Effectiveness of gamified team competition as mHealth intervention for medical interns: a cluster micro-randomized trial

Jitao Wang[1], Yu Fang[2], Elena Frank[2], Maureen A Walton[2,3,4], Margit Burmeister[2,4], Ambuj Tewari[5], Walter Dempsey[1,6], Timothy NeCamp[7], Srijan Sen[2,3], Zhenke Wu[1]

Affiliations

1. Department of Biostatistics, University of Michigan, Ann Arbor, MI.
2. Michigan Neuroscience Institute, University of Michigan, Ann Arbor, MI.
3. Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI.
4. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI.
5. Department of Statistics, University of Michigan, Ann Arbor, MI.
6. Institute of Social Research, University of Michigan, Ann Arbor, MI.
7. Data Bloom Consulting LLC, Cincinnati, OH.

Corresponding author

Zhenke Wu, PhD.
Department of Biostatistics,
University of Michigan, Ann Arbor
1415 Washington Heights
Ann Arbor, MI, 48109-2029, USA
Phone: (734) 764-7067
E-mail: zhenkewu@umich.edu

ABSTRACT

Gamification, the application of gaming elements to increase enjoyment and engagement, has the potential to improve the effectiveness of digital health interventions, while the effectiveness of competition gamification components remains poorly understood. To address this gap, we evaluate the effect of smartphone-based gamified team competition intervention on daily step count and sleep duration via a micro-randomized trial. In 1,797 interns, competition intervention significantly increases the mean daily step count by 111.5 steps (SE 40.4, p=0.01) relative to the no competition arm, while competition does not significantly affect the mean daily sleep minutes (p=0.69). Moderator analyses indicates that, the causal effects of competition on daily step count and sleep minutes decrease by 9.1 (SE 11.6) steps (p=0.43) and 1.9 (SE 0.6) minutes (p=0.003) for each additional week-in-study, respectively. Intra-institutional competition negatively moderates the causal effect of competition upon daily step count by -114.9 (SE 93.7) steps (p=0.22). Our results shows that gamified team competition delivered via mobile app significantly increased daily physical activity which suggests that team competition can function as a mobile health intervention tool to increase short-term physical activity level. Future improvements on strategies of forming competition opponents and introducing occasional competition breaks may improve the overall effectiveness.

INTRODUCTION

Sufficient physical activity and sleep are associated with lower risk for numerous health conditions, including cardiovascular disease, obesity and depression[1–3]. However, only one in four US adults meets the recommended 150 minutes of moderate-intensity activity per week[4], and over one-third of US adults do not achieve the recommended seven hours of sleep per night[5,6].

Recent technical advances in wearable devices and mobile phones provide a new integrated platform to deliver interventions with minimal expense and user burden[7] with the additional advantage of temporal and spatial flexibility. Mobile devices can collect real-time and objective measurements of a user's physical activity and geographic location to provide personalized just-in-time adaptive interventions (JITAI)[8]. To date, many previous studies have shown the effectiveness of wearable and smartphone-based intervention on health outcomes[9–11]. Some studies included gamification, a strategy that attempts to enhance user enjoyment and engagement[12] by introducing game mechanics into a non-game environment[13–15]. Theories of health behavior change suggest that gamification elements that prompt self-monitoring, such as performance feedback, progress monitoring, and social comparison have the potential to motivate changes in behavioral outcomes[16,17].

Team competition is one such potential gamification strategy, however to our knowledge, its effectiveness at improving health behaviors has not been formally assessed.

Micro-randomized trials (MRT) can be used to address scientific questions about whether and under what circumstances JITAI components are effective, with the ultimate goal of developing effective and efficient JITAI[18–20]. In this study we conduct a cluster MRT using principles of health behavior change and gamification to deliver a mobile app-based weekly team competition to evaluate the effectiveness of this type of mHealth intervention on individual physical activity, sleep duration in the population of medical intern. We also explore the effectiveness of this mHealth intervention on user's engagement and individual self-reported mood score, inspired by previous work showing that increased sleep opportunity and physical

activity may improve individual's mood in this depression-vulnerable population[21,22]. Although we mainly focus on the effect of team competition on short-term (proximal) outcomes including step count and sleep minutes in this study, our ultimate goal is to improve interns' long-term (distal) mental health by increasing their short-term physical activity and sleep duration.

The study is conducted among a national cohort of first-year medical residents. Medical internship, a one-year-long physician training program, is highly stressful, which may lead to reduced health functioning and mental health symptoms. This study also assesses potential effect moderators, variables that increase or decrease the effectiveness of mHealth intervention, to inform future research incorporating personalized team-competition into mHealth intervention.

RESULTS

Study cohort

Between April 1, 2020, and June 16, 2020, a total of 4,791 incoming interns received the invitation email and 2,286 (47.7%) of interns enrolled in the study. Of those who enrolled, 84.7% (1,936/2,286) of participants could be grouped in a team with at least five interns and were included in the competition arm, with a total of 191 teams. These eligible competition-arm participants were randomized according to Figure 2. Of the 1,936 participants, 139 (7.2%) participants did not have any fitness tracker data available during the study and 18 (0.9%) did not have sufficient pre-internship survey data and baseline data, which were excluded from the analysis (see Figure 1 for details of subject inclusion). All remaining interns represented 90 residency institutions and 12 specialties. Among the 1779 (91.9%) participants included in the analysis, the mean age of the participants was 27.6 (SD 2.6), Males and females were nearly equally represented (54.5% female). See detailed demographic information in Table 1. Of the 1,779 medical interns who were eligible for intervention, all of them were assigned to the competition arm at least once during the study and the mean number of weeks a participant was in the competition arm was 5.8 (SD 1.9) weeks.

Main analysis

Main-effect analysis indicated that intervention of competing on step count had a significant positive causal effect on proximal daily step count compared to the non-competition arm (see detailed parameter estimates in Table 2). The number of daily steps increased by 111.5 (SE 40.4) steps for participants in the competition-step arm, compared to the non-competition arm (p=0.01). While no statistically significant effect on sleep duration was observed in response to competition on sleep. The estimate for the competition-sleep effect is -0.7 (SE 1.8) minutes (p=0.69).

Moderation analysis

Moderation analyses were performed by adding linear interaction terms between effect moderator and intervention into the model (see detailed parameter estimates in Table 2). A negative though non-significant association was observed between the number of weeks in the study and competition-step intervention (-9.1 steps/day; SE 11.6; p=0.43). That is, the moderation analysis (not the main-effect analysis) indicated that being in a competition-step week resulted in about 161.5 additional steps/day during the first week of the study, about 115.8 additional steps/day during the sixth week of the study, and about 61.0 additional steps/day during the twelfth week of the study. Similarly, a significantly negative interaction between the competition-sleep intervention and number of weeks in the study was identified: the causal effect of being in a competition-sleep week changed by -1.9 minutes/day (SE 0.6) with each additional week in the study (p=0.003); note that at the beginning of the study (the first week), the causal effect was significantly positive 9.9 minutes/day (SE 3.6, p=0.008), but then decreased. Plots of estimated causal effects of competition on proximal step count or sleep duration at different weeks and a sensitivity analysis to the linearity assumption (that the causal effect changes linearly by additional weeks in the study) was provided in Supplementary Figure 3. We also assessed whether the causal effect of competition upon step count or sleep minutes (relative to no competition) would vary by the opponent team being from the same or a different institution or specialty (see Supplementary Table 4). Non-significant intra-institution negative moderation (-114.8 steps/day; SE 93.7; p=0.22) and intra-specialty positive moderation (26.1 steps/day; SE 74.7; p=0.73) of causal effect of competition on step were observed. Similarly, no statistically significant intra-institution

(0.4 minutes/day; SE 3.2; p=0.90) or intra-specialty moderation (-1.9 minutes/day; SE 3.2; p=0.57) of causal effect of competition on sleep duration were observed.

Exploratory analysis

Exploratory analyses assessed the causal effect of being in a competition week on the proximal daily participation rates of step count and sleep minutes, averaged over all weeks (see Supplementary Table 5). The positive causal effect on daily participation rates of step count and sleep minutes were a 0.4% (SE 0.3%, p=0.13) and 0.9% (SE 0.3%, p=0.003) respectively. That is, if 1,779 participants were all in the competition week, there would be additional 50 (1,779 * 0.4% * 7) person-day records of step count and 112 (1,779 * 0.9% * 7) person-day records of sleep minutes recorded within this week, compared to a non-competition week. We also assessed whether the causal effect of team competition had a positive impact on team-averaged mood score. Non-significant positive effect (0.02 units/day, SE 0.02, p=0.35) of causal effect of team competition on mood score was observed (see Supplementary Table 6).

DISCUSSION

This study answered two questions: 1. Is gamified competition delivered via mobile app effective in the field of mHealth intervention? 2. If it is effective, how to personalize and optimize the efficacy of competition intervention? The main-effect analysis indicated that the gamified competition administered through smartphones can lead to increased proximal daily step count. Positive causal effect suggested inclusion of competition via mobile app is a beneficial component of mHealth intervention. The moderator analysis demonstrated that week-in-study negatively moderates the efficacy of team competition, suggesting a waning causal effect of competition over time. Also, intra-institutional competition decreased the efficacy of competition, suggesting potential improvements in strategies of assigning opponent teams to boost the effect of competition interventions.

The finding on the beneficial effect of mobile-based gamified competition upon physical activity is consistent with previous studies that have shown that the mHealth intervention with gamified components

can increase physical activity[23–25]. However, the effect size from our study is smaller than previous studies[24,25]. Under the highly stressful and intensive working environment, the medical interns may be less responsive to the intervention, which may explain smaller effect size relative to other study populations. On the other hand, our result may be applied to other shift workers or under-stress populations who are similar to medical interns.

One possible explanation for the waning causal effect of competition was that interns might be motivated when study began and get tired later so that they were less responsive to the competition assignment. The phenomenon of waning treatment effect is common in the field of mHealth intervention[10,26,27] and further studies are needed to investigate how to extend the mHealth intervention effect. For example, a break could be given to the interns after an episode of competition assignment to decrease their fatigue, and then intervention can be reintroduced to them after some rest time to regain the benefits from intervention. Adding novel competition-related elements such as levels, scoreboard, and prizes could be another option.

