Published in final edited form as:
Health Informatics J. 2020 December ; 26(4): 2315–2331. doi:10.1177/1460458220902330.

Refining an algorithm-powered just-in-time adaptive weight control intervention: A randomized controlled trial evaluating model performance and behavioral outcomes

Stephanie P Goldstein, **J Graham Thomas**
The Warren Alpert Medical School of Brown University, USA

Gary D Foster,
WW (Weight Watchers), USA; University of Pennsylvania, USA

Gabrielle Turner-McGrievy,
University of South Carolina, USA

Meghan L Butryn,
Drexel University, USA

James D Herbert
University of New England, USA

Gerald J Martin, **Evan M Forman**
Drexel University, USA

Abstract

Suboptimal weight losses are partially attributable to lapses from a prescribed diet. We developed an app (OnTrack) that uses ecological momentary assessment to measure dietary lapses and relevant lapse triggers and provides personalized intervention using machine learning. Initially, tension between user burden and complete data was resolved by presenting a subset of lapse trigger questions per ecological momentary assessment survey. However, this produced substantial missing data, which could reduce algorithm performance. We examined the effect of more questions per ecological momentary assessment survey on algorithm performance, app utilization, and behavioral outcomes. Participants with overweight/obesity (n = 121) used a 10-week mobile weight loss program and were randomized to OnTrack-short (i.e. 8 questions/survey) or OnTrack-long (i.e. 17 questions/survey). Additional questions reduced ecological momentary assessment adherence; however, increased data completeness improved algorithm performance. There were

Corresponding author: Stephanie P Goldstein, Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University and The Miriam Hospital/Weight Control and Diabetes Research Center, 196 Richmond Street, Providence, RI 02909, USA. stephanie_goldstein@brown.edu.

Declaration of conflicting interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: G.D.F. reports salary support from Weight Watchers International during the conduct of the study. All other authors report no competing interests.

no differences in perceived effectiveness, app utilization, or behavioral outcomes. Minimal differences in utilization and perceived effectiveness likely contributed to similar behavioral outcomes across various conditions.

Keywords

diet; machine learning; mHealth; mobile health; weight loss

Background and significance

The worldwide prevalence of overweight and obesity among adults (aged 18+) is 39 and 13 percent, respectively.[1] Negative health correlates associated with overweight/obesity can be improved with modest weight loss through behavioral weight loss (BWL) programs.[2,3] BWL programs are successful to the extent that participants are able to adhere to core recommendations, in particular, dietary prescriptions.[4,5] However, participants in BWL have frequent “lapses” (i.e. instances of eating that violate dietary recommendations), which have been associated with poorer weight loss outcomes and higher attrition.[6–8]

To address lapses within BWL, our team developed a smartphone app (OnTrack) that predicts lapses from dietary recommendations and provides a targeted intervention(s) tailored to specific risk factors. OnTrack uses ecological momentary assessment (EMA), repeated assessment of variations in cognitive, emotional, physical, and environmental states, to assess dietary lapses and relevant triggers identified in previous literature.[7,9–11] OnTrack prompts participants to complete six EMA surveys per day. Each of the six surveys contains nine questions: one dietary lapse measure and eight questions drawn quasi-randomly (e.g. time of day and variable restrictions imposed based on theory and research support) from a pool of 17 questions. Each question assesses one potential trigger for lapse that had been identified in prior literature (e.g. mood, social environment, hunger). Thus, OnTrack assesses 17 unique triggers for lapse by repeatedly asking different lapse trigger questions throughout the day via EMA surveys.

As these data are collected in real time, OnTrack uses machine learning algorithms to continuously assess lapse risk. When the algorithm identifies that a user has crossed a predetermined threshold for risk of lapsing, a risk alert notification is delivered to the individual. The risk alert provides users with brief psychological or behavioral interventions (approximately 150–250 words), designed to address up to three identified risk factors. All interventions are made available as a “library” for users to access at any time. An open trial supported the feasibility, acceptability, and preliminary effectiveness of OnTrack.[12]

Problem statement

One disadvantage of OnTrack is that not all 17 lapse trigger questions are asked at each EMA survey, which means that OnTrack is not regularly evaluating all of the potential triggers for lapse. When developing OnTrack, we assumed that answering 17 lapse trigger questions, six times per day over a period of 10 weeks would be too burdensome.[13–15] We anticipated the burden would result in low adherence to the EMA surveys which would lead to fewer data from which to model lapse predictions. After completing the open trial, it became clear that adherence to the EMA surveys was much higher than anticipated (85.1%)[12] and that asking only a subset of the 17 lapse trigger questions created large amounts of systematic missing data. Given that the missing data could impair the OnTrack algorithm's performance,[16,17] we sought to formally evaluate an alternative procedure in which all 17 lapse trigger questions are presented at every EMA survey.

Objective

We examined the effect of administering additional lapse trigger questions per EMA survey on algorithm performance (primary outcome). Secondary outcomes of interest were app utilization (e.g. adherence to EMA surveys, intervention access) and behavioral outcomes (e.g. lapse frequency and percent weight loss (PWL)). We recruited individuals with overweight/obesity who were seeking to lose weight. Participants were prescribed the WW (formerly known as Weight Watchers) digital BWL program via mobile app and were randomly assigned to one of two OnTrack versions for 10 weeks: OnTrack-short (OT-S) contained 8 lapse trigger questions at each EMA survey as described above, and OnTrack-long (OT-L) contained all 17 lapse trigger questions at each EMA survey. We evaluated the following hypotheses: (1) OT-L would have superior algorithm performance (e.g. perceived effectiveness, model accuracy, sensitivity, and specificity), (2) OT-L would be non-inferior to OT-S regarding EMA survey adherence and superior with regard to app utilization (e.g. percentage of opened risk alerts, library interventions accessed), and (3) OT-L would have superior behavioral outcomes (e.g. PWL, lapse frequency) compared to OT-S.

Materials and methods

Participants

Individuals were recruited through listservs, press releases, radio, and social media advertisements from May 2017 through January 2018. All study procedures were conducted remotely. Eligible participants were those who (1) had a current body mass index (BMI) of 25–50 kg/m^2, (2) owned an internetwork operating system (iOS) device with data plan, (3) were willing to purchase or lent a wireless scale, (4) were 3 months stable on medication known to affect weight/appetite, (5) were aged 18–70 years, and (6) were living in the United States. Exclusion criteria were (1) enrollment in another structured weight loss program, (2) pregnancy or planning to become pregnant during the study, (3) endorsing disordered eating symptoms on a semi-structured clinical interview, (4) medical conditions contraindicating weight loss, (5) self-reporting weight loss of >5 percent over the previous 6 months, or (6) history of bariatric surgery.

