Toward Usable Evidence: Optimizing Knowledge Accumulation in HCI Research on Health Behavior Change

Predrag Klasnja[1], Eric B. Hekler[2], Elizabeth V. Korinek[2], John Harlow[2], Sonali R. Mishra[3]

[1] Group Health Research Insttute
Seattle, WA, USA
klasnja.p@ghc.org

[2] Arizona State University
Phoenix, AZ, USA
{ehekler, evkorine, jharlow}@asu.edu

[3] University of Washington
Seattle, WA, USA
smishra@uw.edu

ABSTRACT

Over the last ten years, HCI researchers have introduced a range of novel ways to support health behavior change, from glanceable displays to sophisticated game dynamics. Yet, this research has not had as much impact as its originality warrants. A key reason for this is that common forms of evaluation used in HCI make it difficult to effectively accumulate—and use—knowledge across research projects. This paper proposes a strategy for HCI research on behavior change that retains the field's focus on novel technical contributions while enabling accumulation of evidence that can increase impact of individual research projects both in HCI and the broader behavior-change science. The core of this strategy is an emphasis on the discovery of causal effects of individual components of behavior-change technologies and the precise ways in which those effects vary with individual differences, design choices, and contexts in which those technologies are used.

Author Keywords

Evaluation methods; behavior change; health informatics; user studies.

ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User interfaces (Theory and methods); J.3. Computer applications: Life and medical sciences (Health).

INTRODUCTION

Imagine that you are designing a mobile application for stress management. During literature review, you came across UbiFit [9,10], MILES [24], UbiGreen [15], and BeWell [29,32], and inspired by the idea of a glanceable display you decide to incorporate it into your application. But you face a problem: you are developing for iOS, which gives you neither the control of the lock screen nor a mostly empty home screen on which to put your display. You have some ideas—you can create the display as a widget on the iPhone's Today screen, for instance—but before you put in the time and resources to build and test the display, you'd like to use prior research to gauge if it's worth it to use a glanceable display for your target users.

Specifically, you'd like to know the following: (1) How well did the overall UbiFit, MILES, UbiGreen, and BeWell systems do at influencing physical activity, eco-friendly behavior, and social wellbeing? (2) How impactful was the glanceable display component in particular for producing these outcomes? (3) What design differences between the glanceable displays of UbiFit, MILES, UbiGreen, and BeWell were meaningful, and would those design differences matter for your target user? (4) What are the essential features that make an interface a useful and usable glanceable display? For example, does a glanceable display require a passive sensor for the target variable of interest, such as stress, to work effectively? What other design options fit with the concept of a glanceable display (e.g., would a Today widget on iOS count)? (5) What are the characteristics of the users and the contexts of use in UbiFit, MILES, UbiGreen, and BeWell that may have influenced the value (or lack thereof) of their glanceable displays? (6) How likely is it that the glanceable display concept is useful for a target like stress? For example, will frequent reminders of stress end up being more aversive compared to physical activity and thus undermine the usefulness of the glanceable display? (7) How might the design of a glanceable display need to be adjusted for it to be useful for stress management? And, (8) can those adjustments feasibly be implemented on iOS (e.g., the Today screen would be seen a few times a day, but not as often as the lock screen or the home screen. Is this likely to be enough, or should you explore other options, such as an expensive smartwatch)? You reread the papers and gain some insights on the first question, but have only limited guidance related to the other questions, beyond the plausible design suggestions made by the authors with varying degrees of confidence.

The above scenario highlights an important issue related to accumulation and translation of evidence in HCI research on behavior change. While HCI researchers use sophisticated design methods to create innovative technologies for supporting behavior change, how those technologies are evaluated limits the ability to accumulate knowledge needed to make evidence-based decisions during the development of new interventions. In particular, details are missing to understand, with precision, how and why our technolo-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI 2017, *May 06 - 11, 2017, Denver, CO, USA*
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3026013

gies influence human behavior. One core issue is that commonly used study methods, such as simple pre-post designs and between-person randomized controlled trials, can usually not answer the types of subtle questions (like those above) that could sufficiently inform the design of new systems. Another issue is that our empirical findings and design guidelines are rarely specified at a granular enough level to support effective accumulation of evidence across projects about what kinds of interventions work for whom and under what circumstances. And finally, those findings often do not withstand plausible challenges to causal inference and generalizability that come up when researchers from other disciplines try to take up our work or when HCI grant proposals are evaluated for funding.

The purpose of this paper is to articulate the concept of "usable evidence" that aims to address those limitations of the current HCI evidence base and to propose a research process that can be used to generate such evidence. Key contributions of this paper include: 1) delineation of the concept of usable evidence as a desirable target for HCI research that can facilitate more robust knowledge accumulation in HCI and knowledge transfer to other disciplines; 2) specification of a research process for generating usable evidence that fits with the incentives and constraints of HCI research and that melds lessons from HCI with the new evaluation methodologies being developed in behavioral science; (3) the description of an on-going case-study of this approach to ground the discussion in a pragmatic use-case; and (4) an epistemological discussion about the purposes of scientific evidence in HCI research and the ways that our concept of usable evidence addresses those needs.

RELATED WORK

In its aims and approach, the current paper builds on a growing literature in behavioral science and HCI that calls for better evaluations of behavior-change interventions. In HCI, Klasnja et al. [25] and Hekler et al. [18] have argued for the value of interview data for understanding whether behavior-change technologies (BCTs) are working as intended and how user experience and context of use affect the ability of a technology to effectively support the behavior-change process. More recently, Kay et al. [22] have argued for the use of Bayesian statistics as a way to improve the level of quantitative evidence obtained from deployments of technologies, as many field studies that HCI researchers conduct are underpowered to provide high-quality evidence using frequentist methods. The current paper adds to these suggestions by arguing that the use of proximal outcomes in evaluations of BCTs enables HCI researchers to conduct rigorous efficacy evaluations of their systems within resource constraints of HCI field studies.

In behavioral science, the call for better evidence is rooted in the realization that randomized controlled trials provide evidence of limited utility for translating experimental findings into practice [23] and for informing design of more effective interventions [38], especially digital interventions that leverage mobile technology [28,37]. A key response to this challenge has come from "optimization" methods inspired by Linda Collins's multiphase optimization strategy (MOST) [5,6]. MOST emphasizes evaluations that aim to understand functioning of individual intervention components within a complex intervention, in order to enable creation of streamlined interventions that only include components that have been shown effective. While MOST still sees a randomized controlled trial (RCT) of an optimized intervention as the final step in this evaluation pipeline, its biggest impact in behavioral science has been to affirm the notion that researchers should be focusing on opening the "black box" of intervention effectiveness to better understand how interventions work for particular individuals in context. Our paper embraces the emphasis on the evaluation of individual intervention components, and we argue that this emphasis fits well both with the research questions that HCI researchers working on behavior change care about and the resource constraints of HCI research.