One possible explanation for the negative impact of intra-institutional competition on the causal effect of competition on step count was that perhaps interns felt less competitive within their institution relative to extra-institutional members because they see their fellow institution members as colleagues. The negative moderation of intra-institutional competition suggested avoiding intra-institutional competition assignment in the future application to maximize the competition effect.

We also explored the causal effect of team competition on user's engagement and mood score. A significant and positive effect of competition on participation rate of sleep minutes was observed, indicating that the competition might have potential to increase user's engagement. In addition, the positive while non-significant causal effect of competition on mood score was observed.

Our study has multiple strengths. First, compared with standard single-time-point randomized controlled trial design, which can only inform moderation of causal effect by baseline variables (e.g., age, gender), micro-randomized trial enabled us to assess both causal effects of intervention components and time-

varying moderation of these effects. Second, a relatively large sample size (1,779 participants in 191 teams) and long study period (12 weeks) allowed us to detect the causal effect of intervention, as well as effect moderators of interest. Third, the unique study population, medical interns with inherent hierarchical structure (by institutions and specialties), allowed us to assess the moderation of social connection and cooperation on the causal effect of gamified competition. Fourth, the analytical approaches we used in the study, the weighted and centered least squares estimator and multiple imputation, allowed us to assess the causal effect moderation consistently and robustly without requiring strong assumptions.

However, there are several unanswered questions that should be addressed in future research. It remains unclear why competition did not affect participant's sleep duration in the same way as step count. One of our conjectures is that the highly demanding working schedule during medical internship makes interns have little control over their sleep schedule, leading to insensitivity to competition intervention. Also, the reason that intra-institutional competition leads to negative moderation of causal effect of competition needs to be addressed in the future study.

Our study does have several limitations. The first is the data missingness and imputation. More than 30% data were missing for daily step count and 50% for daily sleep duration on individual level. Multiple imputation was used to impute the missing entries under the assumption of missing at random, however, the imputed values of a participant borrowed information from participants of other teams due to limited information in each team, which may result in attenuated estimate when assessing the moderation of competing within the same institution on causal effect of competition since the difference among teams can become smaller after imputation. Second, heterogeneity between Apple Watch and Fitbit charge activity monitors was not accounted for during the analysis. Previous studies have shown that Fitbit Charge 2 and Apple Watch 2 had similar accuracy in terms of estimating step counts[28,29], however, due to longer battery life, the Fitbit Charge series is more likely to be worn continuously, thus more likely to yield a higher step count and longer sleep duration in real life settings, compared with Apple Watch. Third, the results of this study may not extrapolate to a more general population because medical interns are different from the

general population in terms of age, education level, and stress level. Individuals in the general population may be more responsive to the mHealth intervention than medical interns due to more flexible time. Therefore, to validate the generalizability of the results in and out of Intern Health cohort, these suggested interventions should be further refined and replicated in additional studies and cohorts. Fourth, note that instead of individual level analysis, cluster level analysis, where each team was treated as the unit of analysis, was used to avoid ignoring the inherent clustering (team) structure using the team-level summary measures. The summary measures were calculated by taking the average of individuals' measurements of the same team, which did not account for the heterogeneity among members within the team, resulting in reducing the power of the study. Further methodological research on statistical tools allowing analysis at the level of the individual while accounting for the clustering in the data in the field of MRT is needed.

In summary, through this smartphone-wearable-based prospective micro-randomized trial, we were able to identify the positive causal effect of competition on proximal step count; exploratory analysis also suggested competition may improve user's engagement with the study app. In addition, the causal effects of competition were negatively moderated by week-in-study and intra-institutional competition. The competition intervention had no significant causal effect on the sleep duration. These results suggest that gamified competition is worthy of inclusion in the mHealth intervention. Effect of gamified competition may be further boosted by introducing occasional breaks to mitigate waning effects over time and optimizing opponent assignment.

METHODS

Study design and participants

We conducted a three-month MRT to investigate the causal effects of team competition upon proximal weekly average daily step counts, minutes of sleep, participation rate and mood score via the Intern+ mobile app as part of the Intern Health Study, a prospective cohort study assessing stress and depression during the first year of residency training in the USA[30]. Training physicians, who began their internship in July 2020,

were invited via email to participate in the study within one to three months prior to the start of internship. Ownership of an iPhone supporting iOS 10.0 or later or an Android device supporting version 6.0 or later was required. Upon enrollment, participants were provided with a Fitbit Charge 3 to collect sleep and activity data if they did not already own a compatible Fitbit or Apple Watch. All participants provided informed consent electronically and were compensated $80 to $130. The University of Michigan institutional review board approved the study. And the trial is registered with ClinicalTrials.gov, NCT05106439, November 3, 2021.

To protect participant anonymity, we required a minimum of five participating interns per team to be eligible for the competition arm of the study. Programs with at least five interns were grouped into program-based teams (e.g., "Michigan Psychiatry"). Interns within the same residency institution in programs that did not meet this criterion were grouped into institution-based teams (e.g., "Michigan Programs"), also with a minimum of five participants per team. All the remaining enrolled study subjects were considered ineligible for the competition arm.

All the eligible subjects were onboarded before their internships started on July 1, 2020. Baseline surveys that assessed interns' stress; also, baseline step counts, and sleep minutes were recorded via the Fitbits/Apple Watch. The competition assignment started on the first Monday following the start of internship (July 6, 2020) and ended on Sep 27, 2020 (Sunday of the 12th week), which lasted nearly three months. Each competition episode was one week, starting on Monday 00:00 and ended on Sunday 23:59.

Randomisation and masking

Each week, we repeatedly randomized interns by three factors: competition status (in competition or not), opponent team, and competition type (on step count or sleep minutes). Such a factorial design enables inference of causal effects of one or multiple factors based on the same study data. In particular, first, each team was randomized with equal probabilities to the competition or non-competition arm every Monday - this is the main randomization of the study. Second, every week, teams in competition were randomly

assigned an opponent team: 1) total randomization, where the opponent team was assigned regardless of institution and specialty (e.g., Michigan Pediatrics vs Yale Emergency Medicine); 2) intra-institutional randomization, where two competing teams were from the same institution (e.g., NYU Internal Medicine vs NYU Surgery); 3) intra-specialty randomization, where two opponent teams were from the same specialty (e.g., Northwestern Psychiatry vs OSU Psychiatry). All the three rules for opponent assignment had equal probability (1/3) to be selected for each week. If there were an odd number of teams inside the randomization pool, then the team left over would be put back into the non-competition arm. Third, for each pair of opponent teams, there was a 50/50 chance of competing on average daily step counts or average daily sleep minutes. Figure 2 details the randomization scheme. Due to the nature of the intervention, participants could not be masked from the competition assignment. Although investigators were not masked to intervention allocation, all data collected from participants was through the app or wearable device.

Procedures

After completing consent and downloading the study app, the wearable devices started to record daily step count and time spent asleep. Participants were prompted to report their daily mood (a score of 1 corresponded to the lowest and a score of 10 corresponded to the highest mood) every day at a user-specified time between 5 PM and 10 PM (default was 8PM) in the study app. In addition to collecting data, the study app aggregated and displayed visual summaries of participant's historical data, including daily step count, sleep minutes and mood score, through a dashboard which participants could access at any time via the app (see Supplementary Figure 1). Separate from the competition component of the app, each user also had a 50/50 chance each day to receive a push notification at 3pm which contained a message summarizing their personalized data feedback, a relevant fact or tip for improving mental health and well-being, or a general supportive statement. Furthermore, baseline and quarterly follow-up surveys were administered through the app.

The competition intervention was conducted for 12 weeks (Monday July 6, 2020, to Sunday September 27, 2020). Each team was randomly assigned to competition (intervention) and non-competition arm during

each competition episode. Four competition-related smartphone push notifications were sent to participants in each competition week:1) an alert of competition type (steps or sleep) and opponent team (Sunday 9 pm prior to the competition week), 2) two competition score updates (Wednesday 9 pm and Saturday 11 am during the competition week), and 3) the final competition results (Monday 12 pm following the competition week). Examples of messages are included in Supplementary Table 1. Participants could view their current competition scoreboard and competition history at any time via the Intern+ app. Supplementary Figure 1 shows three representative screenshots of the app interface involving competition.

Outcomes

The primary outcomes of the study were proximal weekly average daily step count and sleep minutes, by taking average values of team members within a competition episode, which were measured by wearable devices (Fitbit or Apple Watch). The exploratory outcome was proximal weekly average daily participation rates of step count and sleep minutes, which was defined as the proportion of days that the participants in the team provided daily step/sleep minutes within a competition week. The daily step count or sleep duration would be missing if the user was not wearing wearable devices during the daytime or nighttime. Demographic information and psychology-related scores were collected from baseline surveys.

Missing data

Missing data occurred throughout the trial for various reasons: forgetting to wear a fitness tracker, only wearing fitness tracker during the day, technical glitches, and so on (see missingness information in Supplementary Figure 2). Therefore, we used multiple imputation, a robust method for dealing with missing data, to impute the daily step count and minutes of sleep. For each day, the daily step count, sleep minutes and mood score were imputed with predictor variables including step count, hours of sleep and mood score from the previous three days, weekly average step count, sleep minutes and mood score from previous week and individual's baseline characteristics including gender, depressive symptoms score (PHQ-9), neuroticism, early family environment. To accommodate the heterogeneity between different institutions

and specialties, individual's institution and specialty were added to the predictor variable list of imputation. R version 4.0.2 and *mice* function from R library *mice* were used to do the multiple imputation and predictive mean matching was selected as the imputation method. Results were pooled using 20 imputed dataset following Rubin's rules.

Statistical Analysis

Note that the competition assignment was randomized on the team level, therefore all the competition-related analyses in this paper were performed on the team level using summary measurements from each team, that is, the team was treated as the unit of analysis instead of individual. Weekly team-based summary measurements were calculated by taking the average of individuals' measurements within each team.