Procedures

Interested participants completed an online screening questionnaire. Eligible participants were asked to purchase (or were lent by request) a low-cost Bluetooth wireless scale to automatically transmit weight data to the research team. Participants provided informed consent and completed the baseline survey after receiving their scale. Recruitment occurred on a rolling basis and participants were randomly assigned to condition (OT-S vs OT-L) using simple random assignment stratified by gender. Participants completed a 60-min phone call in which they set up their wireless scale and downloaded the apps onto their personal smartphone. Baseline weights were collected via the wireless scale under the researcher supervision during this phone call. Participants then completed a 60-min online tutorial for using OnTrack in conjunction with the WW app. Participants were instructed to use the mobile apps for 10 weeks. Unless participants reported technical issues, researcher contact was limited to one weekly reminder email to weigh-in. Participants received access to WW and OnTrack at no cost during the study. All procedures were conducted in accordance with the ethical standards of the Drexel University Institutional Review Board, which approved this study.

WW intervention

The WW mobile app, an evidence-based BWL program,[18,19] automatically determined a dietary points goal based on age, sex, weight, height, and activity levels. The WW SmartPoints® system assigned point values to foods based on calories, saturated fat, sugar, and protein. Some foods are assigned a zero point value to encourage the consumption of these foods. The SmartPoints goal was further segmented into point targets for daily meals and snacks within our study, which is common practice within BWL[20,21] and assisted in creating an objective definition of dietary lapse to aid participant reporting. Based on WW recommendations and successful pilot testing,[12,22] participants were allotted 15 percent of SmartPoints for breakfast, 25 percent for lunch, 40 percent for dinner, and two snacks of 10 percent each.

An unanticipated change occurred in the WW program mid-way through the study. The first set of participants received a program called “Beyond the Scale” which sets a daily “points” goal to achieve weight loss and emphasizes overall health and lifestyle improvements. The zero-point foods in that program were limited to fruits and vegetables. In December 2017, WW introduced a new program, “Freestyle,” which included a wide variety of zero-point foods (e.g. eggs, non-fat yogurt, seafood, lentils, corn) in addition to fruits and vegetables, reduced the points goal by approximately 20 percent, and added an option to roll over up to four unused points to the next day. A total of 37 ($n_{\text{OT-L}}$ = 17; $n_{\text{OT-S}}$ = 20) participants exclusively received Beyond the Scale, 43 ($n_{\text{OT-L}}$ = 22; $n_{\text{OT-S}}$ = 21) participants received a combination of Beyond the Scale and Freestyle, and 41 ($n_{\text{OT-L}}$ = 22; $n_{\text{OT-S}}$ = 19) participants exclusively received Freestyle. WW program type was added as a covariate in all analyses.

OnTrack mobile app

EMA survey procedure.—

EMA surveys measured dietary lapses and 17 lapse triggers (with each question assessing a different domain). Participants were prompted to complete EMA surveys six times per day and were encouraged to self-initiate an EMA survey after having lapsed. Question order was varied to discourage participants from adopting a habitual response pattern. See Figure 1 for an example of EMA survey. Wake/sleep settings were adjustable, so EMA surveys were delivered during waking hours. EMA survey data was uploaded to a web service for immediate remote access. In the OT-S condition, EMA surveys contained 1 dietary lapse question and 8 lapse trigger questions (that were quasi-randomly drawn from the pool of 17 lapse trigger questions).[23] In the OT-L condition, 1 dietary lapse question and all 17 lapse trigger questions were asked at each EMA survey.

Risk alerts.—The first 2 weeks of the study period were used to collect EMA surveys to personalize the lapse prediction algorithm, and risk alerts were enabled at the start of week 3. We used C4.5 to generate a decision tree algorithm of the OnTrack mobile app using Weka machine learning software (Version 3.9.3).[24] The algorithm predicted the likelihood of a lapse report at the next EMA survey (in approximately 2–3 h). Our previous research showed that combining group- and individual-level data resulted in the best model performance.[22] As such, we initiated the study using a base algorithm comprising data from previous trials, which was continuously refined via incoming participant data. When an EMA survey was completed, the algorithm classified responses as no risk (when a prediction was "no lapse"), low risk (probability of lapse >40%), medium risk (probability of lapse between 40% and 70%), or high risk (probability of lapse >70%). Least absolute shrinkage and selection operator coefficients were used to determine the top three lapse triggers. If lapse risk (regardless of level) was predicted by the algorithm, a risk alert notification communicated the top lapse trigger(s) along with three one-sentence descriptions of a coping strategy for each risk factor. Users could also select from the listed strategies to view a more in-depth description from OnTrack's intervention library.[23] See Figure 1 for the risk alert delivery system.

Intervention library.—The library contained 157 interventions, written by clinical psychology doctoral students and licensed clinical psychologists, that were based on empirically supported principles for behavior change[25] and tested in a previous trial.[12] Interventions typically addressed different lapse triggers via cognitive (e.g. restructuring negative thoughts), emotional (e.g. alternative coping strategies), behavioral (e.g. regularizing eating), and/or environmental (e.g. avoid problematic locations) strategies. The library was organized by different lapse triggers (e.g. boredom, fatigue) and within each category there were 8–10 interventions. Interventions were approximately 1–2 app screens and typically contained user interaction such as write-in text or checkboxes.

Measures

Demographics.—Age, sex, ethnicity, occupation, marital status, education level, and technology use were assessed via self-report questionnaire at baseline.

Weight and height.—Baseline, mid-treatment, and end-of-treatment weights were obtained using Yunmai Bluetooth Smart Scales that have been evaluated for validity and reliability.[26] The scales transmitted weights securely in real time through a web portal, which allowed the research team to track compliance. Participants were given the following instructions for self-weighing: remove shoes and extra layers of clothing, keep the scale on a hard surface and minimize movement, and weigh in the morning after using the restroom but before eating or drinking. If they did not weigh on the morning of their scheduled weigh-in, they were emailed a reminder to weigh the next day. Participant's self-reported height at baseline was a valid method for evaluating BMI.[27,28] PWL from baseline weight was calculated at mid- and end-of-treatment for use in analyses.

Dietary lapses.—Lapses from the WW diet were measured via OnTrack EMA survey. Participants self-reported the time and date of a lapse. A lapse was defined as an eating episode in which participants exceeded their allotted point target for a meal or snack. Overall lapse frequency for weeks 1, 5, and 10 were used for baseline, mid-, and end-of-treatment, respectively.

Lapse triggers.—A pool of 17 lapse triggers was assessed via OnTrack EMA surveys. Each lapse trigger was assessed via a single question. See Table 1 for a list of which lapse triggers were assessed and the scale of responses.

Adherence.—OnTrack automatically recorded the number of EMA surveys that were completed and delivered. Adherence was the ratio of total prompts completed to the total delivered.

Retention.—Participants were considered withdrawn if one or more of the following conditions were met: (1) submitted a written request for removal from the study, or (2) did not record a weight after week 1 and did not use OnTrack beyond the first week. Retention was the ratio of completers to total participants enrolled.

App utilization.—OnTrack recorded when a risk alert was delivered and viewed. Percentage of risk alerts viewed was calculated. OnTrack recorded library access by tracking the number of interventions accessed through the library.

Perceived effectiveness.—Participants were asked one end-of-day survey question via OnTrack inquiring about the helpfulness of the risk alerts delivered that day (i.e. “How did you feel about the risk alerts you received today?”). Response options were as follows: (1) “I received one or more risk alerts, and they were helpful”; (2) “I received one or more risk alerts, and only a portion were helpful”; (3) “I didn’t receive any risk alerts, and that seemed right”; (4) “I received one or more risk alerts, but they weren’t helpful”; and (5) “I didn’t receive any risk alerts, but I think I should have.” Response options (1)–(3) were coded as perceiving the alerts to be “helpful” and options (4)–(5) were coded as “unhelpful.”