MOST studies can be conducted using a range of experimental designs, including factorial and fractional factorial experiments [4,7], SMART trials [30,35], micro-randomized trials [26,31], system ID experiments [17], and single-case experimental designs [11,12,27]. Many of these designs are useful for testing BCTs, and, as we argue later in the paper, some of them (factorial designs, single-case experiments) are very well suited for HCI BCT evaluations.

Another response to the call for better evidence has been the emphasis on understanding the dynamics of behavior change in order to develop more precise behavioral theories [37,39]. Insofar as BCTs embody theoretical constructs hypothesized to activate specific mechanisms of change, the types of studies we are proposing in this paper advance this goal very directly. In fact, our emphasis on articulation of causal pathways is intended to help researchers generate evidence that can be directly used for theory development.

Finally, our paper builds directly on our work on Agile Science [19], which synthesizes a range of these ideas into a framework for effective knowledge generation and sharing. As we argue below, agile science can be of great utility to HCI researchers by providing a process for clearly articulating how technologies that HCI researchers are developing are supposed to function, enabling rigorous evaluation of the ideas underlying these technologies.

TERMINOLOGY

In this section, we introduce key terms that we will be using in this paper, in order to create a shared vocabulary for the discussion of different types of evidence and their uses.

Usable evidence

We define *usable evidence* as empirical findings about the causal effects of BCTs and how those effects vary with individual differences, context of use, and system design. Such findings are "usable" in that they enable designers, researchers, and even end-users to make evidence-based

decisions about what design elements to include in new systems for specific user groups and how exactly individual features of a technology should work to maximize its potential to be effective for those users. The applicability of usable evidence is achieved via an explicit focus on discovering how causal effects of individual components of a technology vary with time, in different contexts, and for different people—differences that are usually washed out in studies that aim for "on average" evidence across groups, such as classical randomized controlled trials. As such, usable evidence can be seen as a BCT equivalent of the evidence sought in precision medicine [2] in order to develop treatments finely tailored to individual genotypes.

Intervention idea

An *intervention idea* is an account of how a technology can bring about behavior change. It describes something a technology can do to help people change behavior. Intervention ideas can come at various levels of specificity. They can be based on hunches—rough ideas about what might work to produce desired outcomes—that come from formative work with target users or anecdotal experience (e.g., helping people remember they feel good when they exercise might make them want to exercise more), or they can be formalized hypotheses with a great deal of empirical support (e.g., self-monitoring daily step count can help people walk more). The process we describe later in the paper is intended, in part, to help researchers better specify their intervention ideas and develop them from hunches to hypotheses that can be formally tested in empirical studies.

Intervention components

By *intervention component* we mean a concrete piece of functionality of a technology that enacts a specific intervention idea for supporting behavior change. Many intervention components implement well-supported intervention ideas (e.g., goal-setting, planning, feedback on progress), such as the behavior-change techniques taxonomized by Susan Michie and colleagues [1,33] through an expert-consensus-based review of components used in complex behavior-change interventions. Other components, like some developed by HCI researchers, implement novel intervention ideas. Either way, as concrete pieces of code, intervention components include the many design decisions that have to be made to translate abstract ideas into a working system. For instance, a goal-setting component of a physical-activity intervention incorporates decisions about units used to set goals (e.g., step count vs. minutes of activity), time frame (daily vs. weekly), a choice of default goals, and various design details of the user interaction of setting a goal (where in the application goal-setting is done, how it is presented, etc.). It is such intervention components, as implemented, that are directly evaluated in individual studies of BCTs.

Distal and proximal outcomes

Every BCT is intended to help individuals attain some long-term goal: lose weight, increase physical activity, manage stress, improve medication adherence, or improve outcomes such as reduced cardiovascular disease, diabetes, or cancer risk. Such outcomes are *distal outcomes*, and they are the kinds of things that are evaluated in RCTs to determine if an intervention is effective. Yet, insofar as they happen, such changes in behavior come about gradually and over time, and they are often due both to the effects of the intervention and other factors, such as, for instance, exogenous shocks (e.g., seeing a friend have a heart attack).

An intervention contributes to the behavior-change process via specific intervention components, each of which is designed to support behavior change through a particular mechanism of action. The direct outcomes of the provision of an intervention component—effects that link the component delivery to the intervention's distal outcomes—are the component's *proximal outcomes*. They represent elements in the causal pathway that the intervention uses to facilitate a desired change in the distal outcomes.

Sometimes, proximal outcomes are just small-scale versions of the intervention's distal outcomes. For example, a proximal outcome for a planning component in a physical activity intervention that helps users plan when and where they will be active the next day might be a single bout of physical activity—a micro component of a habit of being active. Other times, proximal outcomes are different than distal outcomes. For instance, there is a fair bit of literature showing that social support can help individuals struggling with substance use remain sober [16]. An intervention targeting substance use might have a component that encourages communication with sobriety-supporting family and friends. A proximal outcome for such a component might be the number of daily interactions with sobriety-supporting social network.

What's important to note is that the proximal outcome is a desired outcome of a single provision of an intervention component—of a single session of planning, single delivery of a suggestion to be active, single reminder, and so on. Proximal outcomes are the most direct changes that an intervention component is intended to create, changes which, over time, should create the more global changes in behavior and health as defined by distal outcomes. As such, proximal outcomes are key to understanding if an intervention is working as intended because they allow researchers to test individual branches of the causal pathway through which the intervention is hypothesized to change behavior and distal health outcomes. We will return to this point.

Finally, although the distinction is not clear-cut, there are two types of proximal outcomes: behavioral and mechanistic. *Behavioral proximal outcomes* are the types of proximal outcomes we have been discussing—bouts of activity, reaching out to a friend, engaging in craving-surfing, etc.—concrete behaviors that intervention components encourage users to do, which, in turn, move them toward the desired outcome(s). *Mechanistic proximal outcomes* refer to the physiological and psychological changes that intervention components aim to induce, which mediate the behavior-

change process. For instance, a reminder to engage in a breathing exercise might be intended to reduce stress, which, in turn, might help a smoker resist lighting a cigarette. A stress sensor, such as cStress [21], can allow researchers to check if sending the reminder affects the user's stress level and, thus, to test the assumption that stress reduction mediates smoking cessation.

