Published in final edited form as:

GetMobile. 2018 June ; 22(2): 11–14. doi:10.1145/3276145.3276149.

Rethinking Evaluations Of mHealth Systems For Behavior Change

Predrag Klasnja and
University of Michigan

Eric B. Hekler
University of California San Diego

Much of the recent research in mobile health (mHealth) has focused on the development of apps and wearables for promoting healthy behavior changes, such as losing weight, increasing physical activity, or adhering to a medication regimen. These interactive systems help users make changes in their behavior by, for instance, tracking health-related activities and states, providing feedback, helping users set and track goals, and facilitating supportive social interactions. We refer to the features that implement such functionality as the system's "intervention components," as they are designed to actuate psychosocial mechanisms (e.g., modeling, self-efficacy, positive reinforcement, etc.) thought to mediate the behavior change process. As with any other type of behavioral intervention, mHealth systems are only effective for some users and some of the time, but insofar as they do work, they do so mainly through the mechanisms of change that are activated via users' interactions with the system's intervention components.

But while mHealth systems are typically designed to target specific mechanisms of change, whether those mechanisms are successfully activated is rarely explicitly evaluated. In the best case scenario, an mHealth system is evaluated in a randomized controlled trial (RCT) intended to establish whether the system as a whole—as an "intervention package"—is effective at changing target behaviors, at least temporarily. And many mHealth systems are only evaluated in pre-post studies that are not able to robustly establish efficacy even at the package level. In either case, insights about which specific aspects of the system contributed to its effectiveness are obtained only indirectly, via analyses of system logs or through qualitative data, rather than being a primary focus of the evaluation. This creates a problem, however. Without a clear understanding about which intervention component worked or didn't work, it is difficult to make evidence-based decisions about what to change in the next version of the system to make it more effective or which design ideas found in an effective system to appropriate when designing a new mHealth intervention. Similarly, if a new system adopts an intervention idea from an effective system and the new system turns out not to be effective itself, it is hard to tell whether the borrowed intervention idea was bad, whether it was designed in an ineffective way, or if that intervention just doesn't work for the behaviors or population that the new system was trying to support. In other words, traditional mHealth evaluations generate evidence that is of limited usefulness for informing future work, either one's own or that of others.

In the remainder of this article, we describe an alternative strategy for evaluating mHealth systems for health behavior change, one that is rooted in the research on intervention optimization in behavioral science [1] and our own work on Agile Science [2,4]. Specifically, we advocate for assessing individual intervention components as the primary focus of mHealth evaluations, using proximal outcomes as evaluation metrics, and employing experimental designs that can efficiently assess causal effects of distinct intervention components and their interactions. This evaluation strategy, we suggest, results in evidence that can more directly inform decisions that mobile computing researchers and designers need to make in their work: which features to discard and which to keep, what needs to be redesigned, what components to incorporate in a new system, and how to design them.

Focus on individual intervention components

Modern mHealth systems are highly complex. Consider mConnect [3] (see Figure 1), a research mHealth app for encouraging walking. mConnect includes seven components: passive activity tracking, graphs of user’s activity, problem solving, reinforcement, a message board, social norms feedback, and an ambient display. Many research mHealth systems have just as many components, and commercial systems such as Fitbit are still more complex, consisting of dozens of intervention components. But complexity does not stop with the number of intervention components. A single component, such as a message board, can support multiple potential mechanisms of change (social affirmation, accountability, emotional support, and instrumental support, among others), and multiple intervention components (e.g., ambient display and activity graphs) can support a single mechanism (self-monitoring feedback).

Understanding how exactly an mHealth system such as mConnect influences its target behaviors—and, therefore, what needs to be changed to make the system more effective, as well as what can be learned from it to inform the design of future mHealth systems—requires researchers to try to unpack this complexity. An efficient way of doing this is via studies that focus on assessing effects of individual intervention components contained in an mHealth system.