The primary aim of this study assesses whether there was a main causal effect of being in the competition arm on the team's average proximal weekly average daily step count and sleep minutes, compared to not being in the competition arm. The primary analysis was done by fitting linear regression models using generalized estimating equation with independent working correlation matrix (R version 4.0.2; *geeglm* function from R library *geepack*) for average daily step count and sleep minutes separately, with competition assignment, number of weeks in study and control variables. Daily step count and sleep minutes were treated as continuous variables. The competition assignment variable was binary, with a value of 1 for being in a competition week and 0 for a non-competition week. The week-in-study was a continuous variable, with 0 for the first week of the study and 11 for the last week. Percentage of female, team average pre-intern measures, including daily step count, sleep minutes, psychology-related scores (e.g., PHQ-9 score), as well as team average previous week's outcomes (i.e., previous week's step count will be included if the outcome is current week's step count), were included as control variables to increase the statistical power. The procedure implements a weighted and centered least square estimator (WCLS, details included in Supplementary Notes), proposed by Boruvka et al[19].

The secondary aims included moderation analyses to assess potential time-varying effect moderators, aimed at informing the design of real-time personalized and optimized delivery of mHealth intervention. Time-varying effect moderators are moderators that can change the treatment effect and are time-varying because the values of moderators can vary across time (e.g. week-in-study and opponent team)[18]. Two potential time-varying moderators of causal effect of competition were examined. The first moderation analysis is motivated by the hypothesis that the longer a participant was in the study, the more they may be accustomed to the competition intervention or become overburdened, leading them to become less responsive. Interaction terms between number of additional weeks in study and intervention variable were included in the model to evaluate the effect moderation. The second moderation analysis is motivated by the hypothesis that participants in the same institution or specialty tended to have stronger social connection, which may result in fiercer competition to boost the competition effect. Therefore, moderator analysis for whether intern was competing within the same institution or specialty was done by including two additional interaction terms between the intervention indicator and the intra-institution and intra-specialty indicators, respectively.

Exploratory analyses included assessing the causal effects of team competition upon user's engagement and mood score averaged by team respectively. Linear probability model was used to assess the main causal effect of competition on proximal weekly average daily participation rates of step count and sleep minutes, motivated by the hypothesis that competition intervention can improve user's engagement to the study app (see details in Supplementary Notes). The casual effect of team competition on team-averaged self-reported mood score was evaluated similarly as our analysis on step count (sleep minutes) in the main and secondary aims and detailed in Supplementary Notes.

All the analyses above were based on the weekly aggregated data because every week was treated as a complete episode of competition or not. For the moderation analyses, the results of these effects were reported from the models with linear moderators. The significance of regression coefficients for all analyses were tested through two-sided Wald test.

Sensitivity analysis

To investigate the robustness of the results, we performed three types of sensitivity analysis. First, we compared the results from complete-case analysis and multiple imputation to evaluate the sensitivity of missing mechanisms. Second, we used a linear model for the moderation in the main text, that is, the model for the treatment effect was specified as a linear function of the moderator. To assess the sensitivity of the linearity assumption, we explore potential non-linearity in the interaction term between the causal effect of competition and additional weeks in the study by replacing the linear function with a non-linear function *f*. Here we fit *f* using penalized basis spline by *gam* function from *mgcv* R package and natural cubic spline from *ns* function in R. The penalized basis spline models were fit using restricted maximum likelihood method (REML) and thin plate regression spline as smoothing basis. Third, to assess the sensitivity of different missingness patterns, the main results, which were from multiple imputation analysis, were compared with complete-case analysis after introducing a certain missingness pattern. Here we examined two different missingness patterns: dropout and weekly missingness. For dropout complete-case analysis, we removed imputed data from interns who dropped out from the study early. For example, if a user does not have any data points after Sep 1, 2020, all the imputed data points for this user after Sep 1, 2020, were removed. For weekly missingness complete-case analysis, we removed weeks with a large percentage of missing data in the outcome of interest. For example, we eliminated all weeks where more than 5 data points were missing before performing complete-case analysis. The results of sensitivity analyses were detailed in Supplementary Notes.

DATA AVAILABILITY

Deidentified data supporting the results and figures in this manuscript are available from the corresponding author upon reasonable request and completion of a data agreement with the Intern Health Study team.

ACKNOWLEDGEMENTS

This study was supported by grants from the National Institute of Mental Health (R01 MH101459) to S.S. and Z.W., and an investigator grant from Precision Health Initiative at University of Michigan, Ann Arbor to Z.W. and S.S.. We thank the interns and residency programs who took part in this study.

COMPETING INTERESTS

T.N. is co-founder of a consulting company. He was also a former member of the IHS research team. His work on this paper was unpaid and not affiliated with his consulting company. All other authors have no competing interests to disclose.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions to the study conception and design. S.S. made substantial contributions to the acquisition of data. J.W., Y.F., Z.W. conducted statistical analysis. J.W., Y.F., E.F., M.W., M.B., A.T., W.D., T.N., S.S., Z.W. made substantial contributions to the interpretation of data. J.W., Y.F., E.F., S.S., Z.W. drafted the first version of the manuscript. All authors contributed to critical revisions and approved the final version of the manuscript.

CODE AVAILIBILITY

Code for data preprocessing and statistical analysis will be made available online at https://github.com/jtwang95/IHS_competition.

REFERENCES

1. Chin, S. -H., Kahathuduwa, C. N. & Binks, M. Physical activity and obesity: what we know and what we need to know*. *Obesity Reviews* **17**, 1226–1244 (2016).
2. Dinas, P. C., Koutedakis, Y. & Flouris, A. D. Effects of exercise and physical activity on depression. *Irish Journal of Medical Science* **180**, 319–325 (2011).
3. Miller, T. D., Balady, G. J. & Fletcher, G. F. Exercise and its role in the prevention and rehabilitation of cardiovascular disease. *Annals of Behavioral Medicine* **19**, 220–229 (1997).
4. Piercy, K. L. *et al.* The Physical Activity Guidelines for Americans. *JAMA* **320**, 2020 (2018).
5. Centers for Disease Control and Prevention (CDC). Effect of short sleep duration on daily activities--United States, 2005-2008. *MMWR. Morbidity and mortality weekly report* **60**, 239–42 (2011).
6. Hirshkowitz, M. *et al.* National Sleep Foundation's updated sleep duration recommendations: final report. *Sleep Health* **1**, 233–243 (2015).
7. McCallum, C., Rooksby, J. & Gray, C. M. Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review. *JMIR mHealth and uHealth* **6**, e58 (2018).
8. Nahum-Shani, I. *et al.* Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine* **52**, 446–462 (2018).
9. NeCamp, T. *et al.* Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: Micro-randomized trial. *Journal of Medical Internet Research* **22**, (2020).
10. Shcherbina, A. *et al.* The effect of digital physical activity interventions on daily step count: a randomised controlled crossover substudy of the MyHeart Counts Cardiovascular Health Study. *The Lancet Digital Health* **1**, e344–e352 (2019).

11. Qian, T. *et al.* The Micro-Randomized Trial for Developing Digital Interventions: Data Analysis Methods. (2020).
12. Edney, S. M. *et al.* A Social Networking and Gamified App to Increase Physical Activity: Cluster RCT. *American Journal of Preventive Medicine* **58**, e51–e62 (2020).
13. Sardi, L., Idri, A. & Fernández-Alemán, J. L. A systematic review of gamification in e-Health. *Journal of Biomedical Informatics* **71**, 31–48 (2017).
14. Patel, M. S. *et al.* Effect of Goal-Setting Approaches Within a Gamification Intervention to Increase Physical Activity Among Economically Disadvantaged Adults at Elevated Risk for Major Adverse Cardiovascular Events. *JAMA Cardiology* (2021) doi:10.1001/jamacardio.2021.3176.
15. Xu, L. *et al.* The Effects of mHealth-Based Gamification Interventions on Participation in Physical Activity: Systematic Review. *JMIR Mhealth Uhealth* **10**, e27794 (2022).
16. Edwards, E. A. *et al.* Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. *BMJ Open* **6**, e012447 (2016).
17. Michie, S. *et al.* The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine* **46**, 81–95 (2013).
18. Klasnja, P. *et al.* Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* **34**, 1220–1228 (2015).
19. Boruvka, A., Almirall, D., Witkiewitz, K. & Murphy, S. A. Assessing Time-Varying Causal Effect Moderation in Mobile Health. *Journal of the American Statistical Association* **113**, 1112–1121 (2018).
20. Qian, T. *et al.* The Micro-Randomized Trial for Developing Digital Interventions: Experimental Design and Data Analysis Considerations. (2021).

21. Kalmbach, D. A. *et al.* Effects of Sleep, Physical Activity, and Shift Work on Daily Mood: a Prospective Mobile Monitoring Study of Medical Interns. *Journal of General Internal Medicine* **33**, 914–920 (2018).
22. Khan, W. A. A., Jackson, M. L., Kennedy, G. A. & Conduit, R. A field investigation of the relationship between rotating shifts, sleep, mental health and physical activity of Australian paramedics. *Scientific Reports* **11**, 866 (2021).
23. Shameli, A., Althoff, T., Saberi, A. & Leskovec, J. How Gamification Affects Physical Activity. in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion* 455–463 (ACM Press, 2017). doi:10.1145/3041021.3054172.
24. Garde, A. *et al.* Assessment of a Mobile Game ("MobileKids Monster Manor") to Promote Physical Activity Among Children. *Games for Health Journal* **4**, 149–158 (2015).
25. Patel, M. S. *et al.* Effectiveness of Behaviorally Designed Gamification Interventions With Social Incentives for Increasing Physical Activity Among Overweight and Obese Adults Across the United States. *JAMA Internal Medicine* **179**, 1624 (2019).
26. Klasnja, P. *et al.* Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine* **53**, 573–582 (2019).
27. Hamari, J., Koivisto, J. & Sarsa, H. Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. in *2014 47th Hawaii International Conference on System Sciences* 3025–3034 (IEEE, 2014). doi:10.1109/HICSS.2014.377.
28. Bai, Y. *et al.* Comprehensive comparison of Apple Watch and Fitbit monitors in a free-living setting. *PLOS ONE* **16**, e0251975 (2021).
29. Fuller, D. *et al.* Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. *JMIR mHealth and uHealth* **8**, e18694 (2020).

30. Sen, S. *et al.* A Prospective Cohort Study Investigating Factors Associated With Depression During Medical Internship. *Archives of General Psychiatry* **67**, 557 (2010).

31. Thomas, N., Harel, O. & Little, R. J. A. Analyzing clinical trial outcomes based on incomplete daily diary reports. *Statistics in Medicine* **35**, 2894–2906 (2016).