Causal pathway

A causal pathway refers to a diagram of the sequence of changes—causes and effects—that are hypothesized to link provision of an intervention to its desired distal outcomes. More simply, a causal pathway is a way to visualize how the intervention is intended to work. In behavioral science, causal pathways are often represented as directed graphs—boxes and arrows showing what is changing what. They represent various effects that each intervention component is hypothesized to have and how those effects influence other downstream outcomes, and, ultimately, desired distal outcomes. Drawing a causal pathway forces one to clearly articulate effects of and relationships among different parts of an intervention, making it possible to check whether the intervention, as conceived, has a hope of working. Causal pathways also enable researchers to design studies that can assess postulated effects and relationships, improving the evidence that results from system evaluations.

Causal pathway diagrams can—and should—include both effects on behaviors and the psychosocial and physiological mechanisms postulated to mediate them, as well as the effects that different parts of the interventions might have on user engagement with the intervention. In addition, causal pathways can help specify factors such as context (e.g., weather), traits (e.g., novelty-seeker), and states of the individual (e.g., being stressed) that might moderate the influence of the intervention on proximal and distal outcomes. Thus, causal pathways are key for developing hypotheses that can be tested in system deployments to determine to what extent the effectiveness of intervention components and systems generalizes [20].

FROM A HUNCH TO A TESTABLE INTERVENTION

As we alluded above, the kind of evidence we are advocating is based on careful experimental tests of intervention components that embody clearly articulated intervention ideas and that have clearly articulated outcomes. But we often start with intervention ideas that are far vaguer, ideas that are based on designer intuitions, hints from formative work with target users, or ideas that emerged from reading the literature. How do we go from such hunches about what we think might work to intervention ideas that are fleshed out enough to be implemented and tested with real users?

To answer this question, we draw on our work on Agile Science [19], a set of methodologies and principles intended to improve knowledge generation and sharing in behavior-change sciences. In particular, a part of the process is called the *Generate Phase*, which is a set of iterative activities that researchers can do to develop their early intervention ideas into testable intervention components with well-specified mechanisms and outcomes. In this section, we provide an outline of the part of that process that targets specification of intervention ideas and outcomes (see [19] for other activities in the generate phase, including user-centered design and simulations studies). In the next section, we illustrate this process with a case study of the development of a walking intervention.

Hunch articulation

The generate phase prototyping work begins with a "hunch," articulated by completing the phrase: "Based on theory and/or understanding potential users, I'm not sure, but I think that (fill in intervention idea) will influence (fill in behavioral outcome of interest)." "I'm not sure" is an important modifier on confidence. If one is sure that the intervention idea will work (or that it won't), then there is no need to test it. Notions about the intervention idea will often come from preparatory work reviewing theory and empirical evidence in the scientific literature, or engaging with users through observation or interviews. With preparatory work complete, researchers can begin to link the intervention idea to behavioral and mechanistic outcomes that the intervention idea is hypothesized to affect.

Niche specification

The initial hunch articulation will usually involve a variety of unstated "niche" specifics. To explore the niche of an intervention idea, researchers should clarify for whom they are designing (e.g. smartphone users, heart attack survivors, Phoenix residents, etc.). Similarly important is whom the intervention idea is not targeting (e.g. highly physically active people, people in a specific climate, people whose culture is mismatched with the intervention, etc.). Beyond specifying the expected users of an intervention, the setting—both in terms of the anticipated context of use and planned temporal patterns of intervention delivery—can help to make an intervention's niche explicit. Two important reasons for specifying the niche in this way is (1) to begin generating information about individual- and contextual-level variables that might affect intervention use and effectiveness, and (2) to generate a list of feasibility checks for potential implementations/prototypes of the intervention idea that is being developed.

Causal pathway mapping

After expressing a hunch and defining its niche, the specific relationship between an intervention idea and its behavioral and mechanistic outcomes can be explored through causal pathway mapping. As we discussed, a causal pathway is a visual representation of the cause and effect chain(s) that link an intervention with its outcomes. This exercise helps clarify exactly *how* an intervention idea is presumed to impact its proximal outcome(s) and, through them, its intended distal outcome(s). Causal pathway mapping begins by listing various outcomes that the intervention idea might affect, as well as drawing on prior literature and niche specification, factors that might influence the impact of the intervention of these outcomes. Outcomes and moderators are

then arranged in a directed graph that describes how they relate to one another. For instance, the graph can show that sending someone a suggestion to take a walk has steps over the next 30 minutes as a proximal behavioral outcome, but that this outcome might be dependent on the person's busyness level and weather. The graph could also show that user annoyance (from being pinged) is another potential, albeit undesired, outcome of the walking suggestions. At this stage, researchers should try to be as comprehensive in specifying the causal pathway as their knowledge allows. This allows the causal pathway to become a visual representation of the beliefs about exactly how the intervention is supposed to work (or might not work), which enables generation of hypotheses about the intervention's operation that can be formally tested in empirical studies.

Hunch revision

Mapping the causal pathway will likely evolve the intervention idea, which should be updated in the hunch statement. This is also an opportunity to decide on the primary proximal outcome. When the hunch is first written down, it is usually in terms of a distal (or vague) behavioral outcome. Causal pathway mapping helps identify relevant proximal outcomes that would be more directly influenced by the intervention idea. The revised hunch should reflect these updates in the understanding of the intervention idea and the main outcomes that the idea is intended to impact.

Variation generation

Once the hunch is reasonably well specified, researchers proceed to variation generation—the activity of generating various ways that the intervention idea can affect its proximal outcome. We have found that this works well in the form of group brainstorming, where team members rapidly write on post-it notes, individually first and then collectively, all ideas that come to mind for different forms that the intervention idea can take. Once these ideas are generated, they need to be categorized. Many of the ideas will have the same theme or mechanism, and they can be grouped together. Our preliminary experience with this method has generally produced three to five distinct categories of ideas. The groupings are contender intervention-idea variations, and researchers can then decide to focus on just one of them (thus narrowing how they understand their intervention idea), or they can decide to compare different variations—maybe because they are interested in their relative effectiveness. Either way, the ideas generated in this activity begin to concretize the intervention idea by articulating it in forms that are potentially implementable.

Prototype specification

Building on prior work on parallel development and prototyping [14], there is great value in testing multiple variations of an intervention idea, particularly in an early stage of the process when the goal is to discover which intervention ideas might be potentially useful. To test these variations, they need to be implemented. By a *prototype* we refer to an implementation of an intervention idea that can be tested with real users. Typically, a variation of an intervention idea can be implemented in a number of ways, and resource constraints, feasibility, and interest in particular design aspects will guide how a researcher decides to implement the ideas he/she wants to test. On one end of the spectrum, a prototype can be minimal—a bare-bones implementation that is purely designed to assess whether there is any value at all to the intervention idea that the researcher is considering. For new intervention ideas for which there is little prior literature on which a researcher can draw, testing of such bare-bones prototypes can provide an efficient way to assess whether it is worth investing any more time and resources on the idea (see [13] for advertising probes that are at this level). On the other end of the spectrum, an idea can be implemented as a refined component within a complex system. In either case, once prototypes are created, they enable the researcher to test whether, for whom, and in what contexts these implementations of the idea that the researcher is investigating have their intended effects. As we will shortly see, these kinds of tests are at the core of the concept of usable evidence.

