yang, et al. Success breeds success: Weight loss dynamics in the presence of short-term and long-term goals. Technical report, 2017.

[84] Kevin G Volpp, Leslie K John, Andrea B Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein. Financial incentive–based approaches for weight loss: a randomized trial. *Jama*, 300(22):2631–2637, 2008.

[85] Frank Windmeijer. A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of econometrics*, 126(1):25–51, 2005.

[86] Dawn Winters and Gary P Latham. The effect of learning versus outcome goals on a simple versus a complex task. *Group & Organization Management*, 21(2):236–250, 1996.

[87] Linda Y Wong. Structural Estimation of Marriage Models. *Journal of Labor Economics*, 21(3):699–727, 2003.

[88] Jeffrey M Wooldridge. Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139, 2002.

[89] Kaiquan Xu, Jason Chan, Anindya Ghose, and Sang Pil Han. Battle of the channels: The impact of tablets on digital commerce. *Management Science*, 63(5):1469–1492, 2016.

[90] Hema Yoganarasimhan. Impact of Social Network Structure on Content Propagation: A Study using YouTube Data. *Quantitative Marketing and Economics*, 10(1):111–150, 2012.

[91] Hema Yoganarasimhan. Impact of social network structure on content propagation: A study using youtube data. *Quantitative Marketing and Economics*, 10(1):111–150, 2012.

[92] Hema Yoganarasimhan. The Value of Reputation in an Online Freelance Marketplace. *Marketing Science*, 32(6):860–891, 2013.

[93] Jessica Yu. Search, Selectivity, and Market Thickness in Two-Sided Markets. *Working Paper*, 2018.

[94] Juanjuan Zhang. The Sound of Silence: Observational Learning in the US Kidney Market. *Marketing Science*, 29(2):315–335, 2010.

[95] Juanjuan Zhang and Peng Liu. Rational Herding in Microloan Markets. *Management Science*, 58(5):892–912, 2012.

[96] Xiaoquan Michael Zhang and Feng Zhu. Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, 101(4):1601–15, 2011.

Appendix A

A GAMIFIED ONLINE WEIGHT-LOSS COMMUNITY

A.1 The Difference-in-Difference Approach

In a difference-in-difference approach, the estimation equation for user i in month t is:

$$weightLoss_{it} = \delta_1 adopt_i + \delta_2 after_t + \delta_3 adopt_i \times after_t + X_{it}\Phi + m_t + e_{it} \qquad \text{(A.1)}$$

where, $weightLoss_{it}$ is user is average weight-loss (kg) at month t. The covariate $adopt_i$ is a binary variable indicating if user i ever participates in at least one challenge, and it controls for the time-invariant differences between challenge adopters and non-adopters. The covariate $after_t$ denotes the time period after the introduction of challenges on the platform, and it controls for the potential temporal factors that may simultaneously affect potential challenge adoption and weight-loss outcomes. The interaction term, $adopt_i \times after_t$ captures the difference-in-difference estimator which shows how the adopters weight-loss outcome is affected after the introduction of challenges on the platform. The variables X_{it} and m_t are the same variables that I discussed in §3.3.1. As shown in model A1, Table A.1, the estimated coefficient of the interaction term δ_3 is reported as positive but insignificant.

Note that in this analysis, there are 253 adopters and 792 non-adopters. As earlier illustrated in Table 3.4, the observed characteristics of these adopters and non-adopters before the introduction of challenges on the platform are statistically different. Next, I employ propensity score matching (PSM) to reduce the potential differences across adopters and non-adopters, and I form a sample of adopters and non-adopters who are similar in terms of the pretreatment observed characteristics. I use a static matching approach in which the matched pool of non-adopters remains the same for all months within the study period. Propensity score is defined as the conditional probability of receiving the treatment

[67]. Thus, propensity score is calculated as:

$$propensityScore_i = Pr(adopt_i = 1|X_i = x) \tag{A.2}$$

where X_i is the set of user-specific pretreatment variables. I use the psmatch2 Stata command to match pairs. This command executes a probit model to estimate the propensity score (see equation A.4). The results of this probit model are summarized in Table A.3. Using the propensity scores, and nearest neighbor matching, I match adopters and non-adopters with common support. Now, as summarized in Table A.2, there are 247 adopters and 247 non-adopters with similar pretreatment attributes.