Focusing evaluations on individual intervention components has a number of benefits: (1) It enables researchers and designers to identify which intervention components are contributing to the system’s effectiveness and which are not, informing decisions about components that may be eliminated to simplify the system and make it less burdensome; (2) It enables studying how different components interact: whether the effect of a component is stronger when another one is present or being used, or if the effect of a component is weakened by the presence of another; (3) It allows studying how design influences a component’s effectiveness. Researchers can create two versions of a component and directly assess whether one design works better than the other; and (4) it enables researches to evaluate not only whether included intervention components are influencing the target health behavior but also whether they are functioning through the mechanisms they are intended to activate. For instance, a study can assess not only whether inclusion of a social norm intervention component increases walking but also whether that component is, in fact, changing social norms related to health and exercise. In this way, studies focused on individual components can deepen our understanding not only of what works but also of why and how it works.

Proximal outcomes

Behavior change is a long-term process, and convincingly showing that a health behavior was adopted and is being maintained can take many months or even years. Yet, evaluating mHealth systems in the way we have been describing can often be done much more quickly —within weeks or a couple of months. What makes this possible is the concept of proximal outcomes [4]. Proximal outcomes refer to the intended short-term effects of a single provision of an intervention component or another meaningful minimal dose (e.g., a week of motivational text messages). For a medication reminder, a proximal outcome might be whether the user took her medicine within some time window of receiving the reminder. For daily step goals, a proximal outcome might be whether the user met her goal by the end of the day or came within some threshold of the goal. And for rewards for goal attainment (e.g., the fireworks displayed by a Fitbit when the user meets her daily step goal), a proximal outcome might be whether the user meets the goal on the following day. Proximal outcomes, in other words, capture the small effects through which a particular intervention component is intended to influence the macro behaviors the mHealth system is trying to support (routinely walking 10,000 steps a day, losing 20 pounds, taking hypertension medication every day, etc.). A practical way of evaluating individual intervention components is by assessing whether they are effectively influencing their proximal outcomes.

It is useful to distinguish two types of proximal outcomes: behavioral and mechanistic proximal outcomes. Behavioral proximal outcomes assess the specific behaviors that a particular intervention component is intended to encourage. Some intervention components work by encouraging small-scale versions of the health behaviors that the mHealth system as a whole is trying to support —a single bout of walking, a single day of sticking to one’s calorie goal, or a single act of taking medications. Such intervention components support behavior change through accumulation of target behavior over time, and their proximal behavioral outcomes just are individual bouts of target behavior those components are trying to influence (e.g., a single act of medication taking). Other intervention components encourage behaviors that are known to support the behavior change process but are not the target health behaviors themselves. For example, a component of an mHealth app for helping individuals with problem drinking might encourage the user to reach out to family and friends who support her sobriety. For that component, its behavioral proximal outcome might be the number of sobriety-supporting interaction that the user has each day. Note that for both types of behavioral proximal outcomes, the outcome is actual observable behavior.

Mechanistic proximal outcomes intend to capture the psychosocial or physiological processes that are thought to mediate health behavior change. These are things like self-efficacy, stress, outcome expectancies and other such psychological or physiological constructs. In our example of the problem drinking app, the component that encourages interactions with sobriety-supporting friends may do so in order to enhance the user’s feelings of being supported. To assess whether this component actually has this effect, researchers might assess, usually via questionnaires, participants' perception of social support. This would be done in addition to assessing their social interactions, as the two outcomes might not covary closely and researchers might be interested in learning which of the two is more predictive of abstinence. Similarly, rewards for attaining daily step goals might be postulated to work both as positive reinforcement—increasing the frequency of the rewarded behavior in the future—and as a way of increasing users' self-efficacy. To test if the rewards are working in this way, researchers might assess participants' self-efficacy for physical activity each day, as well as their walking behavior.

Assessing causal effects

While proximal outcomes enable capturing of behaviors and processes that components of a system are intended to support, to robustly establish that those outcomes are actually being changed by those components and not by some other process requires the use of study designs that can experimentally control the delivery of the intervention components under investigation. Luckily, over the last fifteen years, a range of such experimental designs have been developed in behavioral science, and they are as useful for evaluating mHealth technologies as they are for other types of behavioral treatments. A class of experimental designs that are particularly well suited for mHealth research are factorial experiments [1]. Factorial experiments work by randomizing individuals to versions of an mHealth system that contain different subsets of its intervention components. Researchers can then test the efficacy of individual components by assessing proximal and distal outcomes for participants who received a particular component vs. those who did not receive that component. What makes factorial designs particularly efficient is that multiple intervention components can be evaluated in the same study without increasing the sample size over what would be needed to evaluate a single component (albeit one with the smallest expected effect size).