Specification of the experimental design

The final step of the process is a clear delineation of the experimental design that can be used to efficiently test aspects of the causal pathway via the prototypes. There are several options available for this, which we discuss in the Towards Usable Evidence section. For more information on various experimental designs, see [8].

STEP GOAL VARIABILITY: A CASE STUDY OF THE GENERATE-PHASE PROTOTPYING PROCESS

The process we described in the previous section is likely to be too abstract on first reading. To make the process easier to understand, in this section we provide a case study of how the Generate Phase prototyping process was used to develop an intervention idea of using goal variability to help individuals increase their physical activity.

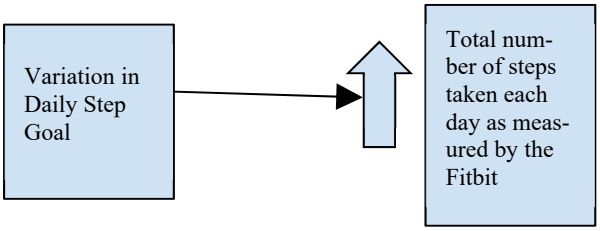
The idea behind this case study emerged in the course of testing of an Android mHealth application, which encouraged regular walking via an adaptive daily step goal and financial incentives for meeting this goal. During exit interviews in the study, participants indicated that they enjoyed that their step goals changed each day, raising the possibility that introduction of variation might be a meaningful way to increase walking behavior. We developed this hunch of an intervention idea using the process described in the last section. The process proceeded as follows:

Initial hunch: "Based on theory and understanding potential users, I'm not sure, but I think that step-goal variability (intervention idea) will help people walk more (behavioral outcome of interest)."

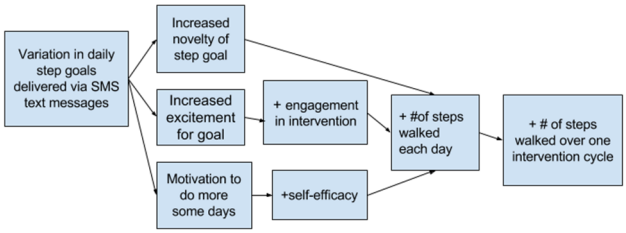
Niche specification: The research team was specifically targeting higher SES, overweight, sedentary, middle-aged adults who used smartphones. The team was not designing for adults with high current activity levels, children, or those with a health condition that might interfere with their

ability to walk for exercise. Recruitment was planned to occur through a medium-sized consulting firm and would include employees working remotely. Employees were described as holding a mostly sedentary job, with some employees having access to a sit-to-stand desk.

Causal pathway: The direct relationship between step goal variability and walking behavior was specified visually as:



This pathway was hypothesized to be mediated by the following factors:



Proximal outcomes: From this model the proximal outcomes were identified as: (1) an increase in the number of steps walked each day, as measured by Fitbit; (2) satisfaction with and enjoyment in the intervention component, as measured by a weekly Qualtrics survey; and (3) an increase in the total number of steps walked over one intervention cycle (one week).

Variations: In this exercise the research team focused on different ways they could vary a step goal to help the population of interest walk more and achieve the target proximal outcomes identified above. Initial brainstorming yielded over 20 different ideas for varying a daily step goal. For example, "graded increasing goals," "surprise days of rest," "one day per week of an extreme goal," "contextually tailored goal," etc. From this list, the research team organized the ideas into three contender groups that represented distinct ways to vary a daily step goal: (1) epic goals; (2) user choice goal; and (3) high amplitude goals.

Prototypes: Due to resource constraints, the research team decided to implement the idea of goal variability via text messages sent to participants each day of the study. Furthermore, the team decided to test all three variations on the variable-goal idea, and it further defined how each variation would function. Intervention variations were defined as follows: *Epic goals*: let user plan for an "epic" goal—defined as more steps than they have ever walked on a single day in the previous month. *User choice goal*: let user choose their own step goal on one weekday and one weekend day. *High amplitude goals:* user receives some atypically high goals and some atypically low goals.

At the end of this process, the research team ended up with prototypes of three versions of the variable-goal idea that could be tested in the field to (1) validate the idea that goal variability could be useful for encouraging walking; (2) compare whether one version/design of this idea works better than others; and (3) assess how this intervention idea impacts different types of outcomes both on average and over time. From this, we developed an experiment (a variant of a single-case design called a Latin Square), which allowed us to study the concept of goal variability. As we'll see next, studies that can answer such questions are central to the concept of usable evidence.

WHAT IS EVIDENCE FOR?

Research is, ultimately, a form of communication. We build our work on the work of others and hope that others, in turn, will benefit from what we have done. As such, to maximize the impact of our work, the evidence we should aim to create through our research should be in the form that makes it maximally useful to others, both in our own field and in related disciplines that would be interested in building on or using our work. In other words, we should aim to create *usable evidence*, evidence that most readily helps other researchers pursue their research goals. In this section, we outline the chief forms that these goals take, and their implications for the forms of evidence we should be striving to generate to support those goals.

Adapting intervention ideas from prior research

An important form of work is adapting intervention ideas previously deployed for new interventions. The scenario that we started this paper with, on using a glanceable display for stress management, is an example of this research goal. In HCI research, adoption of intervention ideas from prior work usually involves introducing some form of novelty (design, technical underpinnings, etc.), but the question a researcher is asking is more general: can I use intervention idea A, potentially with modifications, for context X, where context X might differ from the context in which A was originally used in terms of the behavior, population, circumstances of use, and/or technological constraints?

To answer this question with reasonable confidence, a researcher needs to know several pieces of information: (1) What mechanism of change does A embody, and is that mechanism applicable for the context in which the new intervention would be used? (2) Did the intervention component that implemented intervention idea A have an effect, and if so, for what outcomes did it work? (3) If the intervention component produced an effect, what were the dynamics of the effect, and are those dynamics acceptable for the new context for the intervention? (for instance, reminders that work for a couple of weeks but then slowly stop working might be perfectly fine for increasing adherence to antibiotics, but would not be suitable for adherence to birth

control pills); (4) Did idea A work particularly well/poorly for particular types of people or in particular contexts, and how does that translate to the population and circumstances where the new intervention needs to be used? And (5) were there evaluations of different versions or designs of intervention idea A? If so, what does the evidence suggest about how the new intervention may need to be designed to maximize its effectiveness? For instance, do reminders presented as coming from the clinic work differently than reminders that appear to come from a consumer application?