FIGURES

Figure 1: Study flow diagram detailing subject inclusion from enrollment through analysis

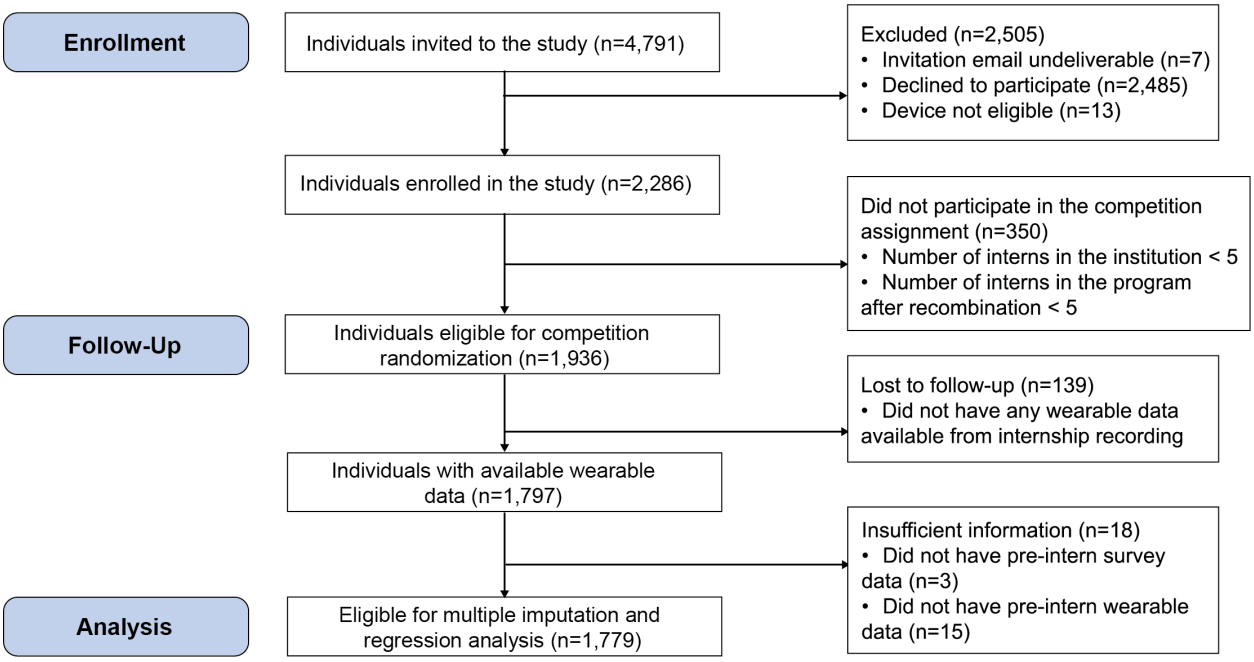
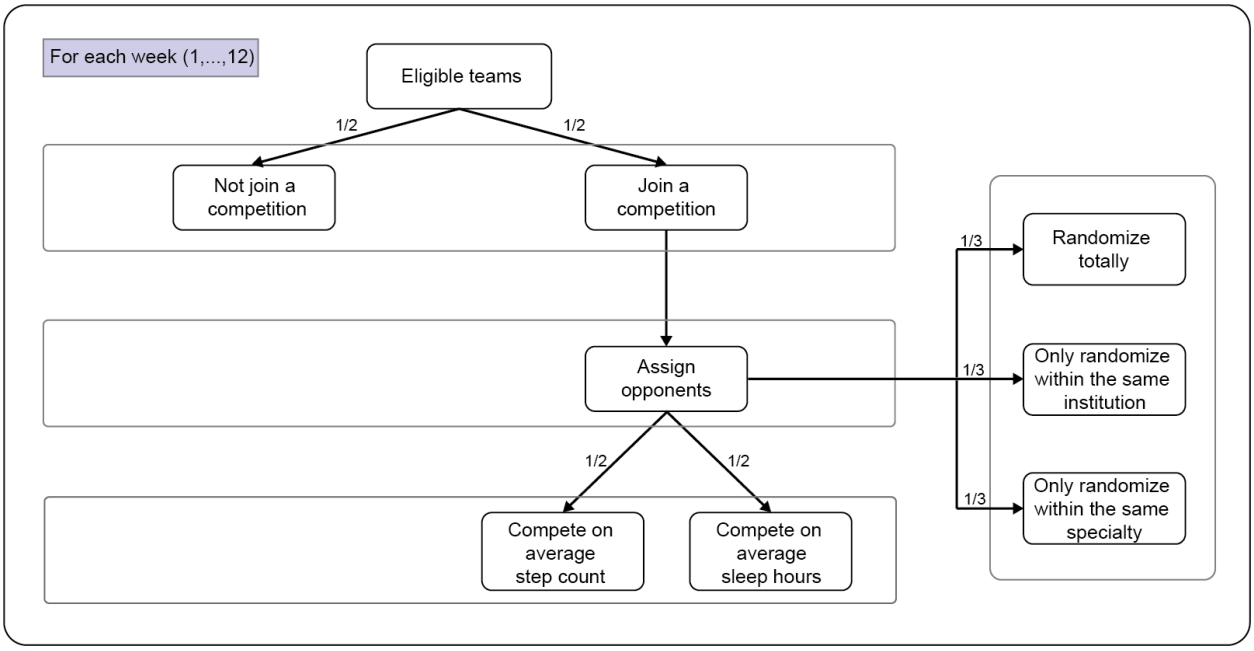


Figure 2: Study randomization scheme of Intern Health micro-randomized trial



TABLES

Table 1: Demographics characteristics and specialty for study participants (N=1779).

Demographic characteristics		Specialty	N (%)
Age (years), mean (SD)	27.6 (2.6)	Internal medicine	492 (27.7)
Sex (Female), N (%)	969 (54.5)	Surgery	240 (13.5)
Pre-internship baseline steps/day, mean (SD)	8121.0 (3228.9)	Pediatrics	215 (12.1)
Pre-internship baseline sleep minutes/day, mean (SD)	420.6 (107.5)	Emergency medicine	145 (8.2)
Internship average days of competition, mean (SD)	40.7 (13.5)	Psychiatry	126 (7.1)
Race, N (%)		Ob/Gyn	104 (5.8)
White	957 (53.8)	Anesthesiology	94 (5.3)
Black/African American	106 (6.0)	Family medicine	78 (4.4)
Hispanic/Latino	96 (5.4)	Neurology	49 (2.8)
Asian	420 (23.6)	Med/Peds	38 (2.1)
Arab/Middle Eastern	31 (1.7)	Transitional	21(1.2)
Other/Multiracial/Not reported	169 (9.4)	Other	177 (9.9)

N number of subjects, *SD* standard deviation.

Table 2: Parameter estimates for linear models assessing marginal and time-varying causal effect of team competition on daily step count and sleep duration.

			Main Effect Analysis		Time-varying Effect Analysis	
		Parameter	Estimate	95% CI	Estimate	95% CI
Outcome & Competition type	Step	Intercept	7679.3	7552.0, 7806.6	7664.5	7527.6, 7781.4
		Week	-18.0	-28.6, -7.4	-15.1	-28.8, -1.4
		Competition	111.5	32.2, 190.8	161.5	17.2, 305.8
		Week: Competition	-	-	-9.1	-32.0, 13.8
	Sleep	Intercept	416.6	411.8, 421.5	413.5	408.3, 418.8
		Week	0.0	-0.3, 0.3	0.5	0.1, 1.0
		Competition	-0.7	-4.3, 29	9.9	2.7, 17.1
		Week: Competition	-	-	-1.9	-3.1, -0.7

CI confidence interval

Supplementary Materials

Supplement to: Effectiveness of gamified team competition as mHealth intervention for medical interns: a cluster micro-randomized trial

List of Supplementary Materials

Supplementary Notes

A weighted and centered least square estimator (WCLS)

To estimate the coefficients of interest with the existence of time-varying moderators, we used a weighted and centered least squares estimator proposed by Boruvka et al[1]. The method uses both the estimating equation method and robust "sandwich" estimate to provide consistent estimates and robust inference. The main advantage of the proposed method is that it does not require correct specifications to the terms that do not interact with treatment, to provide consistent estimates for the parameters of interest.

WCLS can provide valid inference for treatment-related variables when the treatment assignment probabilities are time-varying. Since the treatment assignment probabilities were constantly 0.25 across weeks in the IHS 2020 (each team had 50/50 chance to be in a competition week and for teams in competition week, they had 50/50 chance to compete on step count or sleep minutes), the weights in the estimating equation are all 1. In addition, the centering term for treatment variable is always a constant 0.25.

The estimating equation method with robust error estimation has two main advantages: 1. It does not require distributional assumptions on continuous outcomes. 2. It allows dependence between observations in the data, where the observations for the same team in our dataset are correlated due to repeated measurements.

Models for assessing causal effect of competition on participation rate

To investigate the causal effect of competition on intern's participation rate on daily step count and sleep duration, linear probability model was used to model the mean structure. More specifically, we modeled the probability that the daily step count or sleep duration was measured in week t as $E(Y_t|X_t, W_t, Z_t) = \beta_0 + \beta_1 Z_t + \beta_2 W_t + \beta_3 X_t$ for marginal-effect model, and $E(Y_t|X_t, W_t, Z_t) = \beta_0 + \beta_1 Z_t + \beta_2 W_t + \beta_3 Z_t W_t + \beta_4 X_t$ for time-varying-effect model, where Y_t is a continuous variable between 0 and 1 and the value of Y_t represents the proportion of available data points during week t for each team. $Z_t$is a binary treatment variable, where $Z_t = 1$implies a competition week and $Z_t = 0$ implies a non-competition week. X_t is the set of control variables, including team-average baseline data measured before the weekly randomization, for the purpose of reducing variation in the outcome of interest Y_t. The set of variables X_t consists of percentage of female, team-average pre-intern daily step count, sleep minutes and psychology-related scores.