Next, I run the difference-in-difference model on the matched sample. As shown in model A2, the estimated coefficient of the interaction term δ_3 is reported as positive but insignificant. It is necessary to emphasize that in this setting, the treatment or participation in challenges is time varying. In other words, although challenges are introduced at month $t = 5$, users can participate in challenges at different points of time. Therefore, instead of comparing adopters (and non-adopters) before and after the challenge introduction, I implement a difference-in-difference analysis with time-varying treatment coupled with propensity score matching, following [89, 64]. In this approach, the estimating equation for user i in month t is:

$$weightLoss_{it} = \delta_1' adopt_i + \delta_2' participate_{it} + \delta_3' adopt_i \times participate_{it} + X_{it}\Phi + m_t + e_{it} \tag{A.3}$$

where, $participate_{it}$ is a binary variable indicating the user i's challenge-participation period for each treated user and the matched untreated user. The interaction term shows how the weight-loss outcome of the adopters change while participating in a challenge in contrast to that of control group in the same period.

As shown in model A3, using the difference-in-difference model with time-varying treatment, coupled with propensity score matching, the estimated coefficient of the interaction term δ_3' is reported as positive and significant. The magnitude of the difference-in-difference estimation in model A3 is very similar to the fixed effects model reported in Table 3.5. For a detailed comparison of matching estimators and fixed effects estimators, see [44].

A.2 The Set of Assumptions in System GMM

The validity of the system GMM estimates relies on satisfying a set of assumptions. The set of assumptions in a dynamic panel model with the dependent variable y_{it}, and endogenous variable x_{it}, using a system GMM approach are the followings:

1. $|\alpha| < 1$: This assumption points to the stationarity of the dynamic process.

2. $e_{it} \sim iid(0, \sigma_i^2) \quad \forall i, t$: This assumption allows e_{it} to be mean-zero, and uncorrelated across all i and across all t (no serial correlation). The validity of assuming no serial correlation is tested and verified using Arellano-Bond test [3]. Also, errors e_{it} can be heteroskedastic across individuals (σ_i^2). I can ensure the robustness of the results to this heteroscedasticity using a two-step system GMM.

3. $\eta_i \sim iid(0, \sigma_\eta^2) \quad \forall i$: This assumption requires the fixed effects η_i to be mean-zero and uncorrelated across individuals.

4. $E(e_{it}.\eta_i) = 0 \quad \forall i, t$: This assumption requires the e_{it} and η_i to be uncorrelated across all i and t.

5. $E(e_{it}.x_{is}) \neq 0 \quad \text{if} \quad t \leq s \quad \forall i, t, s$: This assumption shows that current shocks e_{it} can influence current and future x_{is}.

6. $E(e_{it}.x_{is}) = 0 \quad \text{if} \quad t > s \quad \forall i, t, s$: This assumption shows that current shocks e_{it} are uncorrelated with previous x_{is}.

7. $E(x_{it}.\eta_i) \neq 0 \quad \forall i, t$: This assumption allows user time-varying characteristics to be correlated with user unobserved fixed effects η_i.

8. $E(z_i.\eta_i) \neq 0 \quad \forall i, t$: This assumption shows that user observed time-invariant characteristics z_i can be correlated with user unobserved fixed effects η_i.