Developing new intervention ideas

Although adoption of intervention ideas from prior work is a common research strategy in behavior-change science, within HCI an arguably even more important research task is creation of new intervention ideas—invention of new ways in which technology can support behavior change. As we saw, new intervention ideas often start as hunches, obtained from formative work or anecdotal evidence, that something might work. Such new intervention ideas do not build on prior literature as directly as adaptations of existing intervention ideas do, but prior evidence is still useful. Most new intervention ideas implement some known mechanism of change, which has been previously tested. Understanding this evidence can help to better define—and design—the new intervention idea. A researcher wants to understand the kinds of effects this mechanism of change is likely to produce, their temporal dynamics, and differences in effects previously found for interventions that implemented this mechanism differently. Have there been implementations that have worked particularly well or particularly poorly? If so, what was distinctive about those implementations, and what does that mean about the intervention idea that the researcher is developing? The clearer the evidence from prior work related to these questions, the easier it is for the researcher to develop the new idea from the start in a way that increases the likelihood of success.

Optimizing interventions for efficacy evaluations

Although HCI researchers usually focus on early-stage evaluations, for intervention ideas to have maximum impact (such as uptake in healthcare), they need to be incorporated into systems that can be tested for efficacy for distal health outcomes, and, if found efficacious, made available more broadly. As Collins et al. have argued [3,6], the time and cost of randomized controlled trials used to robustly establish intervention efficacy only make sense for interventions that have been optimized to only include components that have shown preliminary evidence of efficacy in early-stage studies. To make decisions about which components to include and how to combine them in a complex intervention, researchers need good prior evidence about how each component they are considering works. Results from studies that assessed various versions of the components they are considering enable intervention scientists to incorporate in an intervention they plan to deploy in a large trial the most promising versions of only those intervention components for which they have strong preliminary evidence.

Finding limits of generalization for change mechanisms

For behavioral scientists, questions of generalizability are often just as important as development of effective interventions. They are interested in understanding the change process at a more general level, which means that they are interested in understanding mechanisms that facilitate behavior change and the conditions in which those mechanisms work. But mechanisms cannot be studied in the abstract. They can only be studied via interventions that embody them. As such, robust evidence about the effects of individual intervention components with clearly articulated underlying mechanisms is paramount for furthering our understanding of the process of behavior change and its drivers; and the more granular the evidence, the better. Knowing that a planning component of a physical-activity intervention had an effect on increasing activity is good, but knowing how this effect changed over time, whether the effect was greater for certain types of people (e.g., for people high in conscientiousness), and whether the effect was limited to certain contexts (e.g., when planning was done when the participants were not tired) is better. When such evidence is available for different interventions that implement the same hypothesized mechanism of change, researchers can systematically synthesize the literature to uncover conditions under which each mechanism operates [20]. Such knowledge not only increases our understanding of how the behavior change process works, but also provides a guide for the development of new interventions, as components that leverage a particular mechanism can be designed to be provided only to people and in contexts in which they are most likely to be effective.

TOWARD USABLE EVIDENCE

The research goals we outlined in the last section have much in common in terms of the nature of the evidence that these research activities require. This evidence shares the following characteristics: (1) it focuses on individual intervention components rather than whole systems; (2) it focuses on showing component-specific effects—namely, effects that are directly linked to the provision of each intervention component; and (3) it requires granularity in terms of temporal dynamics and effect moderation. All three of these requirements, we want to suggest, can be achieved by focusing on assessment of proximal outcomes.

In this section, we describe how the focus on proximal outcomes can allow HCI researchers to create robust evidence about causal effects of technologies they are developing, evidence that is well suited to support the full range of research goals outlined in the last section. Just as importantly, we suggest that HCI researchers can do this within the time and resource constraints of typical HCI projects, without the need to recruit hundreds or thousands of individuals or run studies that last many months or years.

Proximal outcomes as chief evaluation targets

As we described in the section on terminology, proximal outcomes are direct behavioral or mechanistic outcomes that an intervention component is designed to have. Im-

portantly, they are outcomes that, if present, could be observable for a single intervention provision, such as a single session of planning or an interaction with a support group.[1] Proximal outcomes are inherently *short-term*—they unfold over minutes, hours, or days, rather than months or years. The accumulation of such outcomes over time creates desired patterns of long-term behavior and health outcomes. Proximal outcomes themselves, however, are those patterns' behavioral, psychological, and physiological building blocks that can be observed at much shorter time-scales.

The short-term nature of proximal outcomes provides three key advantages for HCI evaluations. First, proximal outcomes can be assessed in relatively short studies lasting weeks or months. Since they can be assessed longitudinally, even a short 6-week study can produce 40+ data points per person for a proximal outcome that is assessed daily, such as daily step count or caloric intake. For proximal outcomes of intervention components that can be delivered several times a day, such as reminders to measure blood pressure, this number will be much higher. This means that HCI researchers can assess effects of their interventions on proximal outcomes with good power within the time and resource constraints of typical HCI projects, something that is rarely possible for more distal behavior-change outcomes.

Second, the short-term nature of the outcomes and their repeated measurement means that evaluations can assess not only whether a component has an effect at all, but how that effect changes over time. The dynamics of effects of different types of intervention components and their designs is a key type of data that HCI researchers can contribute to the broader behavior-change science, as the data can be interpreted in the light of rich qualitative evidence that HCI researchers are used to collecting in their studies.

Third, since many BCTs are used in different environments, proximal outcomes will reflect any differences in the impact that those technologies have in these different contexts. Interventions that record contextual information at times when individual components are provided or used enable modeling of the influence that the environment and the person's state have on intervention effectiveness, and how that influence changes over time. Proximal outcomes and context-aware technology thus enable a genuinely new form of evidence about conditions under which individual intervention components work, providing a deeper level of understanding of behavior change processes than was possible.

In summary, focus on proximal outcomes would provide HCI researchers a way to conduct rigorous quantitative evaluations of their novel interventions, contributing both to HCI and broader behavior-change science, all within the resource constraints of typical HCI research.