In the Intern Health Study data, whether the intern had a daily step count, sleep duration or self-reported mood survey is considered as a binary outcome. A conventional approach to model the binary outcome is to fit a logistic model. However, in our analysis, we did not adapt conventional logistic regression approach on daily data. Instead, we perform a weekly analysis using linear probability model and use a weighted and centered estimator proposed by Boruvka et al. to estimate the parameters of interest. The reason that we preferred a weekly analysis to a daily analysis is that the competition assignment was randomized on a weekly level and the daily level analysis violates the positivity assumption of the estimating method proposed by Boruvka et al[1]. Besides, the reason that we used linear probability model rather than beta regression model, which is more commonly used than linear probability model for proportional outcomes due to unit interval boundary, was that the theoretical guarantee of using beta regression to assess the time-varying causal effect moderation is not well-established. We calculated the predicted probability for each week in our dataset and all the estimated values fell between unit interval, suggesting that the linear probability model is appropriate in our case.

Mood-score related analysis

Mental health disorders such as anxiety and depression are considered to be closely related to insufficient physical activity and sleep duration[2,3]. Therefore, we performed analyses on assessing the marginal and time-varying causal effect of competition on intern's mental health outcome: self-report mood score. We also did similar analyses on the participation rate of daily mood surveys. During mood-related analyses, the treatment variable competition was defined as either an intern was in competition step or sleep group. Here, we considered the competition indirectly affected the intern's mood score since interns were competing on daily step count and sleep minutes, rather than directly on mood score.

The parameter estimates for linear models assessing the marginal and time-varying causal effect of competition on weekly average daily mood score were shown in Supplementary Table 6. From the marginal-effect model, we concluded that on average competition tended to improve the daily mood score with an estimated effect of 0.02 (SE 0.02, p=0.35). We also concluded that the additional weeks in the study was a significantly negative moderator of the causal effect of competition on daily mood score with an estimated moderation of -0.01 (SE 0.01, p=0.05) from the time-varying-effect model. The time-varying-effect plot showed that being in the competition arm had a significant positive effect on daily mood score at the early stage of the study and the effect waned over time.

The parameter estimates for linear models assessing the marginal and time-varying causal effect of competition on participation rate of daily mood survey were shown in Supplementary Table 5. Also, the estimated causal effect of competition on participation rate of daily mood survey at different weeks was shown in Supplementary Figure 4. We concluded that the competition did not affect the participation rate of daily mood surveys marginally. From Supplementary Figure 4, we can observe an interesting fact that the competition decreased the intern's participation rate of self-report mood surveys early in the study. Other than push notifications including life insight and tips received by all interns, the ones assigned to the competition arm received additional competition-related messages (see Supplementary Table 1) four times per week, which might make those less responsive to the push notifications and more possible to ignore the mood survey completion reminder 8:00 pm every night. Daily step and survey data were collected objectively through the fitness tracker, which was less sensitive to push notification fatigue. Intensive mHealth push notifications (overtreatment) may lead to inferior treatment effect; therefore, this gives rise to the need for just-in-time adaptive intervention (JITAI), which can deliver mHealth intervention optimally.

Results of sensitivity analysis

Sensitivity of complete-case analysis

For primary aim, the estimate of the marginal causal effect of competition on step count from multiple imputation analysis was 111.5 (SE 40.4, p=0.01) steps, compared to 102.6 (SE 46.9, p=0.03) steps from complete-case analysis. The estimate of the causal effect of competition on sleep duration from multiple imputation analysis was -0.7 (SE 1.8, p=0.69) minutes, compared to -0.2 (SE 2.0, p=0.93) minutes from complete-case analysis. We can conclude that the conclusions for primary aim were mildly sensitive to missingness mechanisms.

For secondary aims, the estimate of the moderation of additional weeks in the study on the causal effect of competition on step count was -9.1 (SE 11.6, p=0.43) steps/week from multiple imputation analysis, compared to -6.9 (SE 11.7, p=0.55) steps/week from complete-case analysis. The estimate of the moderation of additional weeks in the study on the causal effect of competition on sleep duration was -1.9 (SE 0.6, p=0.003) minutes/week from multiple imputation analysis, compared to -1.1 (SE 0.6, p=0.06) minutes/week from complete-case analysis. The estimate of the moderation of competing within the same institution or specialty on the causal effect of competition on step count was -114.9 (SE 93.7, p=0.22) steps and 26.1(SE 74.7, p=0.73) steps from multiple imputation analysis, compared to -172.7 (SE 104.1, p=0.10) steps and 71.2 (SE 77.7, p=0.36) steps from complete-case analysis. The estimate of the moderation of competing within the same institution or specialty on the causal effect of competition on sleep duration

was 0.4 (SE 3.1, p=0.90) minutes and -1.9 (SE 3.2, p=0.57) minutes from multiple imputation analysis, compared to -1.7 (SE 4.1, p=0.68) minutes and 2.1 (SE 4.4, p=0.63) minutes from complete-case analysis. We can conclude that the conclusions for moderation of additional weeks in the study on causal effect of competition were insensitive to missingness mechanisms, while the conclusions for moderation of competing within the same institution or specialty on causal effect of competition were sensitive to missingness mechanisms. The size of the estimated moderation of competing within the same institution was enlarged when performing complete-case analysis and the sign of the moderation remained negative, matching the conclusions made in the main text.

For mood-related analysis, the estimate of the marginal causal effect of competition on mood score from multiple imputation analysis was 0.02 (SE 0.02, p=0.35), compared to 0.01 (SE 0.02, p=0.32) from complete-case analysis. The estimate of the moderation of additional weeks in the study on the causal effect of competition on mood score was -0.01 (SE 0.01, p=0.05) from multiple imputation analysis, compared to -0.01 (SE 0.00, p=0.05) from complete-case analysis. We can conclude that the conclusions for mood-related analysis were insensitive to missingness mechanisms.

The estimates of all the models mentioned above can be obtained through Supplementary Figure 2-4,6.

Sensitivity of non-linear moderation of treatment effect

The estimated causal effect of competition on step count or sleep duration at different weeks from nonlinear regression was plotted in Supplementary Figure 5. From the plots, we can see that linearity assumption is appropriate for our analysis.

The estimated causal effect of competition on participation rate of step count, sleep duration and mood survey at different weeks from nonlinear regression was plotted in Supplementary Figure 6. From the plots, we can notice some evidence of non-linearity, while linearity assumption is still appropriate for easy interpretation.

Sensitivity of missingness patterns

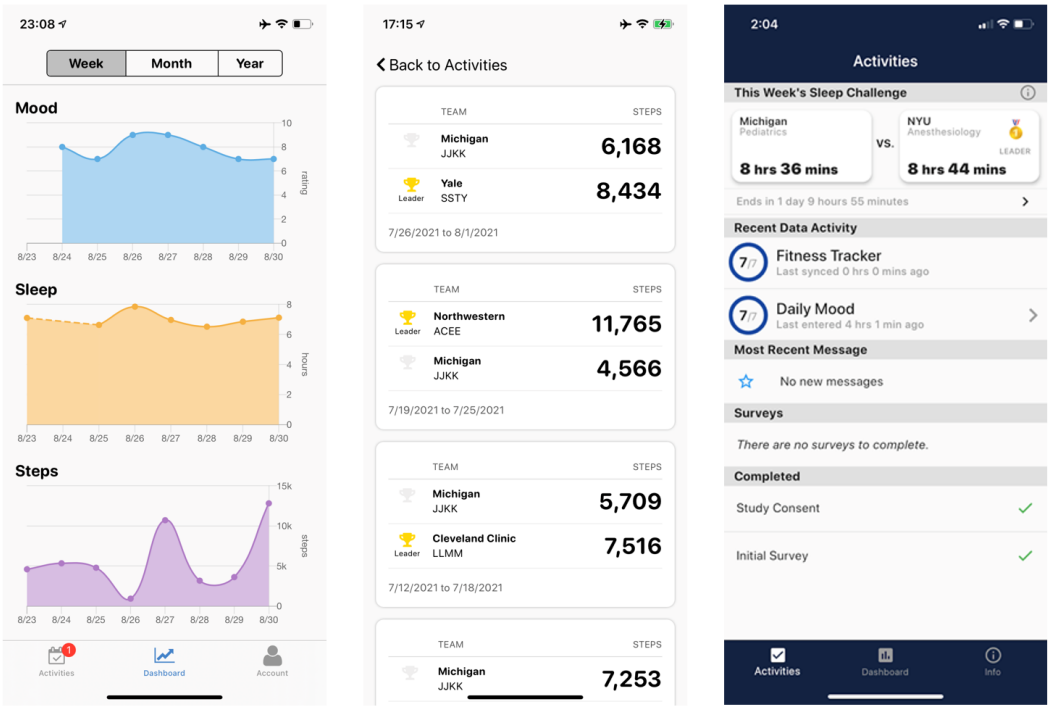
For primary aim, the estimate of marginal causal effect of competition on step count was 111.5 (SE 40.4) steps from multiple imputation analysis, compared to 93.1 (SE 45.7) steps from complete-case analysis with dropout and 91.1 (SE 45.7) steps from complete-case analysis with weekly missingness. The estimate of marginal causal effect of competition on sleep duration was -0.7(SE 1.8) minutes from multiple imputation analysis, compared to -0.5 (SE 1.7) steps from complete-case analysis with dropout and -0.5 (SE 1.8) minutes from complete-case analysis with weekly

missingness. We concluded that the conclusions of the primary aim were robust to both dropout and weekly data missingness.

For secondary aims, the estimate of the moderation of additional weeks in the study on the causal effect of competition on step count was -9.1 (SE 11.6) steps/week from multiple imputation analysis, compared to -7.7 (SE 11.7) steps/week from complete-case analysis with dropout and -2.6 (SE 11.6) steps/week from complete-case analysis with weekly missingness. The estimate of the moderation of additional weeks in the study on the causal effect of competition on sleep duration was -1.9 (SE 0.6) minutes/week from multiple imputation analysis, compared to -1.6 (SE 0.6) minutes/week from complete-case analysis with dropout and -1.3 (SE 0.6) minutes/week from complete-case analysis with weekly missingness. The estimate of the moderation of competing within the same institution or specialty on the causal effect of competition on step count was -114.9 (SE 93.7) steps and 26.1 (SE 74.7) steps from multiple imputation analysis, compared to -171.4 (SE 102.1) steps and 25.4 (SE 77.9) steps from complete-case analysis with dropout and -186.2 (SE 105.9) steps and 27.1 (SE 80.8) steps from complete-case analysis with weekly missingness. The estimate of the moderation of competing within the same institution or specialty on the causal effect of competition on sleep duration was 0.4 (SE 3.1) minutes and -1.9 (SE 0.6) minutes from multiple imputation analysis, compared to -2.0 (SE 3.9) minutes and -1.1 (SE 4.1) minutes from complete-case analysis with dropout and -1.2 (SE 4.2) minutes and 0.4 (SE 4.5) minutes from complete-case analysis with dropout. We can conclude that the conclusions for the effect of two moderators on the causal effect of competition were insensitive to dropout and weekly missingness, except that the moderation intra-institution competition is sensitive to the weekly missingness and dropout. The sign the moderation remained negative; however, the effect sizes were increased.