9. $y_{i1} = \frac{\eta_i}{1-\alpha} + u_{i1} \quad \text{where} E(u_{i1}.\eta_i) = 0 \quad \forall i$

Assumption (9) is a form of stationarity assumption on the initial conditions y_{i1}. This initial condition requires the deviation u_{i1} of the initial observation to be uncorrelated with the fixed effects η_i. This assumption implies that in the initial period, users with larger fixed effects are not systematically further or closer to their steady states than those with smaller fixed effects [91]. All these assumptions are required to have valid IVs. I examine and show the validity of the group of IVs by using Hansen test.

A.3 The Estimation of the Probability of Adoption

In order to employ IPW, I define $adopt_i$ indicating whether user i is a challenge adopter, and I use a probit model to calculate the probability of adoption based on users' observed variables prior to the introduction of challenges, using the following model:

$$\begin{aligned} adopt_i = c_0 + c_1 avgNumReoprt_i + c_2 avgJournal_i + c_3 avgGeneralForum_i \\ + c_4 avgGroupForum_i + c_5 tenure_i 1 + c_6 goal_i + \epsilon_{it}. \end{aligned} \tag{A.4}$$

As shown in Table A.3, Ifind that users with high engagement in the platform (those who report their weight frequently and post in journals and group forums) are more likely to adopt challenges.

	(A1)	(A2)	(A3)
	Diff-in-Diff	Diff-in-Diff + PSM	Diff-in-Diff + PSM
Variable			(time-varying treatment)
$adopt_i$	-0.139	-0.135	-0.164
	(0.131)	(0.154)	(0.134)
$after_t$	0.008	-0.305	-
	(0.136)	(0.199)	
$adopt_i \times after_t$	0.026	0.157	-
	(0.152)	(0.180)	
$participate_{it}$	-	-	0.170
			(0.165)
$adopt_i \times participate_{it}$	-	-	0.314*
			(0.189)
Controls	✓	✓	✓
Constant	0.611***	0.767***	0.315*
	(0.104)	(0.155)	(0.190)
Observations	3159	1656	1656
Individuals	1045	494	494
Adopters	253	247	247
Nonadopters	792	247	247

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.1: Difference-in-Difference estimates.

	Mean			Mean (matched sample)		
	adopter	non-adopter	$p > \|t\|$	adopter	non-adopter	$p > \|t\|$
Variables	(n=253)	(n=792)		(n=247)	(n=247)	
$avgNumReoprt_i$	0.883	0.755	0.001	0.861	0.829	0.553
$avgJournal_i$	0.431	0.176	0.000	0.390	0.337	0.342
$avgGeneralForum_i$	0.287	0.099	0.000	0.248	0.210	0.417
$avgGroupForum_i$	0.016	0.001	0.000	0.006	0.003	0.241
$tenure_{i1}$	5.549	4.879	0.029	5.551	6.182	0.106
$goal_i$	68.399	68.648	0.785	68.436	69.128	0.572

Table A.2: The comparison of adopters and non-adopters before introduction of challenges.

Variables	Coef.	Std. Err.
$avgNumReoprt_i$	-0.038	0.092
$avgJournal_i$	0.417***	0.119
$avgGeneralForum_i$	0.164	0.140
$avgGroupForum_i$	4.715***	1.727
$tenure_{i1}$	0.028***	0.010
$goal_i$	-0.001	0.003
constant	-0.888***	0.251
observations	1045	
R^2	0.056	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: The estimation of the probability of adoption using a probit model.

Appendix B

A GAMIFIED ONLINE DATING PLATFORM

B.1 Robustness Checks

B.1.1 Effect of Stars on Preference-Rankings - Linear Model

I consider the following linear model:

$$pref_{ijt} = \beta_1 star1_{it} + \beta_2 star3_{it} + \gamma z_i + \eta_i + \epsilon_{ijt} \tag{B.1}$$

The main difference of these coefficients and variables to what I discussed in §4.4.1 is that here they relate directly to the observed outcome instead of the latent variable $pref^*$. Hence, even though I use the same variable names for expositional convenience, the interpretation of the coefficients in the two models is different. In short, the magnitude of the coefficients from the two models cannot be directly compared.