Looking for causes: Study designs for causal inference

Ultimately, a key question that all developers of BCTs care about is whether their intervention produced the desired effect. This is fundamentally a *causal* question [36], as it asks whether any observed changes in behavior are actually due to the intervention rather than to some other factor. Insofar as behavior change happens, it almost always comes about from a confluence of factors that might include changes in social relationships (a smoker starts dating a non-smoker), exogenous shocks (having a baby or a friend having a heart attack), changes in the physical environment (moving to a more walkable part of town) and so on. In traditional randomized controlled trials, researchers use between-subject random assignment to "balance out" such factors and thus increase confidence that any differences in outcome between the control arm and intervention arm was in fact due to the intervention, and not some other confounder that researchers didn't measure. Use of proximal outcomes, paired with appropriate study designs, enables HCI researchers to apply similarly rigorous casual inference to early-stage evaluations of novel technologies as well. Although a detailed review of experimental designs is beyond the scope of this paper, we briefly review two types of design, factorial experiments and single-case designs, that fit within the constraints of HCI research and can be used to generate evidence of the kind that we argue the field needs.

Factorial designs [4,7] are extensions of the RCT paradigm where intervention provision is randomized to support causal inference. Traditional factorial designs use between-subject randomization to allocate each participant to receive a version of the intervention that contains only certain "levels" of each component. Level might be the presence or absence of a component, different intensity or type of the component and so on. For instance, some participants would get reminders to plan their activity and others would not; some participants would get a daily step goal and others would not. In a factorial design, multiple components can be tested simultaneously, and participants are randomized into cells representing combinations of component levels (e.g., 2x2 cell design for a goals on/off and planning on/off study). Comparing participants in all cells that received one level of a particular component with participants in all cells that received another level of that component provides a way to assess causal effects for that component. In a balanced design (like our 2x2 example), this means that the same number of participants is used to assess effects of each component. The use of proximal outcomes makes traditional factorial designs more efficient, since repeated measurement of these outcomes increases power.

Micro-randomized trials (MRTs) [26,31] are a form of sequential factorial design in which components that can be repeatedly administered (e.g., medication reminders) are

[1] Note that this requirement means that proximal outcomes and intervention components might need to be defined in tandem. If, say, motivational text messages are intended to increase self-efficacy, but changes in self-efficacy can only be seen weekly, the intervention component might be several text messages over a week rather than a single text.

randomized, for each participant, each time that they can be provided. For example, every morning each participant in a walking study could be randomized to receive or not receive a daily step goal. Insofar as the component that is being micro-randomized has a proximal outcome that can be observed both when the component is provided and when it's not provided (e.g., daily step count), the difference in the proximal outcome between all times when the component was provided and all times when the component was not provided gives an estimate of the causal effect of that intervention component. Since in micro-randomized trials this contrast can pool across both between-person and within-person differences in outcome, MRTs require far fewer participants than trials that rely on between-person randomization alone [31]. Studies that last two to three months can often be done with as few as 30 to 40 people. A sample size calculator for MRTs is available at: https://jisun.shinyapps.io/SampleSizeCalculator/

What makes factorial designs particularly powerful for HCI research, is that between-subject and within-subject randomization can be combined in the same study to estimate effects of several components. Some components, like glanceable displays, are either part of a system or they are not; they cannot be given and taken away repeatedly over time. Other "pull components" that individuals access at will, such as educational resources, have the same property. Such components, if they need to be tested, can be randomized at baseline. Other components, such as reminders, messages of encouragement, etc. are inherently designed to be delivered repeatedly over time. Such components can be sequentially randomized. By using such a factorial design, HCI researchers can assess causal effects of their technologies far more rigorously than is possible with observational methods typically used in HCI field studies.

Single-case experimental designs [11,12] provide a different methodology for causal inference by systematically varying when participants have access to an intervention. For instance, after a week of baseline, participants in a walking study might be sent step goals for a week, after which the goals would be withdrawn for another week (often called an A/B/A or reversal design). A researcher would then look to see if the participants' daily steps are different for the week when they received the goals than the weeks when they were not given goals. Such designs provide an efficient way of initially assessing whether an intervention has an effect, often with just a handful of participants (as "control" occurs within an individual, not between). While single-case studies have traditionally evaluated whole interventions, and not individual intervention components, their efficiency makes them a great candidate for doing exploratory studies of simple systems that only include one component that the researcher is developing (such as our goal variability example). Such quick-and-dirty evaluations can provide initial evidence for "is anything here?" questions—providing researchers with a check on whether the intervention idea they are considering is worth exploring further.

Types of evidence

Factorial experiments (and, when thoughtfully used, single case designs), combined with measurement of proximal outcomes, can help generate evidence that is an excellent fit for the research goals we outlined above. They allow researchers to understand functioning of individual intervention components in their systems at a very granular level, allowing them to answer four types of questions:

Is the component having an effect at all?

The most basic type of evidence is gathered by assessing whether, on average, there is a difference in the proximal outcome when the component is present vs. when it's absent, averaging over the duration of the study and other factors. The result, a marginal main effect, tells us if there is any impact of the component on its proximal outcome. In the case of our example of daily step goals, the main effect would tell us whether individuals walk more, on average, when they have a step goal vs. not. Contrasts between multiple versions or designs of a component can also be assessed. Insofar as a component has multiple relevant proximal outcomes, an effect on each can be estimated.

Does the effect of the component change over time?

Since proximal effects are repeatedly measured, factorial studies can also assess if the effect of a component is changing over time. It might be that step goals have a large effect initially, but that this effect slowly dissipates as participants internalize the step goal. Or it might be that participants become habituated to reminders to measure their blood pressure and they stop paying attention to them. This kind of change in the effect over time can be modeled due to the short-term nature of proximal outcomes.

For whom is the intervention working?

A common theme in behavioral science literature is that there is a great deal of heterogeneity of intervention effects, and that they don't work equally well for everyone [20]. Assessing how component effects are moderated by age, gender, or other baseline characteristics thought to be important can create an evidence base that can help researchers tailor interventions to individuals by providing them only with the components or specific designs of those components that are most likely to work for them. For instance, occasionally giving someone a very high activity goal might work well for high novelty seekers (people who are adventurous), but it might turn off low novelty seekers who might do worse if an intervention tries to overly challenge them. Like contextual moderation, which we describe next, moderation by baseline characteristics is a key component of understanding the range of conditions under which an intervention is effective. A thoughtfully selected set of baseline measures could help HCI researchers contribute to this evidence base even in very early-stage evaluations of novel technologies.

In what contexts does the intervention work?

Finally, focus on proximal effects provides an opportunity to assess how the effect of an intervention component is

influenced by the context in which the component is used or provided, both in terms of the environment (location, time of day, etc.) and the person's own state (e.g., stress level). It might be that having a step goal is mostly effective when participants have not walked much on the previous day, or when they are in a new location and they are not able to follow regular routines. Or, as we found in a recent study of one of our own systems, it might be that planning of physical activity for the next day works, for sedentary adults, only on weekdays and not on weekends. For individuals who perceive physical activity as an obligation, weekends are, apparently, a time when they want to get a break both from their jobs and from needing to be active. This level of evidence makes it possible to deeply tailor interventions not only to person-level characteristics but also to contexts, providing interventions only when they are most likely to be effective [34]. Across all components of an intervention, such contextual tailoring allows us to both increase effectiveness of our technologies and to reduce user burden that can result from having to engage with an intervention at times and places when its support is not needed or wanted.