For mood-related analysis, the estimate of marginal causal effect of competition on mood score was 0.02 (SE 0.02) from multiple imputation analysis, compared to 0.03 (SE 0.02) from complete-case analysis with dropout and 0.02 (SE 0.02) from complete-case analysis with weekly missingness. The estimate of moderation of additional weeks in the study on causal effect of competition on mood score was -0.01 (SE 0.01) from multiple imputation analysis, compared to -0.02 (SE 0.01) steps from complete-case analysis with dropout and -0.02 (SE 0.01) from complete-case analysis with weekly missingness. We can conclude that the conclusions of mood-related analysis were insensitive to dropout and weekly missingness.

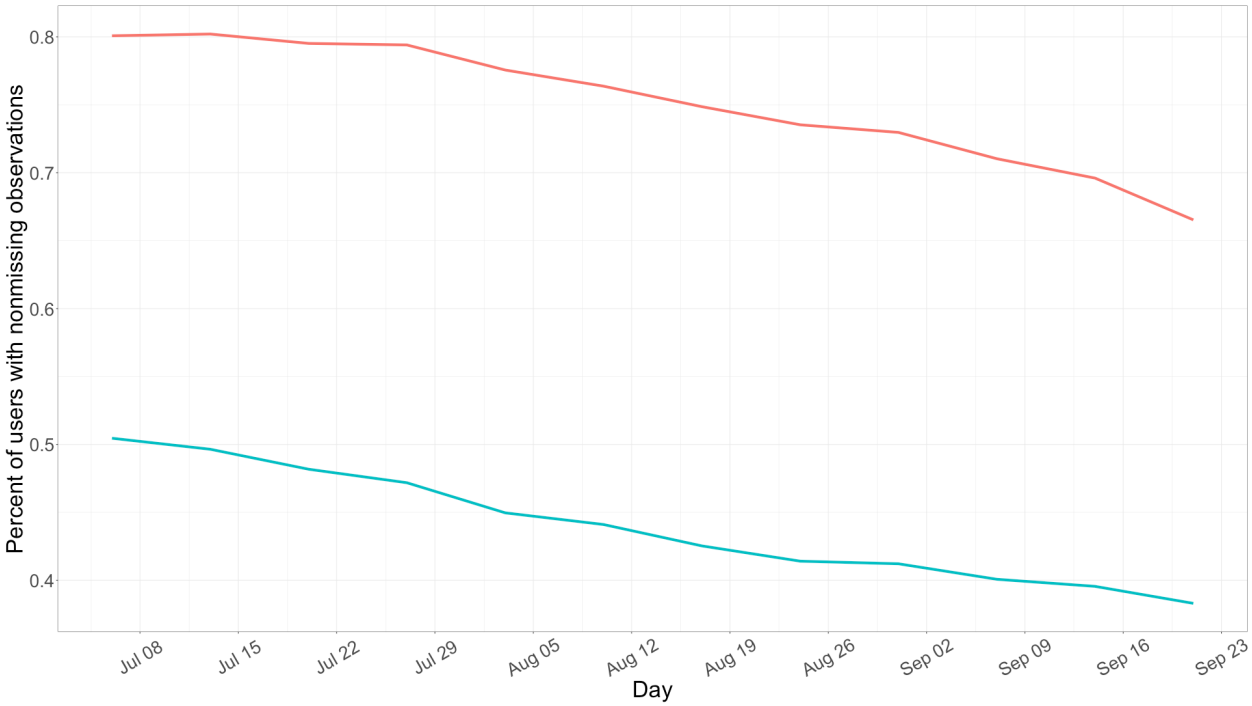
The estimates of all the models mentioned above can be obtained through Supplementary Table 7-10.



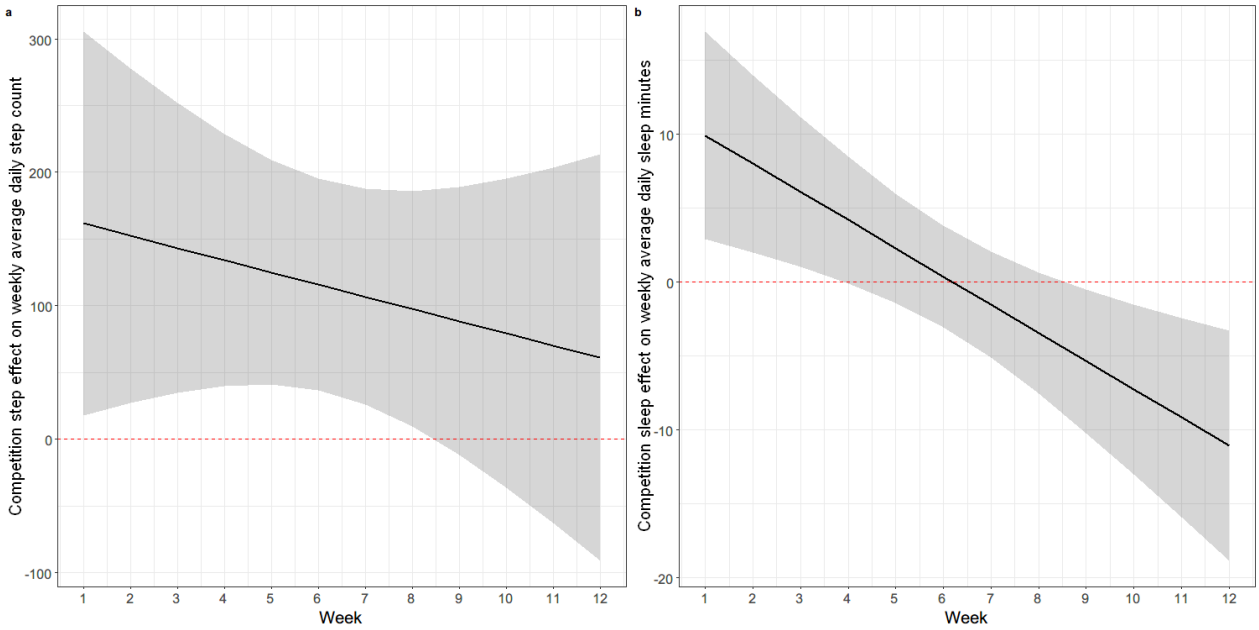
i) dashboard ii) competition history iii) competition assignment

Supplementary Figure 1: Screenshots of the app[1] dashboard, previous competition history and weekly competition assignment. The screenshot of previous competition history contains pseudo program names.

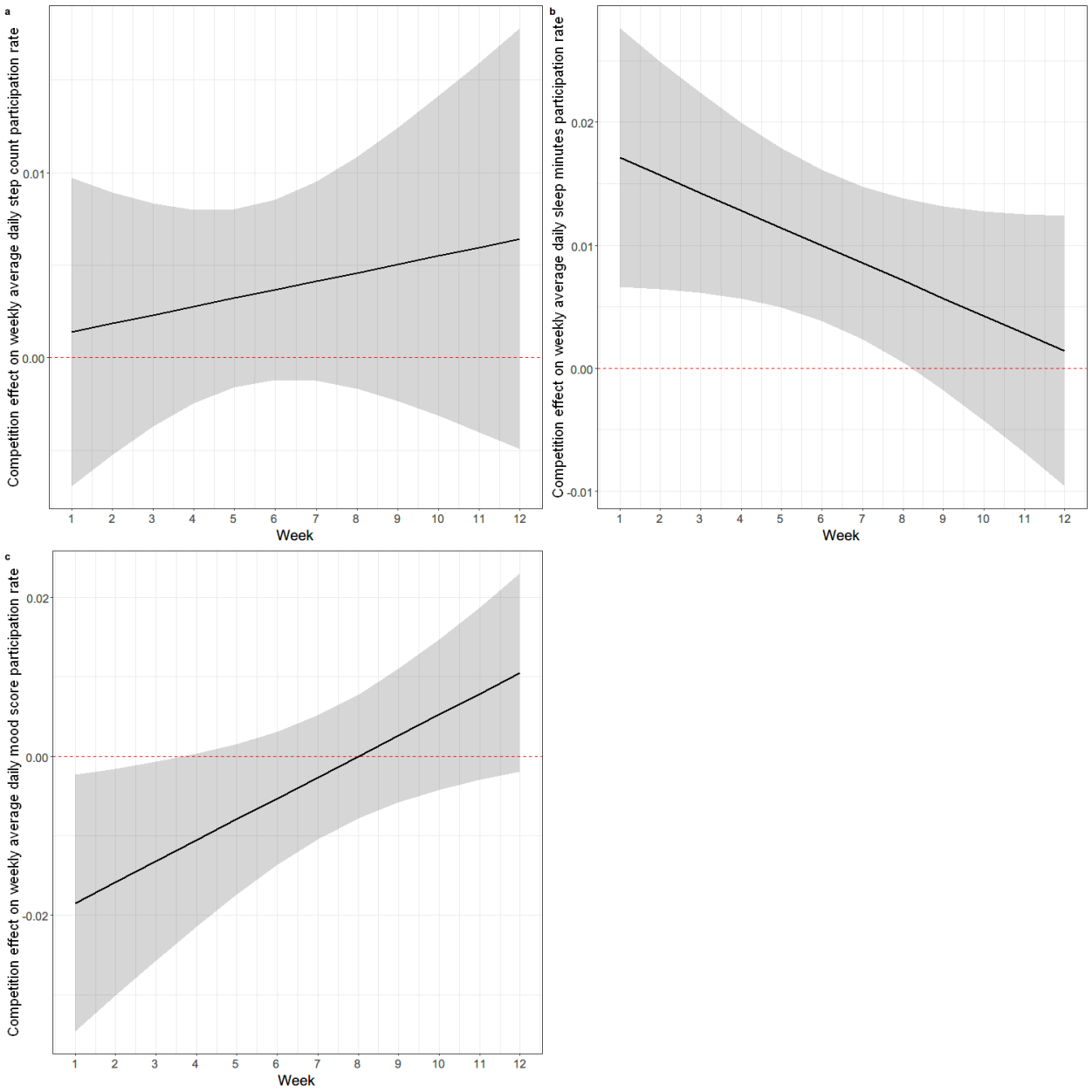
[1](c) 2016-20222 The Regents of the University of Michigan, Intern + Mobile Application; Permission to use by Office of Innovation Partnership at University of Michigan.