There are two possible estimation strategies here: (1) pooled OLS, which ignores the problem of correlated unobservables, and (2) fixed-effects model, which addressed the omitted variable bias due to η_i by employing a "within" transformation to subtract out the time-invariant user-specific variables.

A pooled OLS estimation strategy consists of pooling all the data across games and users, and simply running a multiple regression on this data. I consider two pooled OLS models – (1) a simple model that only includes star-rating variables as the independent variables, and (2) a slightly more elaborate model that includes all the user-specific observables (z_i). The results from both these models are shown in models A1 and A2 in Table B.1.

Next, I discuss the fixed-effects estimation approach. Here, I start with the following averaging equation for each user i:

$$\overline{pref}_i = \beta_1 \overline{star1}_i + \beta_2 \overline{star3}_i + \gamma z_i + \eta_i + \bar{\epsilon}_i, \tag{B.2}$$

where $\overline{pref}_i = \frac{\sum_{t=1}^{T_i}\sum_j pref_{ijt}}{4\times T_i}$, $\overline{star1}_i = \frac{\sum_{t=1}^{T_i} star1_{it}}{T_i}$, $\overline{star3}_i = \frac{\sum_{t=1}^{T_i} star3_{it}}{T_i}$, and $\bar{\epsilon}_i = \frac{\sum_{t=1}^{T_i}\sum_j \epsilon_{ijt}}{4\times T_i}$. z_i, η_i are constant across time periods, and hence their averages are the same as the variables themselves. Next, I subtract Equation (B.2) from Equation (B.1) as follows:

$$pref_{ijt} - \overline{pref}_i = \beta_1 \left(star1_{it} - \overline{star1}_i\right) + \beta_2 \left(star3_{it} - \overline{star3}_i\right) + (\epsilon_{ijt} - \bar{\epsilon}_i) \tag{B.3}$$

Note that all the time-invariant user-specific variables are now subtracted out and the new error term, $\epsilon_{ijt} - \bar{\epsilon}_i$, is no longer correlated with the star-rating variables. The fixed-effects estimator is essentially a pooled OLS estimator for Equation (B.3) and it gives us consistent estimates of β_1 and β_2 under the linearity assumption. The results from this model are shown in model A3 in Table B.1. Note that to keep the comparisons consistent, I only use the first 100 games of users who saw at least one star change during the observation period. Hence, model A3 in analogous to model M3 in Table 4.4.

B.1.2 Estimation Sample

I present two validation checks to confirm that the substantive results in model M3, Table 4.4 are not driven by the estimation sample (which consists of users who saw at least one star change during the observation period).

First, I run the pooled ordered logit model on the subset of users who go through at least one star change and present the results in Table B.2. I find that the magnitude and the direction of the estimates in model A4 are similar to those for the full population model M1.

Second, I compare the distribution of user-specific observables for two groups of users – (1) users who saw no star change during the observation period, and (2) users who saw at least one star change in the observation period. I find that users who go through at least one star change are more likely to be new users who joined the app recently and a vast majority of them had not played any games at the start of the observation period. In contrast, users who do not see a star change are experienced users who had played a large number of games in the past. It is important to note that this difference in past experience does not reflect inherent differences in users, i.e., differences on user characteristics. Rather, it captures the

	(A1) (OLS)	(A2) (OLS)	(A3) (FE)
$star1_{it}$	-0.08946***	-0.07038***	0.01776
	(0.01422)	(0.01838)	(0.01126)
$star3_{it}$	0.03779***	0.03446**	-0.03100***
	(0.00971)	(0.01443)	(0.00913)
Controls		✓	
Constant	2.50031***	2.50513***	2.50006***
	(0.00113)	(0.00948)	(0.00034)
Individuals	24393	11639	3494
Observations	2980148	1580848	630160
R-Squared	0.00003	0.00260	0.00002

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Controls in Model A2 include: age_i, $college_i$, $graduate_i$, pic_score_i, num_pic_i, $employment_i$, and bio_i.