These four types of evidence can greatly enrich the granularity with which HCI researchers understand the technologies they are developing and can help to flesh out the types of questions raised in the introduction. When triangulated with qualitative data, such quantitative assessments of intervention effects can go a long way toward helping us understand how and why BCTs function to help individuals make desired changes in their behavior.

Finally, it is important to note that these four types of evidence provide insights about functioning of intervention components *as they were designed and implemented*. In other words, these evaluations directly assess concrete intervention components contained in concrete systems, not the abstract mechanisms of change that are supposed to underlie them. There are two implications of this: first, this means that different designs of an intervention idea can be directly compared by including them as levels of a component in a factorial experiment. For instance, if a researcher wants to compare two designs of a medication reminder (e.g., how it's presented on the screen or framing of the language), these different designs can be assessed by sequentially randomizing them in a micro-randomized trial. If variants are chosen to reflect important points in the design space for a particular intervention idea, such studies can greatly advance our knowledge about which design elements matter, when they matter, and for whom they matter.

Second, as the evidence of the kind we described here accumulates, it becomes possible to synthesize across studies to assess functioning of more general intervention ideas, such as glanceable displays, as the literature would contain results for various forms of such displays. Similarly, results from studies of systems that include different intervention components that rely on the same underlying mechanisms of change (e.g., priming) could be synthesized to gain deeper insights about the functioning of those mechanisms. Both types of synthesis can greatly increase our understanding of the behavior-change process and how to effectively bring about change through technology.

DISCUSSION AND CONCLUSIONS

In this paper we have argued that HCI researchers who develop and evaluate BCTs should adopt an evaluation strategy that focuses on assessment of proximal outcomes of individual intervention components. This strategy has three key advantages: (1) it provides a rigorous way to evaluate efficacy of individual components of a BCT using the types of studies that HCI researchers are accustomed to running; (2) It connects HCI to other behavior-change sciences by helping HCI researchers clearly articulate intervention ideas and mechanisms of action that their technologies embody, making it easier to interpret their findings and to synthesize them with other evidence to better understand the functioning of the constructs and mechanisms on which behavior-change interventions are based; and (3) it provides a way of rigorously assessing comparative effectiveness of different designs of the same intervention idea, creating a much stronger evidence base for design guidelines.

By enabling rigorous evaluations during early-stage technology development, the approach we outlined allows HCI researchers to focus on inventing novel intervention ideas and novel ways to design intervention components, while at the same time creating robust evidence about whether these new ideas and designs are having their intended effects. As the evidence accumulates across studies, the field will be enriched with new intervention strategies that have been found to be effective, as well as with robust knowledge about which design aspects matter when particular intervention strategies, such as self-monitoring of food, are implemented. Insofar as novel intervention components are found to be effective in preliminary studies, HCI researchers could release reusable versions of those components as intervention modules that other researchers can use in their projects [19]. For behavioral scientists who have little design expertise, gaining access to well-designed intervention modules would be a huge win; for HCI researchers, the use of their modules by others would provide broader impact for their work and enable their designs to be further studied to understand more deeply for whom and in what contexts those designs are effective. Finally, as the approach we advocated relies on the ability to measure proximal outcomes, HCI researchers could contribute by creating innovative ways of capturing, in passive or low-burden ways, a range of behavioral and mechanistic outcomes, thus advancing the ability of behavioral-change scientists to more deeply study and understand the processes that make human beings tick.

ACKNOWLEDGEMENTS

This work was supported by funding from the Robert Wood Johnson Foundation (PI: Hekler, 71995) and from the National Heart Lung, and Blood Institute (PI: Klasnja, 1R01HL125440).

REFERENCES

1. Abraham, Charles and Michie, Susan.. 2008. A taxonomy of behavior change techniques used in interventions. *Health Psychol* 27, 3: 379-87. http://doi.org/10.1037/0278-6133.27.3.379

2. Collins, Francis S and Varmus, Harold.. 2015. A new initiative on precision medicine. *New England Journal of Medicine* 372, 9: 793-795.

3. Collins, Linda M, Chakraborty, Bibhas, Murphy, Susan A and Strecher, Victor.. 2009. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials* 6, 1: 5-15.

4. Collins, Linda M, Dziak, John J and Li, Runze.. 2009. Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods* 14, 3: 202-224. http://doi.org/10.1037/a0015826

5. Collins, Linda M, Murphy, Susan A, Nair, Vijay N and Strecher, Victor J.. 2005. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine* 30, 1: 65-73.

6. Collins, Linda M, Murphy, Susan A and Strecher, Victor.. 2007. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *Am J Prev Med* 32, 5 Suppl: S112-8. http://doi.org/10.1016/j.amepre.2007.01.022

7. Collins, Linda M., Dziak, John J., Kugler, Kari C. and Trail, Jessica B.. 2014. Factorial Experiments: Efficient Tools for Evaluation of Intervention Components. *American Journal of Preventive Medicine* 47, 4: 498-504. http://doi.org/10.1016/j.amepre.2014.06.021

8. Consolvo, Sunny, Bentley, Frank and Hekler, Eric B.. 2017. *Mobile User Research: A Practical Guide.* Morgan & Claypool Publisher, Williston, VT .

9. Consolvo, Sunny, Klasnja, Predrag, McDonald, David W, Avrahami, Daniel, Froehlich, Jon, LeGrand, Louis, Libby, Ryan, Mosher, Keith and Landay, James A.. 2008. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing*, 54-63.

10. Consolvo, Sunny, McDonald, David W, Toscos, Tammy, Chen, Mike Y, Froehlich, Jon, Harrison, Beverly, Klasnja, Predrag, LaMarca, Anthony, LeGrand, Louis, Libby, Ryan, Smith, Ian and Landay, James A.. 2008. Activity sensing in the wild: a field trial of UbiFit garden. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 1797-1806.