Statistical analysis

Statistical Package for the Social Sciences (SPSS) version 24.0[29] and R version 3.4.0[30] were used for analysis. Data were evaluated for normality, outliers, and homoscedasticity to ensure that assumptions were met for general linear models and generalized linear mixed models. Descriptive statistics (means, standard deviations and frequencies, percentages) were calculated for the total sample and for each condition.

Algorithm performance.—Accuracy, sensitivity, and specificity of the OT-L and OT-S algorithm predictions were calculated by comparing outcomes predicted by OnTrack algorithms to lapse versus non-lapse events reported by participants while risk alerts were delivered (weeks 3–10). A complicating factor of assessing model outcomes is that cases in which lapse predictions are followed by a non-lapse could either be false positives or successfully prevented lapses. Consistent with our prior work,[12] we presumed that opening a risk alert likely resulted in preventing a lapse and re-coded non-lapses after opened risk alerts into lapses (2560 instances re-coded in OT-S and 2979 instances in the OT-L conditions). We calculated true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), and a chi-square test of independence was used to examine counts of TP, FP, TN, and FN across conditions.

A generalized linear mixed model with a negative binomial distribution and logit link function was used to evaluate the effect of condition on perceived effectiveness. Fixed effects included treatment condition, WW program type, and number of alerts opened. Random effects were subject and the slope of time (represented as weeks since baseline).

App utilization.—An independent groups two one-sided test (TOST) procedure was used to examine the non-inferiority of OT-L with regard to EMA survey adherence. Based on prior studies, we set an a priori non-inferiority margin of 10 percent for adherence.[31,32] A generalized linear mixed model with a binomial distribution and logit link function was used to evaluate the effect of condition on the likelihood of opening a risk alert. Fixed effects included treatment condition, WW program type, and total risk alerts delivered. Random effects include the subject and the slope of time (represented as weeks since baseline). Time was allowed to interact with treatment condition. A one-way analysis of variance (ANOVA) with fixed effect of condition and controlling for WW program type was used to evaluate the effect of condition on library access.

Behavioral outcomes.—A general linear model was used to evaluate the effect of condition and time on PWL at mid-treatment, and final assessment points, controlling for baseline BMI, age, gender, race/ethnicity (non-Hispanic White vs others), and WW program type. Missing weights were imputed using baseline weight carried forward, thus assuming zero weight loss for participants who dropped out or did not complete end-of-treatment weigh-in.[33] A similar model was used to evaluate the effect of condition and time on lapse frequency at mid- and end-of-treatment assessment points, controlling for baseline lapse frequency and EMA survey adherence.

Power analysis

A post hoc power analysis revealed that our sample yielded 20,504 predictions ($df = 3$, $w = 0.84$) to evaluate algorithm performance with a chi-square test of independence. We were adequately powered (power = 1.00) to detect significant differences between the OT-S and OT-L group in TP, TN, FP, and FN counts with an alpha value of 0.05.

Results

Participants

See Figure 2 for the Consolidated Standards of Reporting Trials (CONSORT) diagram depicting participant flow through the trial. Analyses represent data from 121 participants allocated to treatment condition. Table 2 illustrates participant characteristics, retention, and online tutorial completion.

Algorithm performance

Model outcomes.—OnTrack demonstrated an overall accuracy of 79.8 percent of lapse prediction (79.7% in OT-S vs 79.9% in OT-L), a sensitivity of 74.5 percent (71.6% in OT-S vs 77.7% in OT-L), and a specificity of 83.1 percent (84.4% in OT-S vs 81.7% in OT-L). The OT-L outperformed OT-S with regard to TP (3214 vs 2791) and FN (922 vs 1105) predictions, with OT-S producing greater TN predictions (5710 vs 4659), $\chi^2(3) = 120.29$, and $p < 0.001$.

Perceived effectiveness.—A total of 3321 questions regarding the helpfulness of risk alerts were answered (OT-S = 1761 vs OT-L = 1560). Participants answered, on average, 32.24 questions each (standard deviation (SD) = 15.23) and most (72.84%) identified risk alerts as helpful/accurate. As seen in Table 3, condition did not impact perceived effectiveness.

App utilization

Across the 116 people who used OnTrack, average EMA survey adherence in OT-S was 65.4 percent (SD = 25.5) in OT-S and 60.5 percent (SD = 29.7) in OT-L. See Figure 3 for a depiction of EMA survey adherence by condition over time. The TOST procedure revealed that the upper limits of the 90% confidence interval (CI) for adherence during treatment ($M = 4.9$; 90% CI = −3.8 to 13.5) did exceed the a priori defined non-inferiority margin ($\Delta = 10.0$) and corresponded with a non-significant test result, $t(114) = -0.98$ and $p = 0.16$, indicating that non-inferiority cannot be supported at the 90% CI.

A total of 13,037 risk alerts were delivered ($M = 118.5$ alerts per person; SD = 52.8). On average, participants opened 46.9 percent (SD = 26.2) of risk alerts. See Figure 4 for risk alerts opened by study week across condition ($M = 7.80$, SD = 0.77). There was no statistically significant between-groups difference observed in the likelihood of opening a risk alert (see Table 3).

Participants opened an average of 8.60 library interventions (SD = 11.54). Intervention library access did not differ significantly between OT-S ($M = 6.58$, SD = 8.89 interventions) and OT-L ($M = 10.62$, SD = 13.45), $F(1) = 3.53$, $p = 0.06$, and partial $\eta^2 = 0.04$.

Behavioral outcomes

Participants lost an average of 2.6 percent (SD = 4.51) of their starting weight at mid-treatment and 3.4 percent (SD = 3.72) at end-of-treatment. General linear models (see Table 4) revealed that the effect of condition was not statistically significant at mid-treatment (−2.8 ± 2.5 PWL in OT-S vs −2.3 ± 5.9 PWL in OT-L) or at end-of-treatment (−3.3 ± 3.5 PWL in OT-S vs −3.5 ± 3.9 PWL in OT-L).

A total of 5058 lapses were reported through completion of 4319 prompted EMA surveys and 739 user-initiated EMA surveys. Participants reported an average of 4.36 lapses per week (SD = 1.46). At mid-treatment, models revealed that the effect of condition was not statistically significant (4.5 ± 4.4 lapses in OT-S vs 3.8 ± 4.0 lapses in OT-L) and a similar pattern was observed during end-of-treatment (week 10; 3.3 ± 4.4 lapses in OT-S vs 3.1 ± 4.4 in OT-L). See Table 4 for model outcomes.