Table B.1: Pooled OLS and fixed-effects estimates of the effect of user's star-rating on preference-rankings received. All standard errors are clustered at the user-level.

dynamics of star-ratings. As users play more games, the marginal impact of a new game on their average popularity score is small. Thus, users who have played more games are less likely to experience a star change than new users.

I illustrate the point using Figure B.1, which shows how the change in users' popularity ($popularity_{it} - popularity_{it-1}$) varies as a function of the number of games played ($total\_game_{it}$). Recall that $popularity_{it}$ is simply the average of preference-rankings received by i in all her/his prior $t-1$ games. For the average user, after fifteen games, the expected change in popularity reduces to 0.03. This is simply due to the Law of Large Numbers –

	(A4)
$star1_{it}$	-0.13888***
	(0.02369)
$star3_{it}$	0.05136***
	(0.01590)
μ_2	-1.09054***
	(0.00456)
μ_3	-0.00036
	(0.00420)
μ_4	1.09067***
	(0.00446)
Individuals	3494
Observations	630160

Table B.2: Ordered logit estimates of the effect of star-rating on preference-rankings received (without fixed-effects), for a subset of users who experienced a star-change.

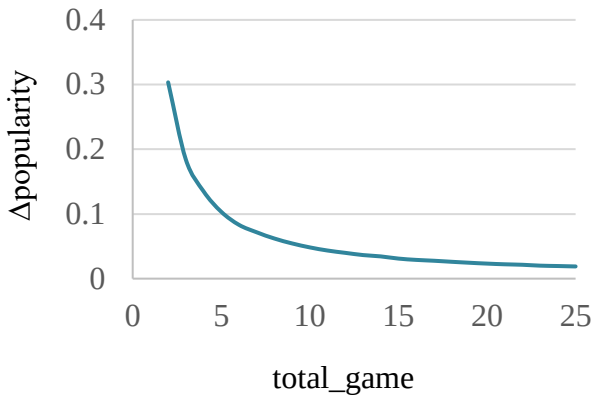

Figure B.1: Change in popularity as a function of number of games played.

for any user i with a set of characteristics z_i, η_i, the popularity measure ($popularity_{it}$) starts converging to a constant value after a few games. Thus, the variation in the number of star-changes a user experiences in the observation period is simply a function of whether

s/he is new to the app or not (and is not driven by her/his attributes).

Next, for users who are new in the app, I find that the two sets of users who go through at least one star change and those who do not experience any star-rating change are very similar on their user-specific variable. The results from this comparison are presented in Table B.3. Overall, there is sufficient empirical evidence to suggest that users who experience at least one star change and those who experience no star changes are similar on many important dimensions. Moreover, the pooled ordered logit estimates for the two subgroups are also similar. Thus, I expect the findings from the fixed-effects model can be interpreted as being largely applicable to the full population of users in the app.

Variables	Star Change	Mean	Std. Dev	Size	$Pr(\|T\| > \|t\|)$
age_i	No	21.950	7.393	2300	0.4487
	Yes	22.113	7.563	2538	
bio_i	No	56.909	168.424	2715	0.368
	Yes	53.045	152.914	2920	
$education_i$	No	1.737	0.512	2420	0.083
	Yes	1.712	0.510	2595	
$employment_i$	No	1.777	1.295	1614	0.758
	Yes	1.791	1.333	1727	
num_pic_i	No	5.355	1.393	2629	0.665
	Yes	5.338	1.426	2828	
pic_score_i $(Male)$	No	-0.092	0.635	1246	0.296
	Yes	-0.066	0.643	1296	
pic_score_i $(Female)$	No	-0.021	0.682	1066	0.077
	Yes	0.031	0.737	1246	

Table B.3: Comparison of attributes between new users who experienced no star change and groups who experienced at least one star change.