11. Dallery, Jesse, Cassidy, Rachel N and Raiff, Bethany R.. 2013. Single-Case Experimental Designs to Evaluate Novel Technology-Based Health Interventions. *Journal of Medical Internet Research* 15, 2: e22. http://doi.org/10.2196/jmir.2227

12. Dallery, Jesse and Raiff, Bethany R.. 2014. Optimizing behavioral health interventions with single-case designs: from development to dissemination. *Transl Behav Med* 4, 3: 290-303. http://doi.org/10.1007/s13142-014-0258-z

13. Dow, Steven.. 2016. Probe to Learn, Probe to Design. *interactions* 23, 4: 22-23. http://doi.org/10.1145/2931079

14. Dow, Steven P, Glassco, Alana, Kass, Jonathan, Schwarz, Melissa, Schwartz, Daniel L and Klemmer, Scott R.. 2010. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-efficacy. *ACM Trans. Comput.-Hum. Interact* 17, 4: 18:1-18:24. http://doi.org/10.1145/1879831.1879836

15. Froehlich, Jon, Dillahunt, Tawanna, Klasnja, Predrag, Mankoff, Jennifer, Consolvo, Sunny, Harrison, Beverly and Landay, James A.. 2009. UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In *Proceedings of the 27th international conference on Human factors in computing systems*, 1043-1052.

16. Havassy, B E, Hall, S M and Wasserman, D A.. 1991. Social support and relapse: commonalities among alcoholics, opiate users, and cigarette smokers. *Addict Behav* 16, 5: 235-46.

17. Hekler, E B, Buman, M P, Poothakandiyil, N, Rivera, D E, Dzierzewski, J M, Aiken Morgan, A, McCrae, C S, Roberts, B L, Marsiske, M and Giacobbi, P R.. 2013. Exploring Behavioral Markers of Long-Term Physical Activity Maintenance: A Case Study of System Identification Modeling Within a Behavioral Intervention. *Health Education & Behavior* 40, 1 Suppl: 51S-62S. http://doi.org/10.1177/1090198113496787

18. Hekler, Eric B, Klasnja, Predrag, Froehlich, Jon E and Buman, Matthew P.. 2013. Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems* (CHI '13), 3307-3316. http://doi.org/10.1145/2466416.2466452

19. 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. http://doi.org/10.1007/s13142-016-0395-7

20. Hekler, Eric B, Michie, Susan, Pavel, Misha, Rivera, Daniel E, Collins, Linda M, Jimison, Holly B, Garnett, C, Parral, S and Spruijt-Metz, Donna.. Advancing models and theories for digital behavior change. *American Journal of Preventive Medicine.*

21. Hovsepian, Karen, al'Absi, Mustafa, Ertin, Emre, Kamarck, Thomas, Nakajima, Motohiro and Kumar, Santosh..

2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 493-504.

22. Kay, Matthew, Nelson, Gregory and Hekler, Eric.. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI '16).*

23. Kessler, Rodger and Glasgow, Russell E.. 2011. A proposal to speed translation of healthcare research into practice: dramatic change is needed. *American journal of preventive medicine* 40, 6: 637-644.

24. 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. http://doi.org/10.1371/journal.pone.0062613

25. Klasnja, Predrag, Consolvo, Sunny and Pratt, Wanda.. 2011. How to evaluate technologies for health behavior change in HCI research. In *CHI '11: Proceedings of the 2011 annual conference on Human factors in computing systems*, 3063-3072. http://doi.org/10.1145/1978942.1979396

26. 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. http://doi.org/10.1037/hea0000305

27. Kravitz, R L and Duan, N.. 2014. *Design and Implementation of N-of-1 Trials: a users guide.* Agency for Healthcare Research and Quality, Rockville, MD. from www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm

28. Kumar, Santosh, Nilsen, Wendy J, Abernethy, Amy, Atienza, Audie, Patrick, Kevin, Pavel, Misha, Riley, William T, Shar, Albert, Spring, Bonnie, Spruijt-Metz, Donna, Hedeker, Donald, Honavar, Vasant, Kravitz, Richard, Craig Lefebvre, R, Mohr, David C, Murphy, Susan A, Quinn, Charlene, Shusterman, Vladimir and Swendeman, Dallas.. 2013. Mobile Health Technology Evaluation: The mHealth Evidence Workshop. *Am J Prev Med* 45, 2: 228-36. http://doi.org/10.1016/j.amepre.2013.03.017

29. Lane, Nicholas D, Mohammod, Mashfiqui, Lin, Mu, Yang, Xiaochao, Lu, Hong, Ali, Shahid, Doryab, Afsaneh, Berke, Ethan, Choudhury, Tanzeem and Campbell, Andrew.. 2011. BeWell: A smartphone application to monitor, model and promote wellbeing. In *5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2011).*

30. Lei, H, Nahum-Shani, I, Lynch, K, Oslin, D and Murphy, S A.. 2012. A SMART design for building individualized treatment sequences. *Annual Review of Clinical Psychology* 8: 21-48.

31. Liao, Peng, Klasnja, Predrag, Tewari, Ambuj and Murphy, Susan A.. 2016. Sample size calculations for micro-randomized trials in mHealth. *Stat Med* 35, 12: 1944-71. http://doi.org/10.1002/sim.6847

32. Lin, Mu, Lane, Nicholas D, Mohammod, Mashfiqui, Yang, Xiaochao, Lu, Hong, Cardone, Giuseppe, Ali, Shahid, Doryab, Afsaneh, Berke, Ethan and Campbell, Andrew T.. 2012. BeWell+: multi-dimensional wellbeing monitoring with community-guided user feedback and energy optimization. In *Proceedings of the conference on Wireless Health*, 10.

33. Michie, Susan, Ashford, Stefanie, Sniehotta, Falko F, Dombrowski, Stephan U, Bishop, Alex and French, David P.. 2011. A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy. *Psychol Health* 26, 11: 1479-98. http://doi.org/10.1080/08870446.2010.540664

34. Nahum-Shani, Inbal, Hekler, Eric B. and Spruijt-Metz, Donna.. 2015. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology* 34, Suppl: 1209-1219.

35. Nahum-Shani, Inbal, Qian, Min, Almirall, Daniel, Pelham, William E, Gnagy, Beth, Fabiano, Gregory A, Waxmonsky, James G, Yu, Jihnhee and Murphy, Susan A.. 2012. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological methods* 17, 4: 457.

36. Pearl, Judea.. 2009. *Causality.* Cambridge university press.

37. Riley, William T, Rivera, Daniel E, Atienza, Audie A, Nilsen, Wendy, Allison, Susannah M and Mermelstein, Robin.. 2011. Health behavior models in the age of mobile interventions: are our theories up to the task? *Transl Behav Med* 1, 1: 53-71. http://doi.org/10.1007/s13142-011-0021-7

38. Riley, William T and Rivera, Daniel E.. 2014. Methodologies for optimizing behavioral interventions: introduction to special section. *Translational behavioral medicine* 4, 3: 234.

39. Spruijt-Metz, Donna and Nilsen, Wendy.. 2014. Dynamic Models of Behavior for Just-in-Time Adaptive Interventions. *IEEE Pervasive Computing*, 3: 13-17.