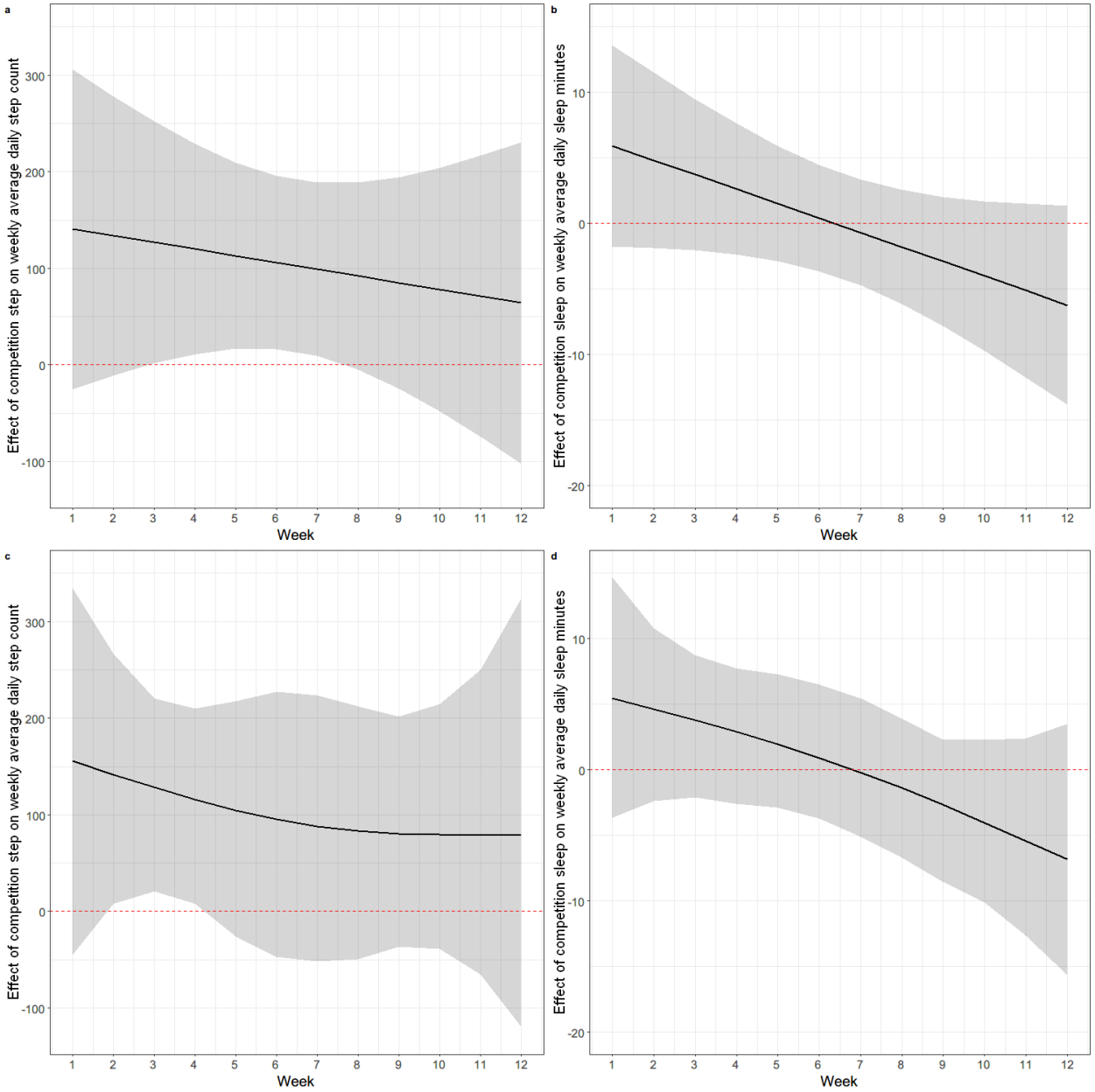
Supplementary Figure 2: Percentage of interns with nonmissing step and sleep observation for each day in the study. Red solid line indicates percentage of non-missing daily step count over time; Blue solid line indicates percentage of non-missing daily sleep record over time.



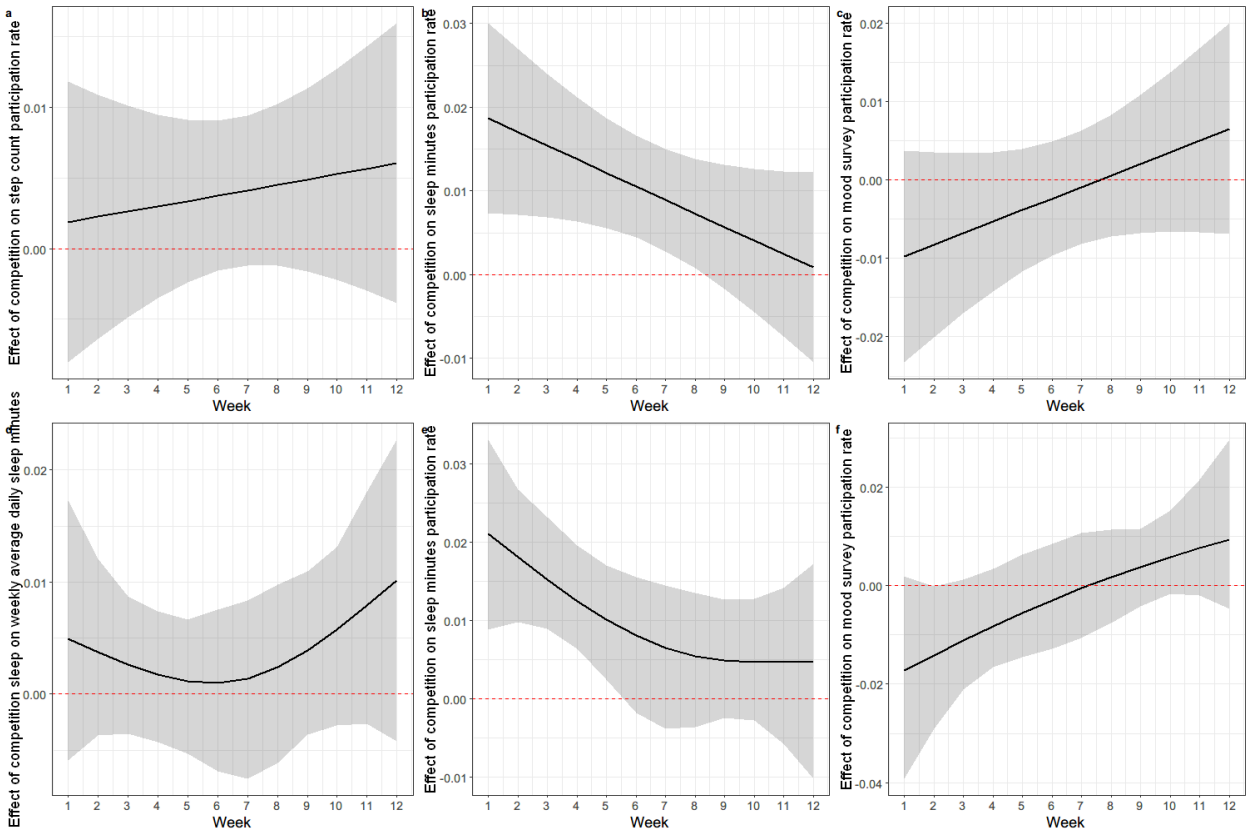
Supplementary Figure 3: a) Estimated causal effect of competition step on weekly average daily step count at different weeks. b) Estimated causal effect of competition sleep on weekly average daily sleep minutes at different weeks. Shaded area indicates 95% confidence interval. Red dotted line indicates no effect.



Supplementary Figure 4: Estimated causal effect of competition on the participation rate of a) daily step count, b) daily sleep minutes c) daily mood survey, at different weeks. Shaded area indicates 95% confidence interval. Red dotted line indicates the effect being 0.



Supplementary Figure 5: a,b)Estimated causal effect of competition on weekly average daily a) step count and b) sleep duration at different weeks fitted using penalized basis spline. c,d) Estimated causal effect of competition on weekly average daily c) step count and d) sleep duration at different weeks fitted using natural cubic spline. Shaded area indicates 95% confidence interval. Red dotted line indicates no effect.



Supplementary Figure 6: a,b,c) Estimated causal effect of competition on participation rate of a) step count, b) sleep minutes, c) mood score at different weeks using penalized basis spline. d,e,f) Estimated causal effect of competition on participation rate of d) step count, e) sleep minutes, f) mood score at different weeks using natural cubic spline. Shaded area indicates 95% confidence interval. Red dotted line indicates no effect.

Supplementary Table 1. Examples of different types of push notifications.

Message types	Time	Examples
Alert of competition types and opponent	Sunday 9:00 pm	MGH Surgery faces off against Northwestern Internal Medicine in this week's step competition!
Competition score status update	Wednesday 9:00 pm and Saturday 11:00 am	Yale Psychiatry is leading in this week's sleep challenge with an average of 8 hrs 41 min. Let's see who will win!
Competition final result	Monday 12:00 pm	Michigan Pediatrics comes out on top of this week's step challenge against NYU Anesthesiology. Great job to both teams!

Supplementary Table 2: Parameter estimates for linear model using complete-case and 20-time multiple imputation dataset, assessing marginal causal effect of competition on daily step count and sleep duration.