	(A5)
$star1_{it}$	0.13535**
	(0.05499)
$star3_{it}$	-0.25426***
	(0.04068)
Individuals	1684
Observations	50668

Table B.4: Ordered logit fixed-effects estimates of the effect of star-rating on preference-rankings received, for a subset of users who initiated a message at least once.

B.1.3 Within Game Correlation

	(A6)
$star1_{it}$	-0.01546
	(0.02713)
$star3_{it}$	-0.07380***
	(0.02135)
Individuals	3430
Observations	248,944

Table B.5: Ordered logit fixed-effects estimates of the effect of star-rating on preference-rankings received, for a subset of games with one competitor who experienced a star change.

B.1.4 Table of Game-level Star Configurations

star2	star1	star3	Games
4	-	-	174,818
3	-	1	4,273
3	1	-	2,606
2	1	1	113
2	-	2	73
2	2	-	24
1	-	3	3
1	1	2	3
1	2	1	1

Table B.6: The star configuration of four competitors in a game.

B.2 Conditional Log Likelihood for the Fixed-effects Logit Model

To study the relationship between the users' messaging behavior with their star-ratings, I consider the following fixed-effects logit formulations:

$$y_{ijt} = \begin{cases} 1, & y^*_{ijt} > 0 \\ 0, & \text{else} \end{cases}$$

where y_{ijt} is a binary variable and it can refer to $first_{ijt}$ or $reply_{ijt}$, and y^*_{ijt} is the corresponding latent variable as follows:

$$y^*_{ijt} = \beta_1 star1_{it} + \beta_2 star3_{it} + \gamma z_i + \eta_i + \epsilon_{ijt}, \tag{B.4}$$

I allow for η_i to be arbitarily correlated to $star1_{it}$ and $star3_{it}$. Further, I assume that $star1_{it}$, $star3_{it}$ and η_i are independent of ϵ_{ijt} since users are randomly assigned to games. Assuming that ϵ_{ijt}s are IID and drawn from an Extreme Value Type I distribution, I can write:

$$Pr(y_{ijt} = 1 star1_{it}, star3_{it}, z_i, \eta_i, \beta_1, \beta_2) = \frac{exp(\beta_1 star1_{it} + \beta_2 star3_{it} + \gamma z_i + \eta_i)}{1 + exp(\beta_1 star1_{it} + \beta_2 star3_{it} + \gamma z_i + \eta_i)} \tag{B.5}$$

I can now write the log-likelihoods of y_{ijt} (the first messages or replies) observed in the data as:

$$LL(\beta_1, \beta_2, \gamma) \;=\; \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{k=0}^{1} \ln\left[Pr(y_{ijt} = kstar1_{it}, star3_{it}, z_i, \eta_i, \beta_1, \beta_2)^{I(y_{ijt}=k)}\right] \quad \text{(B.6)}$$

where N is the total number of users and T_i is the total number of games played by user i. Treating the η_i's as parameters and maximizing this log-likelihood via Maximum Likelihood Estimator (MLE) is inconsistent with large N and finite T due to the well-known incidental parameters problem [60]. As a result, the estimate of β_1, β_2 from this approach will be inconsistent. However, Chamberlain proposes a method to maximize a Conditional Log-Likelihood which gives consistent estimates [15]. Following Chamberlain, I denote s_i as the sum of all received messages (first messages or reply messages) by user i from his/her matches over time, that is:

$$s_i = \sum_{t=1}^{T_i} (y_{ijt} match_{ijt} = 1) \quad \text{(B.7)}$$

and, I denote B_i as the set of all possible vectors of length T_i with s_i elements equal to 1, and $T_i - s_i$ elements equal to 0, i.e. all possible ways that user i could receive s_i messages in total over T_i games, that is:

$$B_i = \{d \in \{0,1\}^{T_i} \sum_{t=1}^{T_i} (d_{jt} = s_i match_{ijt} = 1)\} \quad \text{(B.8)}$$