Discussion

OnTrack is a companion app that was developed to enhance BWL outcomes by predicting lapses from a diet and delivering preventive intervention in moments of greatest need. Preliminary studies showed OnTrack to be feasible and acceptable; however, the EMA procedure of quasi-randomly assessing lapse triggers at each EMA survey (due to concerns related to participant burden) resulted in substantial missing data. We were uncertain about the degree to which missing data impacted algorithm accuracy and hence its effectiveness. Therefore, this trial compared the original version of OnTrack (i.e. eight lapse trigger questions per EMA survey; OT-S) to another version of OnTrack in which all lapse trigger questions were administered at each EMA survey (i.e. 17 lapse trigger questions per survey; OT-L). Given high adherence to EMA surveys in prior trials, we hypothesized that the additional lapse trigger questions would not substantially impact EMA survey adherence. The subsequent increase in data completeness was expected to improve the accuracy of the algorithm (as measured by algorithm performance and perceived effectiveness) and therefore promote greater utilization of the app and improved behavioral outcomes.

Our primary outcome of interest was the performance of the algorithm with regard to lapse prediction. OT-L did show comparable accuracy to OT-S with improved sensitivity (true positive rate). We had hypothesized that increased data completeness with similar EMA survey adherence in the OT-L condition would be the factors that contributed to enhanced algorithm performance. Interestingly, and contrary to our hypotheses, OT-L was statistically inferior to OT-S in EMA survey adherence. Though unexpected, the finding is consistent with prior research showing that increasing participant burden negatively impacts adherence.[13–15] The inferiority of OT-L meant that additional lapse trigger questions in the EMA surveys led to a ~5 percent drop in EMA survey adherence, thus resulting in approximately 1000 more EMA surveys being completed in the OT-S condition (~17 surveys per person). However, the EMA surveys completed in OT-L condition had almost twice the data points as OT-S despite the inferior compliance, which ultimately resulted in more data to model predictions. It is likely that the greater quantity of data produced by the OT-L condition, irrespective of the inferior EMA adherence, is what contributed to enhanced algorithm performance.[34–36]

Despite the objective improvements in algorithm performance, there were no differences in the perceived effectiveness of risk alerts. Therefore, it is logical that there were also no between-groups differences in utilization of features such as risk alerts and the intervention library. Of note, the downward trajectory of app utilization and adherence to EMA surveys appears consistent with prior research on smartphone-based treatments and EMA methods.[13,37,38] Future research may benefit from specifically assessing factors that influence engagement with mobile BWL interventions, such as perceived usefulness, engagement motivations, satisfaction, and perceived burden,[39–41] as accessing app-delivered content can facilitate greater weight loss.[40,42]

With regard to behavioral outcomes, participants lapsed about four times per week on average and lost 3.4 percent of their body weight, which is consistent with our prior work.[7,12,43] There were no significant differences in PWL or lapse frequency between conditions at either assessment point and the effect sizes for the observed associations were small. Given similarities in app utilization, it stands to reason that there would be no clinically meaningful differences in behavioral outcomes.[40,41,44,45]

Overall, results indicate that burden incurred by additional lapse trigger questions did dissuade participants completing EMA surveys; however, the completeness of data improved prediction accuracy. It is possible that the improvement was not clinically meaningful enough to be perceptible to participants, as there were no differences in perceived effectiveness, app utilization, and behavioral outcomes. The benefit of using OT-S in future trials is that individuals will sustain less burden which could conceivably lead to less drop-out, greater EMA adherence, and better app utilization in a longer trial of OnTrack. However, the drawback is that missing data does appear to impact algorithm accuracy which could impair further iterations of model optimization. A compromise could be to examine which lapse trigger questions are most likely to elicit a response from the user and then use this outcome to inform additional streamlining of the EMA surveys in future studies. Another approach could be to individually tailor lapse trigger questions at each EMA survey based on the predictive value of each trigger, and this can be done using additional machine learning algorithms. Both alternatives could be viable methods for enhancing algorithm quality and user engagement.

Strengths and limitations

Within the weight control literature, this is one of the first randomized controlled trials of a just-in-time intervention mobile app. Strengths include the use of Bluetooth scales to collect weights, use of a scalable and evidence-based BWL program, and EMA to measure outcomes of interest (i.e. lapses). This study also had four notable limitations. First, the study was not powered to detect small between-groups effects and most of the effects observed in our sample were small. Second, the study suffered from self-selection bias in that we recruited individuals who were interested in an app-based, self-guided treatment approach. Future research should investigate methods for targeting and engaging treatment-resistant individuals. Third, the method of re-coding lapses likely biased our algorithm by artificially increasing sensitivity to a lapse prediction. One way that we attempted to balance this was to alter the cost-sensitive parameters to reduce false positive predictions. Overall, it was difficult to ascertain algorithm performance when directly intervening on the outcome variable and this complicated our ability to draw firm conclusions about the predictive quality of the algorithm. Future research may consider semi-regular non-intervention periods or micro-randomized designs to properly calibrate algorithms.[46] Fourth, inherent in the use of machine learning is the “black box” element in which the data guide intervention provision rather than a priori defined evidence-based theoretical models.[47] Future studies should consider different statistical procedures and methodologies (such as micro-randomized designs or control systems engineering) to develop and validate a theoretical model for momentary interventions on eating behaviors.[46,48]

Conclusion

Our results demonstrate the importance of conducting optimization research when developing mobile health interventions. Namely, the design allowed us to identify methods for enhancing prediction of lapses while maintaining an acceptable level of participant burden. Future studies may also consider conducting a component analysis to examine which OnTrack features (or combinations of features) are having the strongest impact on behavioral outcomes. Finally, future versions of the app may benefit from automating assessment to relieve burden and improve accuracy. For instance, passive sensing (e.g. global positioning system (GPS), geographical information system (GIS), accelerometer) and wearable technologies could assess the presence of lapse triggers and possibly eating behavior itself.[49,50]

Acknowledgements

We would like to acknowledge Brandon Masterson and Ruslan Askarov for assistance with programming the smartphone applications and online tools used to carry out the current study. We would also like to acknowledge Alexis Wojtanowski for her assistance with disseminating the WW mobile program to participants in this trial.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Weight Watchers Karen Miller-Kovach Research Grant from The Obesity Society (unnumbered) and the Drexel Ventures Innovation Fund (unnumbered).

References

1. World Health Organization. Obesity and overweight, 2018, https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
2. Norris SL, Zhang X, Avenell A, et al. Long-term effectiveness of lifestyle and behavioral weight loss interventions in adults with type 2 diabetes: a meta-analysis. Am J Med 2004; 117(10): 762–774. [PubMed: 15541326]
3. Wing RR, Lang W, Wadden TA, et al. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. Diabetes Care 2011; 34(7): 1481–1486. [PubMed: 21593294]
4. Klem ML, Wing RR, McGuire MT, et al. A descriptive study of individuals successful at long-term maintenance of substantial weight loss. Am J Clin Nutr 1997; 66(2): 239–246. [PubMed: 9250100]
5. Shick SM, Wing RR, Klem ML, et al. Persons successful at long-term weight loss and maintenance continue to consume a low-energy, low-fat diet. J Am Diet Assoc 1998; 98(4): 408–413. [PubMed: 9550162]
6. Drapkin RG, Wing RR and Shiffman S. Responses to hypothetical high risk situations: do they predict weight loss in a behavioral treatment program or the context of dietary lapses. Health Psychol 1995; 14(5): 427–434. [PubMed: 7498114]
7. Forman EM, Schumacher LM, Crosby R, et al. Ecological momentary assessment of dietary lapses across behavioral weight loss treatment: characteristics, predictors, and relationships with weight change. Ann Behav Med 2017; 51(5): 741–753. [PubMed: 28281136]
8. Deiches JF, Baker TB, Lanza S, et al. Early lapses in a cessation attempt: lapse contexts, cessation success, and predictors of early lapse. Nicotine Tob Res 2013; 15(11): 1883–1891. [PubMed: 23780705]
9. Carels RA, Douglass OM, Cacciapaglia HM, et al. An ecological momentary assessment of relapse crises in dieting. J Consult Clin Psychol 2004; 72(2): 341–348. [PubMed: 15065966]
10. Carels RA, Hoffman J, Collins A, et al. Ecological momentary assessment of temptation and lapse in dieting. Eat Behav 2001; 2(4): 307–321. [PubMed: 15001025]

