How to Evaluate Technologies for Health Behavior Change in HCI Research

Predrag Klasnja[1], Sunny Consolvo[3], & Wanda Pratt[1,2]

[1]Information School & DUB group
University of Washington
Seattle, WA 98195, USA
klasnja@uw.edu

[2]Biomedical & Health Informatics
University of Washington
Seattle, WA 98195, USA
wpratt@uw.edu

[3]Intel Labs Seattle
Seattle, WA 98105, USA
sunny.consolvo@intel.com

ABSTRACT

New technologies for encouraging physical activity, healthy diet, and other types of health behavior change now frequently appear in the HCI literature. Yet, how such technologies should be evaluated within the context of HCI research remains unclear. In this paper, we argue that the obvious answer to this question—that evaluations should assess whether a technology brought about the intended change in behavior—is too limited. We propose that demonstrating behavior change is often infeasible as well as unnecessary for a meaningful contribution to HCI research, especially when in the early stages of design or when evaluating novel technologies. As an alternative, we suggest that HCI contributions should focus on efficacy evaluations that are tailored to the specific behavior-change intervention strategies (e.g., self-monitoring, conditioning) embodied in the system and studies that help gain a deep understanding of people's experiences with the technology.

Author Keywords

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

ACM Classification Keywords

H5.2 Information interfaces and presentation (e.g., HCI): User interfaces (Evaluation/Methodology). J.3 Life and Medical Sciences: Medical information systems.

General Terms

Experimentation, measurement.

INTRODUCTION

In recent years, there has been an explosion of HCI research on technologies for supporting health behavior change. HCI researchers have developed systems for encouraging physical activity [2,7,8,24], healthy diet [12,17,23], glycemic control in diabetes [26,39], and self-regulation of emotions [31]. Work in this area is rapidly becoming a staple at many of the field's preeminent publishing venues.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

This work has the potential to make a meaningful impact on society. The prevalence of chronic diseases such as diabetes, obesity, and coronary heart disease continue to rise and are now responsible for over 70% of U.S. healthcare expenditures [20]. Some of the most important risk factors for these conditions are behavioral, including smoking, physical inactivity, excessive food intake, and diets heavy in trans fats. A successful change in these behaviors is a fundamental aspect of both prevention and effective management of chronic conditions, as well as an important contributor to health and wellbeing more broadly. Due to their low cost, high penetration, and integration in people's everyday lives, technologies such as mobile phones, web applications, and social networking tools hold great promise for supporting individuals as they strive to adopt and sustain health-promoting behaviors. HCI research can significantly contribute to the design of innovative and effective tools that help people in these efforts.

However, as HCI researchers increasingly engage in the design of systems for health behavior change, an important question arises: how should interventions for health behavior change be evaluated *within the context of HCI research*? The question is twofold. First, what types of evaluations are appropriate and useful for systems that HCI researchers in this area are developing? And second, how should the research output of this work—primarily in the form of publications—be evaluated? These questions are key, we believe, to moving this area of HCI forward, and their careful consideration should aid both researchers and reviewers working in this area.

In this paper, we argue that the obvious answer to these questions—namely, that the goal of an evaluation of a technology for health behavior change should be to show that the technology brought about the intended change in behavior—is too limited. We argue that behavior change in the traditional clinical sense is not the right metric for evaluating early stage technologies that are developed in the context of HCI research. However, a narrower notion of efficacy, one that tailors outcome measures to the particular intervention strategies a technology employs, can enable HCI researchers to test whether their systems are doing what they are intended to do even at early stages of development. Just as importantly, qualitative studies that focus on people's experiences with the technology could help researchers understand why and how their system is

working—an outcome that we consider a central contribution of HCI work in this domain.

DIFFICULTIES WITH EVALUATING BEHAVIOR CHANGE

For anyone trying to design a system intended to help people change their habits so they can become healthier, the question of whether the system they have built actually does what it was intended to do is of utmost importance. Naturally, we want our systems to work, and we want to carefully assess whether they work, particularly when human lives are at stake. Consequently, the types of evaluations reported in papers on technologies for health behavior change and the claims that authors make about their results often try to address the question of whether the new system was effective. Reviewers who are evaluating research on such systems often ask the same question: did the technology that the authors present really change people's behavior or are reported effects purely due to participation in a study, short-lived interest in a new technology, or similar confounding factors? HCI papers on technologies for health behavior change commonly get some form of this question in the reviewer comments. Both parties tacitly agree that a central outcome of reported studies should be to show that the new system worked—that it helped bring about the intended change in behavior.