For example, if user i plays three games ($T_i = 3$), and receives only one message in total ($s_i = 1$), B_i will be equal to $\{(1,0,0),(0,1,0),(0,0,1)\}$. Now, I can write the conditional probability of y_i given s_i as:

$$Pr\left(y_i star1_{it}, star3_{it}, s_i, \beta_1, \beta_2\right) = \frac{exp\left(y_i.\left(\beta_1 star1_{it} + \beta_2 star3_{it}\right)\right)}{\sum_{d \in B_i} exp\left(d.\left(\beta_1 star1_{it} + \beta_2 star3_{it}\right)\right)} \quad \text{(B.9)}$$

Note that this conditional probability does not depend on η_i's, i.e. s_i is a sufficient statistic for η_i. Thus, I can now specify a Conditional Log-Likelihood that is independent of η_is as shown below:

$$CLL(\beta_1, \beta_2) = \sum_{i=1}^{N}\sum_{t=1}^{T_i} \ln\left[Pr(y_i star1_{it}, star3_{it}, s_i, \beta_1, \beta_2)\right] \quad \text{(B.10)}$$

B.3 Validation Check: Truthfulness Assumption

In §4.6.3, I assume that users state their preference-rankings truthfully during the game. I now formally define this assumption and establish its validity in this setting.

Assumption 1. Truthfulness: I assume that the preference-ranking that user j gives to user i is higher than that she gives to i' during game t, if and only if $EU(pref_{ijt}) > EU(pref_{i'jt})$.

This assumption ensures that the relationship between users' latent expected utilities for any pair of potential partners is consistent with their stated preference-ordering over them. Thus, it allows me to take the empirical patterns observed in the data (from §4.4) and map them to the underlying expected utilities in §4.6.3. In particular, it allows me to take the empirical results established in §4.4 to and express the relative ordering of the underlying expected utilities in Inequality (4.22).

Truth-telling is not always guaranteed in SMPs, and it has been shown that some parties may have an incentive to misrepresent their true preferences depending on the game settings and the matching algorithm used [69].[1] For example, women may have incentive to misrepresent their true preferences if the platform uses a men-optimal stable matching algorithm; recall the discussion from §4.1.4.

In this setting, the ranking game resembles a one-to-one marriage SMP, where: (1) agents have to state their strict preference-rankings (i.e., no indifference rankings), (2) agents cannot truncate their list of preference-rankings (i.e., they cannot strategically choose to only rank their top few choices and refuse to rank their bottom choices), (3) agents cannot collude with each other, (4) agents' preferences are private (i.e,. users know their own preferences but not those of others'). Under such circumstances, it has been shown that, when a men-optimal stable matching mechanism is used, it is the dominant strategy for each man to state his true preferences, and any strategy for a woman is dominated if her stated first choice is not

[1] It has been shown that the incentive to manipulate true preferences is negligible in large markets [23, 63, 50, 49]. However, in this study, there are only four players in a game. So the results from the large-market literature may not generalize.

Match with	State true preferences	1^{st} and 2^{nd} preference misrepresentation	2^{nd} and 3^{rd} preference misrepresentation
	Assuming stated preferences are true preferences		
true 1^{st} choice	49.24	28.23	49.27
true 2^{nd} choice	28.25	49.28	14.90
true 3^{rd} choice	15.00	14.99	28.33
true 4^{th} choice	7.51	7.50	7.50
	Assuming true preferences are random		
true 1^{st} choice	49.48	28.15	49.47
true 2^{nd} choice	28.17	49.54	14.86
true 3^{rd} choice	14.90	14.89	28.21
true 4^{th} choice	7.45	7.42	7.46

Table B.7: Match results if users misrepresent their preferences.

her true first choice [68].[2] (And vice-versa for women-optimal stable matching mechanism.)