11. McKee HC, Ntoumanis N and Taylor IM. An ecological momentary assessment of lapse occurrences in dieters. Ann Behav Med 2014; 48(3): 300–310. [PubMed: 24562984]
12. Forman EM, Goldstein SP, Zhang F, et al. OnTrack: development and feasibility of a smartphone app designed to predict and prevent dietary lapses. Transl Behav Med 2019; 9: 236–245. [PubMed: 29617911]
13. Laing BY, Mangione CM, Tseng C-H, et al. Effectiveness of a smartphone application for weight loss compared with usual care in overweight primary care patients: a randomized, controlled trial. Ann Intern Med 2014; 161(Suppl. 10): S5–S12. [PubMed: 25402403]
14. Siopis G, Chey T and Allman Farinelli M. A systematic review and meta-analysis of interventions for weight management using text messaging. J Hum Nutr Diet 2015; 28: 1–15.
15. Tang J, Abraham C, Stamp E, et al. How can weight-loss app designers' best engage and support users? A qualitative investigation. Br J Health Psychol 2015; 20(1): 151–171. [PubMed: 25130682]
16. Lee JH and Huber J. Multiple imputation with large proportions of missing data: How much is too much? United Kingdom Stata Users' group meetings 2011. Stata Users Group, 2011, https://ideas.repec.org/p/boc/usug11/23.html
17. Ziegler ML. Variable selection when confronted with missing data. University of Pittsburgh, 2006, http://d-scholarship.pitt.edu/9122/1/Ziegler_2006.pdf
18. Thomas JG, Raynor H, Bond D, et al. Weight loss and frequency of body-weight self-monitoring in an online commercial weight management program with and without a cellular-connected "smart" scale: a randomized pilot study. Obes Sci Pract 2017; 3(4): 365–372. [PubMed: 29259794]
19. Thomas JG, Raynor HA, Bond DS, et al. Weight loss in weight watchers online with and without an activity tracking device compared to control: a randomized trial. Obesity 2017; 25(6): 1014–1021. [PubMed: 28437597]
20. Wing RR. Behavioral weight control. Handbook Obes Treat 2002; 2: 301–317.
21. Wadden TA and Foster GD. Behavioral treatment of obesity. Med Clin North Am 2000; 84: 441–461. [PubMed: 10793651]
22. Goldstein SP, Zhang F, Thomas JG, et al. Application of machine learning to predict dietary lapses during weight loss. J Diabetes Sci Technol 2018; 12(5): 1045–1052. [PubMed: 29792067]
23. Goldstein SP, Evans BC, Flack D, et al. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. Int J Behav Med 2017; 24(5): 673–682. [PubMed: 28083725]
24. Chauhan H and Chauhan A. Implementation of decision tree algorithm c4.5. Int J Sci Res Pub 2013; 3: 1–3.
25. Abraham C and Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychology 2008; 27: 379. [PubMed: 18624603]
26. Goldstein SP. Comparing effectiveness and user behaviors of two versions of a just-in-time adaptive weight loss smartphone app. Philadelphia, PA: Drexel University, 2018.
27. Spencer EA, Appleby PN, Davey GK, et al. Validity of self-reported height and weight in 4808 EPIC–Oxford participants. Public Health Nutr 2002; 5(4): 561–565. [PubMed: 12186665]
28. Cui Z, Stevens J, Truesdale KP, et al. Prediction of body mass index using concurrently self-reported or previously measured height and weight. PLoS ONE 2016; 11(11): e0167288. [PubMed: 27898706]
29. Corp I. IBM SPSS statistics for Macintosh, Version 24.0. Armonk, NY: IBM Corp, 2016.
30. Team RC. R: A Language and Environment for Statistical Computing. 3.4.0 ed. Vienna, Austria: R Foundation for Statistical Computing, 2013.
31. Thompson-Felty C and Johnston CS. Adherence to diet applications using a smartphone was associated with weight loss in healthy overweight adults irrespective of the application. J Diabet Sci Technol 2017; 11: 184–185.
32. Wharton CM, Johnston CS, Cunningham BK, et al. Dietary self-monitoring, but not dietary quality, improves with use of smartphone app technology in an 8-week weight loss trial. J Nutrition Educ Behav 2014; 46: 440–444. [PubMed: 25220777]

33. Ware JH. Interpreting incomplete data in studies of diet and weight loss. N Engl J Med 2003; 348(21): 2136–2137. [PubMed: 12761370]

34. Saar-Tsechansky M and Provost F. Handling missing values when applying classification models. J Mach Learn Res 2007; 8: 1623–1657.

35. Enders CK. Applied missing data analysis. New York: Guilford Press, 2010.

36. Acuna E and Rodriguez C. The treatment of missing values and its effect on classifier accuracy. In: Banks D, McMorris FR, Arabie P, et al. (eds) Classification, clustering, and data mining applications. New York: Springer, 2004, pp. 639–647.

37. Shiffman S Designing protocols for ecological momentary assessment. In: Stone AA, Shiffman S, Atienza A, et al. (eds.) The Science of real-time data capture: Self-reports in health research. New York: Oxford University Press, 2007, pp. 27–53.

38. Jacobs S, Radnitz C and Hildebrandt T. Adherence as a predictor of weight loss in a commonly used smartphone application. Obes Res Clin Pract 2017; 11(2): 206–214. [PubMed: 27292942]

39. Kim YH, Kim DJ and Wachter K. A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. Decis Support Syst 2013; 56: 361–370.

40. Turner LD, Allen SM and Whitaker RM. Reachable but not receptive: enhancing smartphone interruptibility prediction by modelling the extent of user engagement with notifications. Pervasive Mobile Comput 2017; 40: 480–494.

41. Suh H, Shahriaree N, Hekler EB, et al. Developing and validating the user burden scale: a tool for assessing user burden in computing systems. In: Proceedings of the 2016 Chi conference on human factors in computing systems, San Jose, CA, 7–12 May 2016, pp. 3988–3999. New York: ACM.