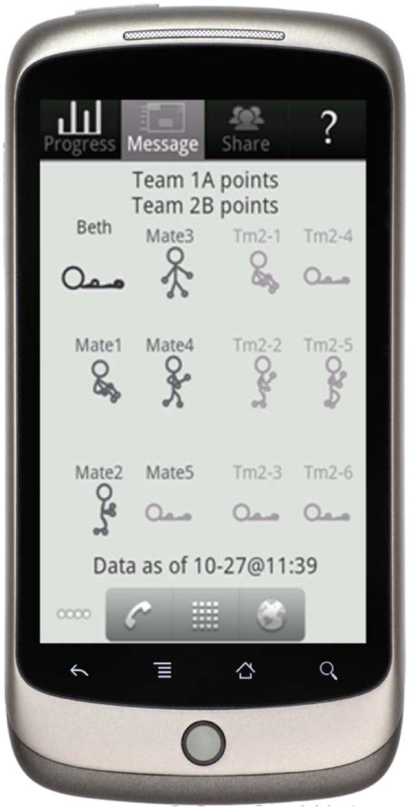
Another efficient experimental design for evaluating mHealth systems are micro-randomized trials [5]. These trials are intended to evaluate intervention components that are "pushed" to users via push notifications, text messages, vibrating alerts, and other methods of gaining users' attention. Reminders, motivational messages, and daily goal assignments are just some examples of push components in mHealth systems. In a micro-randomized trial, each time such a component can be delivered to the user (e.g., each morning for a daily step goal), the system randomizes whether to deliver it or not. Thus, over the course of the study, a push component is randomized many times for each participant. The effect of the component is evaluated by assessing the difference in the component's proximal outcome between the times when participants were randomized to receive the component and the times when they were randomized to not receive it. As the randomization happens repeatedly, analyses of the data can also assess how a component's effects change over time, as well as how those effects are moderated by context (e.g., location) in which the component is delivered. As with factorial experiments, micro-randomized trials can evaluate multiple components in the same study, making them a highly efficient method for testing push interventions.

Summary

We argue that to better understand how their systems are working, mHealth researchers need to focus on assessing effects of individual components of the technologies they are developing. This task is greatly facilitated by formulating proximal outcomes and using study designs such as factorial experiments that were developed to robustly assess causal effects of components of complex interventions. In the context of mHealth, this strategy can generate knowledge that is readily usable for informing decisions that mHealth researchers and designers have to make: what aspects of the system to change, which components to keep or abandon, and how to use findings from evaluations of one system in the development of the next one.

References

1. Collins Linda M, Trail Jessica B, Kugler Kari C, Baker Timothy B, Piper Megan E and Mermelstein Robin J.. 2014 Evaluating individual intervention components: making decisions based on the results of a factorial screening experiment. Translational behavioral medicine 4, 3: 238–251. [PubMed: 25264464]
2. Hekler Eric B, Klasnja Predrag, Riley William T, Buman Matthew P, Huberty Jennifer, Rivera Daniel E and Martin Cesar A. 2016 Agile science: creating useful products for behavior change in the real world. Transl Behav Med 6, 2: 317–28. [PubMed: 27357001]
3. King Abby C, Hekler Eric B, Grieco Lauren A, Winter Sandra J, Sheats Jylana L, Buman Matthew P, Banerjee Banny, Robinson Thomas N and Cirimele Jesse. 2013 Harnessing Different Motivational Frames via Mobile Phones to Promote Daily Physical Activity and Reduce Sedentary Behavior in Aging Adults. PLoS ONE 8, 4: e62613. [PubMed: 23638127]
4. Klasnja Predrag, Hekler Eric B, Korinek Elizabeth V, Harlow John and Mishra Sonali R. 2017 Toward Usable Evidence: Optimizing Knowledge Accumulation in HCI Research on Health Behavior Change. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 3071–3082.
5. Klasnja Predrag, Hekler Eric B, Shiffman Saul, Boruvka Audrey, Almirall Daniel, Tewari Ambuj and Murphy Susan A. 2015 Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychol 34 Suppl: 1220–8.



Team 1A points
Team 2B points

Data as of 10-27@11:39

Figure 1:
The mConnect app