However, this platform does not use either a men-optimal or a women-optimal matching mechanism. Instead, as discussed in §4.1.4, it calculates the set of all possible stable matches and picks the matching with the highest average match-level. Under these conditions, there are no theoretical guarantees on truth-telling for any side of the market. Nonetheless, there are no obvious reasons for users to deviate from truth-telling. While I cannot theoretically prove this, I now empirically establish that, on average, users cannot gain by mis-representing their preferences in this setting.

To do so, I consider two types of deviation checks. In the top panel of Table B.7, I start with the assumption that a player's stated preferences are her/his true preferences. The second column represents the average probability of a player being matched with her/his true first, second, third, and fourth choices if the player ranks truthfully (based on the preference-rankings and match levels observed in the data). I find that truthful revelation

[2]The kind of stability studied in the case of incomplete information is ex post stability, i.e. a stable matching would remain stable even if all the preferences were to become common knowledge [68].

leads to being paired with the first choice 49.24% of the times, the second choice 28.25% of the times, the third choice 15.00% of the times, and the last choice 7.51% of the times. Next, I consider the following deviation: suppose that in game t, everyone except a focal player j plays the same strategy as that observed in the data, and j swaps her/his first and second choices. I then calculate which of her/his true preferences j will be matched with. Then, I aggregate the match outcomes over all players and all games to obtain the average probability of being matched with one's true first choice under this deviation as:

$$\Pr(\text{true first choice}) = \frac{\sum_{t=1}^{T} \sum_{j \in t} \mathrm{I}(match\_level_{jt} = \text{true first choice} | pref_{jt}^{12}, pref_{-jt})}{8T}, \tag{B.11}$$

where $pref_{jt}^{12}$ denotes a strategy where player j swaps her true first and second choices, and $pref_{-jt}$ denotes the preference-rankings observed in the data (i.e., others' strategies). Similarly, I also calculate the average probabilities of being matched with one's true second, third, and fourth choices.

The results from this simulation exercise are shown in the third column. Notice that misrepresenting preferences makes players strictly worse off. When a player ranks her true first choice as second, the probability of being matched with the true first choice drops to 28.23%. In the fourth column, I show the results from an analogous exercise, when a player misrepresents by swapping her second and third choices, i.e., plays $pref_{jt}^{23}$. Again, note that misrepresenting the preferences makes a player strictly worse off compared to truth-telling. Using similar simulations, it is possible to show that all other deviations also make players strictly worse off, compared to truthful revelation.

One possible critique of the above exercise could be that I started with the assumption that players stated preferences are their true preferences. Therefore, I also present results from a general case, where the deviating player's true preferences are drawn randomly (see the bottom panel of Table B.7). Again, I find that deviating from truth-telling makes users strictly worse off. In sum, all these tests confirm the validity of the truth-telling assumption in this setting.

Finally, note that there is no need to make any additional assumption on truth-telling for both *first* and *reply* messages since they are both single-agent decisions, and there is no game involved. Therefore, each player only has to follow her/his expected utilities and doesn't have to worry about the strategic behavior of other players. So, by definition, a player's revealed preferences reflect her/his expected utility.

B.4 Bounded Rationality

I now quantify the likelihood of that user j will receive a reply to her/his first message to a user i as a function of i's star-rating as follows:

$$reply_{jit} = \lambda_1 star1_{it} + \lambda_2 star3_{it} + \eta_i + \epsilon_{ijt} \tag{B.12}$$

The results from this regression are presented in Table B.8.

	(A6)
$star1_{it}$	0.12923^{***}
	(0.02933)
$star3_{it}$	0.09427^{***}
	(0.01662)
Constant	0.11961^{***}
	(0.00272)
Individuals	1972
Observations	9259
Standard errors in parentheses	
$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$	

Table B.8: The effect of matched partner's star-rating on probability of receiving a reply.