42. Turner-McGrievy G and Tate D. Tweets, apps, and pods: results of the 6-month mobile pounds off digitally (Mobile POD) randomized weight-loss intervention among adults. J Med Internet Res 2011; 13(4): e120. [PubMed: 22186428]

43. Goldstein S A preliminary investigation of a personalized risk alert system for weight control lapses. Philadelphia, PA: Drexel University, 2016.

44. Serrano KJ, Yu M, Coa KI, et al. Mining health app data to find more and less successful weight loss subgroups. J Med Internet Res 2016; 18(6): e154. [PubMed: 27301853]

45. Turner-McGrievy GM and Tate DF. Are we sure that mobile health is really mobile? An examination of mobile device use during two remotely-delivered weight loss interventions. Int J Med Inform 2014; 83(5): 313–319. [PubMed: 24556530]

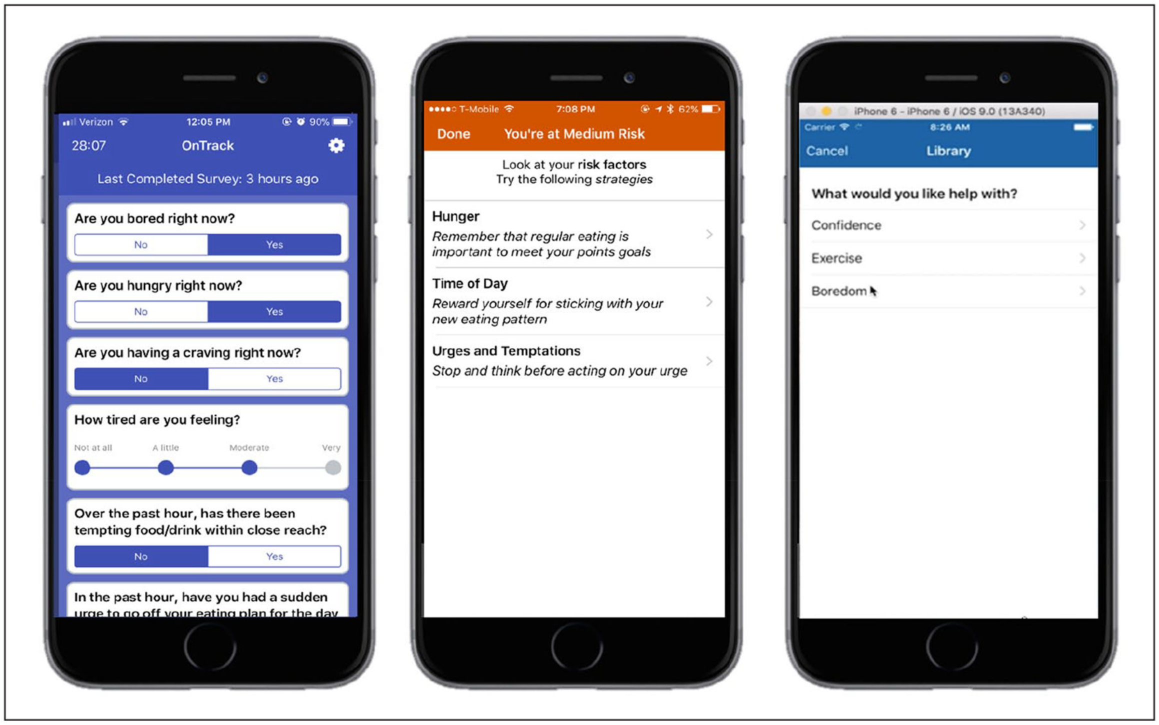
46. Liao P, Klasnja P, Tewari A, et al. Micro-randomized trials in mHealth, 2015, arXiv preprint, arXiv: 1504.00238.

47. London A Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 2019; 49(1): 15–21.

48. Rivera D, Martin C, Timms K, et al. Control systems engineering for optimizing behavioral mHealth interventions. In: Rehg J, Murphy S and Kumar S (eds) Mobile Health. Cham: Springer, 2017, pp. 455–493.

49. Desendorf J, Bassett DR Jr, Raynor HA, et al. Validity of the bite counter device in a controlled laboratory setting. Eat Behav 2014; 15(3): 502–504. [PubMed: 25064306]

50. Alharbi R, Vafaie N, Liu K, et al. Investigating barriers and facilitators to wearable adherence in fine-grained eating detection. In: Proceedings of the 2017 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), Kona, HI, 13–17 March 2017, pp. 407–412. New York: IEEE.



Last Completed Survey: 3 hours ago

Are you bored right now?
No [✓ Yes]

Are you hungry right now?
No [✓ Yes]

Are you having a craving right now?
[✓ No] Yes

How tired are you feeling?
Not at all A little [✓ Moderate] Very

Over the past hour, has there been tempting food/drink within close reach?
[✓ No] Yes

In the past hour, have you had a sudden urge to go off your eating plan for the day

You're at Medium Risk

Look at your **risk factors**
Try the following *strategies*

Hunger
Remember that regular eating is important to meet your points goals

Time of Day
Reward yourself for sticking with your new eating pattern

Urges and Temptations
Stop and think before acting on your urge

Library

What would you like help with?

Confidence

Exercise

Boredom

Figure 1.

OnTrack screenshots of an EMA survey, a risk alert, and the intervention library.