Outcome & Competition type		Complete case		Multiple Imputation	
	Parameter	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Step	Intercept	7779.9	7635.3, 7924.4	7679.3	7552.0, 7806.6
	Week	-12.6	-23.0, -2.2	-18.0	-28.6, -7.4
	Competition Step	102.6	10.8, 194.4	111.5	32.2, 190.8
Sleep	Intercept	415.2	409.2, 421.1	416.6	411.8, 421.5
	Week	0.3	-0.1, 0.8	0.0	-0.3, 0.3
	Competition Sleep	-0.2	-4.2, 3.8	-0.7	-4.3, 2.9

CI confidence interval.

Supplementary Table 3: Parameter estimates for linear model using complete-case and 20-time multiple imputation dataset, assessing time-varying causal effect of competition on daily step count and sleep duration.

Outcome & Competition type		Complete case		Multiple Imputation	
	Parameter	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Step	Intercept	7768.6	7617.4, 7919.8	7664.5	7529.2, 7799.8
	Week	-10.4	-24.4, 3.6	-15.1	-28.8, -1.4
	Competition Step	140.6	0.9, 280.4	161.5	17.2, 305.8
	Week: Competition Step	-6.9	-29.8, 15.9	-9.1	-32.0, 13.8
Sleep	Intercept	413.4	407.2, 419.6	413.5	408.3, 418.8
	Week	0.6	0.1, 1.2	0.5	0.1, 1.0
	Competition Sleep	5.9	-0.8, 12.6	9.9	2,7, 17.1
	Week: Competition Sleep	-1.1	-2.2, 0.0	-1.9	-3.1, -0.7

CI confidence interval.

Supplementary Table 4: Parameter estimates for linear models using complete-case and 20-time multiple imputation dataset, assessing moderation of competing within the same institution on causal effect of competition on step count and sleep duration.

Outcome & Competition type		Complete Case		Multiple Imputation	
	Parameter	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Step	Intercept	7766.8	7614.0, 7919.6	7664.5	7527.6, 7801.4
	Week	-10.4	-24.4, 3.6	-15.1	-28.8, -1.4
	Competition Step	167.2	21.6, 312.8	182.1	31.7, 332.5
	Week: Competition Step	-11.4	-34.9, 12.2	-11.3	-34.9, 12.2
	Competition Step : Same Institution Competition	-172.7	-376.7, 31.3	-114.9	-299.3, 69.6
	Competition Step : Same Specialty Competition	71.2	-81.2, 223.5	26.1	-120.9, 173.2
Sleep	Intercept	413.4	407.3, 419.6	413.5	408.3, 418.8
	Week	0.6	0.1, 1.2	0.5	0.1, 1.0
	Competition Sleep	5.9	-1.4, 13.3	10.1	2.8, 17.5
	Week: Competition Sleep	-1.1	-2.3, 0.0	-1.9	-3.1, -0.6
	Competition Sleep : Same Institution Competition	-1.7	-9.6, 6.3	0.4	-5.8, 6.6
	Competition Sleep : Same Specialty Competition	2.1	-6.5, 10.8	-1.9	-8.3, 4.6

CI confidence interval.

Supplementary Table 5: Parameter estimates for linear models, assessing marginal and time-varying causal effect of competition on participation rate of daily step count, sleep duration and mood score (*100).

Model		Step		Sleep		Mood	
	Parameter	**Estimate**	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Main-effect	Intercept	76.8	76.0, 77.5	43.5	42.8, 44.3	51.2	49.7, 52.6
	Week	-0.4	-0.4, -0.3	-0.2	-0.3, -0.1	-0.6	-0.7, -0.5
	Competition	0.4	-0.1, 0.9	0.9	0.3, 1.5	-0.4	-1.2, 0.4
Time-varying-effect	Intercept	76.9	76.1, 77.7	43.2	42.3, 44.1	51.8	50.3, 53.4
	Week	-0.4	-0.5, -0.3	-0.1	-0.2, 0.0	-0.8	-0.9, -0.6
	Competition	0.1	-0.7, 1.0	1.7	0.7, 2.8	-1.8	-3.5, -0.2
	Week: Competition	0.1	-0.1, 0.2	-0.1	-0.3, 0.0	0.3	0.0, 0.5

CI confidence interval.

Supplementary Table 6: Parameter estimates for linear model using complete-case and 20-time multiple imputation dataset, assessing marginal and time-varying causal effect of competition on mood score (*100).

Outcome & Competition type	Model		Complete case		Multiple Imputation	
		Parameter	Estimate	95% CI	Estimate	95% CI
Mood	Main-effect model	Intercept	720.1	714.8, 725.5	736.3	723.4, 749.2
		Week	-0.2	-0.6, 0.2	-1.7	-2.3, -1.0
		Competition	-1.1	-4.2, 2.0	1.9	-2.1, 6.0
	Time-varying-effect model	Intercept	718.0	712.3, 723.8	733.4	720.1, 746.7
		Week	0.2	-0.4, 0.8	-1.1	-1.9, -0.3
		Competition	3.6	-1.7, 8.9	8.3	1.3, 15.3
		Week: Competition	-0.9	-1.7, 0.0	-1.2	-2.3, 0.0

CI confidence interval.

Supplementary Table 7: Sensitivity analyses for assessing different missing patterns on marginal causal effect of competition on daily step count and sleep duration.

Missing pattern	Outcome & Competition type	Parameter	Estimate	95% CI
Dropout	**Step**	Intercept	7735.4	7604.7, 7866.1
		Week	-14.4	-25.0, -3.7
		Competition	93.1	3.5, 182.8
	Sleep	Intercept	416.1	411.4, 420.8
		Week	0.1	-0.4, 0.5
		Competition	-0.5	-3.9, 2.8
Weekly missingness	**Step**	Intercept	7796.7	7672.9, 7920.5
		Week	-15.2	-26.0, -4.3
		Competition	91.1	1.4, 180.7
	Sleep	Intercept	416.2	411.8, 420.5
		Week	0.4	-0.1, 0.8
		Competition	-0.6	-4.1, 3.0

CI confidence interval.

Supplementary Table 8: Sensitivity analyses for assessing different missing patterns on time-varying causal effect of competition on daily step count and sleep duration.

Missing pattern	Outcome & Competition type	Parameter	Estimate	95% CI
Dropout	**Step**	Intercept	7722.9	7589.9, 7856.0
		Week	-11.9	-25.8, 1.9
		Competition	135.3	-7.0, 277.5
		Week: Competition	-7.7	-30.6, 15.2
	Sleep	Intercept	413.7	408.3, 419.0
		Week	0.5	-0.0, 1.1
		Competition	8.1	1.1, 15.1
		Week: Competition	-1.6	-2.8, -0.4
Weekly missingness	**Step**	Intercept	7792.4	7664.9, 7919.8
		Week	-14.4	-28.6, -0.2
		Competition	105.5	-35.9, 246.9
		Week: Competition	-2.6	-25.5, 20.2
	Sleep	Intercept	414.2	409.4, 419.0
		Week	0.7	0.2, 1.3
		Competition	6.4	-0.2, 12.9
		Week: Competition	-1.3	-2.4, -0.1

CI confidence interval.

Supplementary Table 9: Sensitivity analyses for assessing different missing patterns on moderation of competing within the same institution or specialty on causal effect of competition on daily step count and sleep duration.

Missing pattern	Outcome & Competition type	Parameter	Estimate	95% CI
Dropout	Step	Intercept	7722.8	7588.2, 7857.4
		Week	-12.0	-25.8, 1.9
		Competition Step	167.7	20.8, 314.6
		Week: Competition Step	-10.5	-33.9, 12.9
		Competition Step : Same Institution Competition	-171.4	-371.9, 29.0
		Competition Step : Same Specialty Competition	25.4	-127.6, 178.3
	Sleep	Intercept	413.7	408.4, 419.0
		Week	0.5	-0.0, 1.1
		Competition Sleep	8.7	1.4, 16.1
		Week: Competition Sleep	-1.6	-2.8, -0.4
		Competition Sleep : Same Institution Competition	-2.0	-9.7, 5.7
		Competition Sleep : Same Specialty Competition	-1.1	-9.1, 6.9
Weekly missingness	Step	Intercept	7791.9	7663.1, 7920.7
		Week	-14.4	-28.6, -0.2
		Competition Step	140.8	-6.6, 288.3
		Week: Competition Step	-5.7	-29.1, 17.7
		Competition Step: Same Institution Competition	-186.2	-394.1, 21.6
		Competition Step: Same Specialty Competition	27.1	-131.4, 185.6
	Sleep	Intercept	414.2	409.4, 419.0
		Week	0.7	0.2, 1.3
		Competition Sleep	6.6	-0.5, 13.6

		Competition Sleep : Same Institution Competition	-1.2	-9.5, 7.0
		Competition Sleep : Same Specialty Competition	0.4	-8.4, 9.2

CI confidence interval.

Supplementary Table 10: Sensitivity analyses for assessing different missing patterns on time-varying causal effect of competition on causal effect of competition on daily mood score (*100).

Missing pattern	Model	Parameter	Estimate	95% CI
Dropout	**Main-effect model**	Intercept	730.3	717.2, 743.4
		Week	-1.7	-2.4, -0.9
		Competition	2.7	-2.0, 7.4
	Time-varying-effect model	Intercept	725.8	712.3, 739.3
		Week	-0.8	-1.8, 0.2
		Competition	12.6	4.1, 21.1
		Week: Competition	-1.8	-3.2, -0.4
Weekly missingness	**Main-effect model**	Intercept	739.3	724.2, 754.4
		Week	-1.7	-2.4, -1.1
		Competition	2.1	-2.7, 6.9
	Time-varying-effect model	Intercept	734.6	719.2, 750.0
		Week	-0.9	-1.7, -0.0
		Competition	12.2	0.4, 20.2
		Week: Competition	-1.8	-3.1, -0.6

CI confidence interval.

Supplementary References

1 Boruvka A, Almirall D, Witkiewitz K, Murphy SA. Assessing Time-Varying Causal Effect Moderation in Mobile Health. *J Am Stat Assoc* 2018; **113**: 1112–21.

2 Khan WAA, Jackson ML, Kennedy GA, Conduit R. A field investigation of the relationship between rotating shifts, sleep, mental health and physical activity of Australian paramedics. *Sci Rep* 2021; **11**: 866.

3 Li W, Yin J, Cai X, Cheng X, Wang Y. Association between sleep duration and quality and depressive symptoms among university students: A cross-sectional study. *PLoS One* 2020; **15**: e0238811.