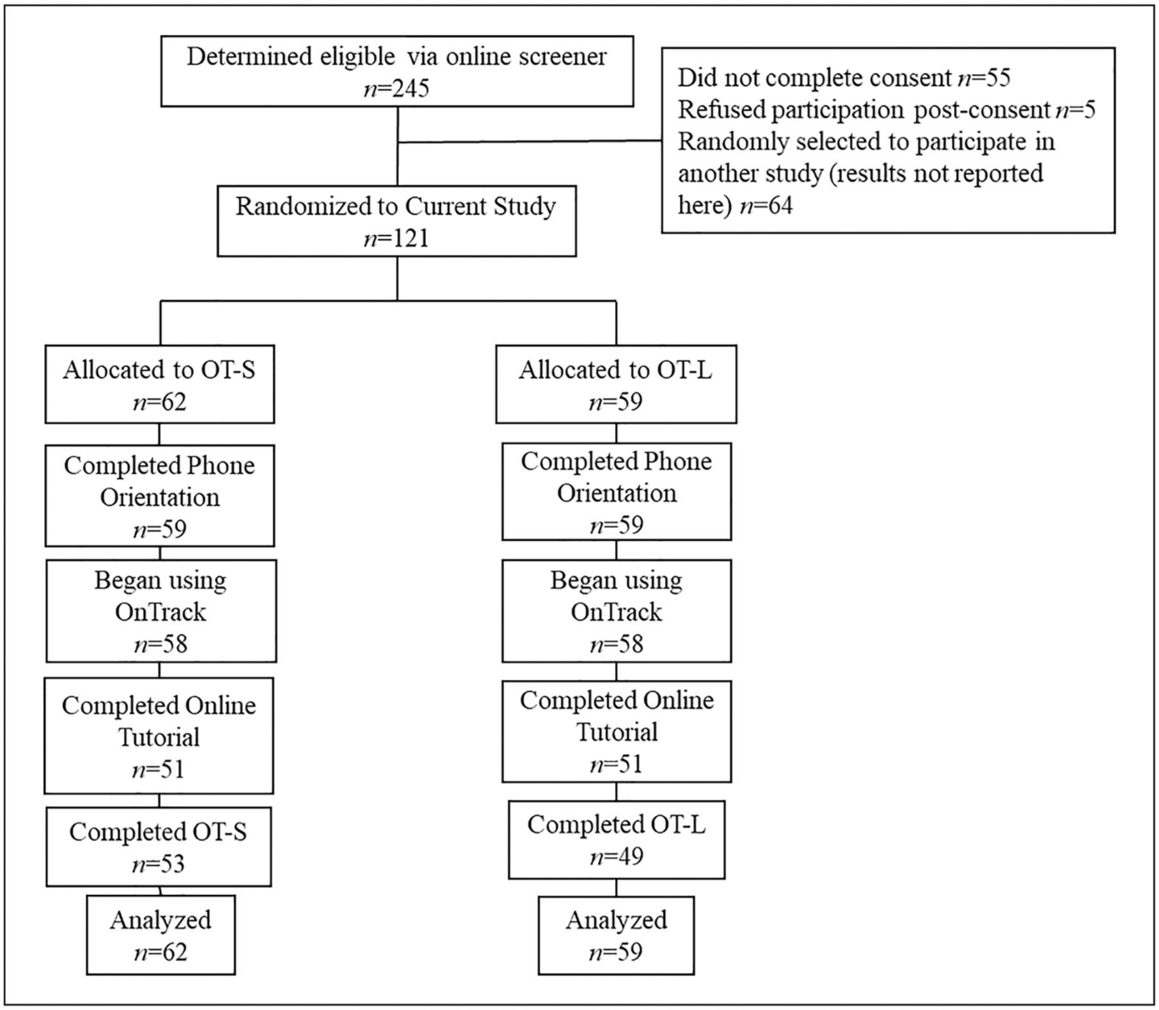

Determined eligible via online screener
n=245
Did not complete consent n=55
Refused participation post-consent n=5
Randomly selected to participate in another study (results not reported here) n=64
Randomized to Current Study
n=121
Allocated to OT-S
n=62
Allocated to OT-L
n=59
Completed Phone Orientation
n=59
Completed Phone Orientation
n=59
Began using OnTrack
n=58
Began using OnTrack
n=58
Completed Online Tutorial
n=51
Completed Online Tutorial
n=51
Completed OT-S
n=53
Completed OT-L
n=49
Analyzed
n=62
Analyzed
n=59

Figure 2.
CONSORT diagram.

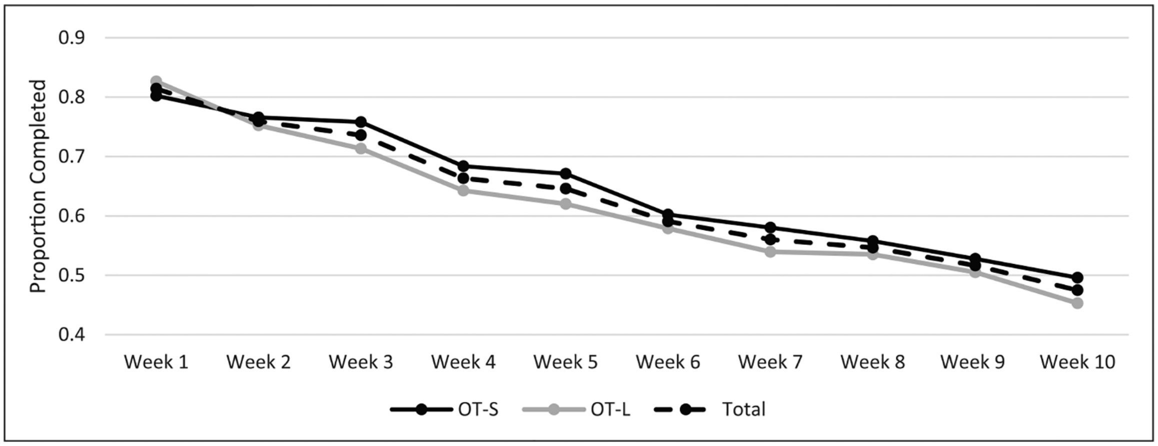


Figure 3.

Average proportion of completed EMA surveys over time.

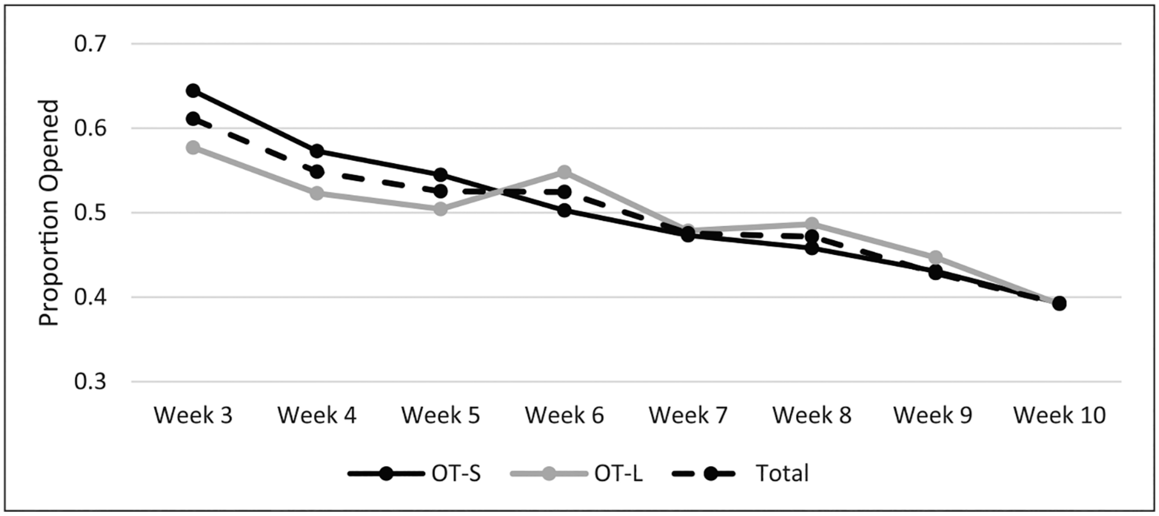


Figure 4.

Average proportion of opened risk alerts over time.

Table 1.

Lapse triggers assessed via EMA.

Lapse trigger	Type	Answer choices
Affect	5-point Likert-type scale	1 = I am in an especially good mood
		2 = I am in a good mood
		3 = I feel slightly stressed/upset
		4 = I feel very stressed/upset
		5 = I feel intensely stressed/upset
Sleep	7-point sliding scale	1 ⩽ 4 h
		7 ⩾ 10 h
Fatigue	7-point sliding scale	1 = Not at all
		7 = Extremely
Hunger	Dichotomous	Yes/No
Motivation to adhere to diet	Dichotomous	Yes/No
Cravings	Dichotomous	Yes/No
Boredom	Dichotomous	Yes/No
Temptation	Dichotomous	Yes/No
Cognitive load	5-point Likert-type scale	1 = Requiring almost no mental effort
		5 = Requiring almost all of my mental effort
Confidence	7-point sliding scale	1 = Not at all
		7 = Extremely
Socializing	3-point Likert-type scale	1 = No
		2 = Yes, without food present
		3 = Yes, with food present
Television	Dichotomous	Yes/No
Negative interpersonal interactions	Dichotomous	Yes/No
Presence of tempting foods	Dichotomous	Yes/No
Food Advertisements	Dichotomous	Yes/No
Planning food	3-point Likert-type scale	1 = Not at all
		3 = Somewhat
		5 = Perfectly
Alcohol	Dichotomous	Yes/No
Time	Automatic, lapses self-reported	None

Table 2.

Participant characteristics at baseline, retention, and tutorial completion.

	OT-L (*n* = 59)	OT-S (*n* = 62)	Total (*n* = 121)
Sex, no. (%)			
Men	8 (13.6)	11 (17.7)	19 (15.7)
Women	51 (86.4)	51 (82.3)	102 (84.3)
Age, mean (SD), years	47.19 (13.09)	47.27 (13.77)	47.23 (13.39)
Race, no. (%)[a]			
Black/African American	8 (12.9)	6 (10.2)	14 (11.5)
African	0 (0.0)	2 (3.4)	2 (1.7)
Asian American	1 (1.6)	2 (3.4)	3 (2.5)
Asian/Pacific Islander	4 (6.5)	0 (0.0)	4 (3.3)
White	42 (67.7)	47 (79.7)	89 (73.6)
European	1 (1.6)	1 (1.7)	2 (1.7)
Other	2 (3.2)	3 (5.1)	5 (4.1)
Ethnicity, no. (%)			
Hispanic/Latino(a)	7 (11.3)	5 (8.5)	12 (9.9)
Non-Hispanic/Latino(a)	55 (88.7)	54 (91.5)	109 (90.1)
Employment, no. (%)			
Full-time	41 (69.5)	44 (71.0)	85 (70.2)
Part-time	7 (11.9)	9 (14.5)	16 (13.2)
Per diem	4 (6.8)	2 (3.2)	6 (5.0)
Disability/SSI	0 (0)	2 (3.2)	2 (1.7)
No income	7 (11.9)	5 (8.1)	12 (9.9)
Student, no. (%)			
Not a student	49 (83.1)	52 (83.9)	101 (83.5)
Full-time	5 (8.1)	6 (10.2)	11 (9.1)
Part-time	5 (8.1)	4 (6.8)	9 (7.44)
Marital status, no. (%)			
Single	9 (15.3)	15 (24.2)	24 (19.8)
Married	35 (59.3)	34 (54.8)	69 (57.0)
Living with partner	4 (6.8)	4 (6.5)	8 (6.6)
Not living with partner	5 (8.5)	5 (4.1)	10 (8.3)
Divorced	4 (6.8)	3 (4.8)	7 (5.8)
Widowed	2 (3.4)	1 (1.6)	3 (2.5)
App use, no. (%)			
Once per month	0 (0.0)	1 (1.6)	1 (0.8)
Daily	30 (50.8)	22 (35.5)	52 (43.0)
More than once/day	29 (49.2)	39 (62.9)	68 (56.2)
Weight, mean (SD), kg	92.87 (17.69)	96.81 (20.15)	94.86 (18.99)
Body mass index, mean (SD), kg/m^2	34.08 (5.14)	34.94 (6.27)	34.52 (5.73)
Retention, no. (%)	49 (83.10)	53 (85.48)	102 (84.29)

	OT-L (*n* = 59)	**OT-S (*n* = 62)**	**Total (*n* = 121)**
Online tutorial completion, no. (%)	48 (82.3)	54 (86.4)	102 (84.29)

OT-L: OnTrack-long; OT-S: OnTrack-short; SD: standard deviation; SSI: Social Security Income.

[a] Proportion not equal to 1 due to multiple selection.

Table 3.

Results of generalized linear mixed models.

	B	**SE**	**z-value**	***p* value**
Opening a risk alert				
Condition[a]	−0.37	0.24	−1.56	0.12
Total alerts delivered	0.01	0.002	4.87	<0.001 **
WW program: mixed[b]	−0.56	0.26	−2.13	0.03 *
WW program: freestyle[b]	−0.91	0.29	−3.13	0.002 **
Study week	−0.22	0.06	−3.89	<0.001 **
Condition[a] × study week	0.02	0.04	0.62	0.54
Perceived effectiveness				
Condition[a]	0.16	0.14	1.11	0.27
Total alerts opened	−0.01	0.001	−4.29	<0.001 **
WW program: mixed[b]	−0.26	0.14	−1.85	0.06
WW program: freestyle[b]	0.06	0.16	0.36	0.72
Study week	−0.03	0.02	−1.67	0.09
Condition[a] × study week	−0.03	0.03	−1.31	0.18

WW: Weight Watchers; SE: standard error.

[a] Reference group: OT-S.

[b] Reference group: beyond the scale.

* $p < 0.05$.

** $p \leqslant 0.01$.

Table 4.

Tests of between-subjects effects for percent weight loss and lapse frequency.

	Mid-treatment (week 5) models					End-of-treatment (week 10) models				
	MS	*df*	*F*	*p*	Partial η^2	MS	*df*	*F*	*p*	Partial η^2
Percent weight loss										
Condition	0.003	1	1.66	0.20	0.02	<0.001	1	0.05	0.82	0.001
Time	0.003	13	1.42	0.16	0.17	0.002	12	1.43	0.17	0.16
Condition × time	0.004	9	1.93	0.06	0.16	0.001	6	0.69	0.66	0.04
Baseline BMI	0.001	1	0.29	0.59	0.003	0.001	1	0.80	0.37	0.009
WW program type	<0.001	1	0.16	0.69	0.002	0.001	1	0.57	0.94	0.006
Race/ethnicity	0.006	1	3.09	0.08	0.03	0.005	1	4.11	0.05*	0.04
Age	<0.001	1	0.11	0.74	0.001	<0.001	1	0.01	0.94	<0.001
Gender	0.011	1	5.63	0.02*	0.06	0.005	1	3.97	0.05*	0.04
Lapse frequency										
Error	0.002	89	–	–	–	0.001	92	–	–	–
Condition	6.92	1	0.59	0.44	0.005	0.47	1	0.04	0.85	<0.001
WW program type	1.34	1	0.12	0.73	0.001	5.91	1	0.47	0.49	0.004
Age	22.49	1	1.93	0.17	0.02	0.17	1	0.01	0.91	<0.001
Gender	16.99	1	1.46	0.23	0.01	12.74	1	1.00	0.32	0.009
Survey compliance	123.99	1	10.62	0.001**	0.09	157.19	1	12.39	0.001**	0.10
Race/ethnicity	1.61	1	0.14	0.71	0.001	38.24	1	3.02	0.09	0.03
Baseline lapses	376.77	1	32.27	<0.001**	0.23	414.49	1	32.67	<0.001**	0.23
Error	11.67	108	–	–	–	0.001	92	–	–	–

BMI: body mass index; WW: Weight Watchers; MS: mean square; *df*: degrees of freedom.

* $p < 0.05$.

** $p \leqslant 0.01$.