This assumption is certainly reasonable. It fits with the emphasis that much of the field of HCI has on quantitative evaluations [4,18], and it reflects the format of evaluation that is common in many subareas of HCI. For example, papers on new interaction techniques attempt to show that the new technique is more effective than current techniques that address the same problem; similarly, HCI systems papers try to show that the new system can do what it is intended to do—help users better understand their privacy settings, manipulate GUI applications by recognizing their widgets at the pixel level [14], and so on. Showing that the system under study is effective at achieving its primary goal is a basic form of evaluation in HCI research. Shouldn't research on technologies for health behavior change be held to the same standards?

In this section, we argue that although the desire to demonstrate behavior change is understandable, as an aim for evaluations in HCI research, it is often not feasible. Behavior change is a complex, long-term process with high relapse rates. To convincingly demonstrate that a technology contributed to such a process requires large-scale, long-term studies that can typically not be done with early-stage and error-prone systems frequently developed in HCI. But such studies are also limited from the perspective of HCI. Even when such studies are done—as they are in health sciences—they often do not reveal *why* the technology that is being evaluated worked or did not work. This knowledge, however, is of central importance to HCI researchers as they attempt to develop novel, yet effective systems. Other types of evaluation are thus also needed.

Behavior change as a long-term process

Based on extensive empirical research and a review of the literature, Prochaska and colleagues [36] have concluded that for behavior change to truly stick, a person has to maintain the target behavior for several years. Prochaska et al cite studies that indicate that for smoking, it is only after five years of abstinence that the relapse rate goes down to around 7%. After a year of abstinence, it is still at 47%. The literature on other health behaviors shows a similar, albeit less extreme pattern. Even among patients who undertake an intensive 12-week cardiac rehabilitation program following a heart attack, six months after the program ends, less than half of patients still exercise regularly [30,37]. The adherence rates for maintenance of low-fat diets after cardiac rehabilitation are similar: one and three years after program completion, adherence rates are 49% and 42%, respectively [41]. And these numbers are for a population for whom changing habits can literally be a question of life or death. In the healthy population, Marcus et al [27] cite data that even among 7,135 regular users of the YMCA, 81% had a break—that is, seven or more days of non-attendance—in their exercise routine over the previous year. The mean number of lapses was 4.8 per year and a "break" lasted on average 36 days.

Thus, for many common health behaviors, changing that behavior is an ongoing process where higher levels of health-promoting activity are interspersed with lapses and setbacks. Due to this protracted nature of behavior change, Marcus and colleagues [27] have called for studies of interventions for encouraging physical activity to include at least 24 months of follow-up. If one truly wants to find out whether a behavior has solidified, the need for such longitudinal studies is unavoidable. Only multi-year studies with repeated follow-ups can accurately assess whether new health-promoting habits have truly been adopted.

This long-term nature of the behavior change process has an important implication for HCI research. Unless we have really conducted such longitudinal studies, we should not make claims that participants in our studies truly changed their habits. We might have seen a change in frequency of a behavior during the course of the study, but evidence suggests that such changes are often short-lived. The changes we observe *might* lead to a more permanent behavior change, but without long-term follow-ups, we simply don't know. A correlate of this point is that in our role as reviewers, we must understand that when we ask authors to demonstrate that their participants formed new habits, we are really asking for such long-term evaluations. However, few of us, we suspect, really believe that a multi-year evaluation is needed as an initial evaluation of a novel HCI technology. Yet, this type of critique is demanding precisely that type of unrealistic evaluation.

Even if it takes several years to demonstrate behavior change *maintenance*, showing that some behavior change took place in the short term is surely more straightforward?

In fact, we argue later that one way that HCI researchers can demonstrate the efficacy of their systems is precisely by showing that certain types of short-term changes in behavior occurred (along with other effects), provided that those changes are properly selected and measured. Nonetheless, we want to suggest that even to conclusively demonstrate that short-term changes in behavior are due to the technology can be infeasible for early HCI research and that another evaluation goal—understanding how and why a technology works or does not work—is an important contribution for this stage of research.

The complexity of the behavior change process

One of the main reasons why behavior change is such a difficult, long-term process is the sheer number of factors that are involved in making a significant change in behavior. Many health behaviors—such as physical activity, diet, smoking, and stress—are deeply ingrained in people's daily routines and the webs of social and institutional relationships. Even a single change, such as increasing physical activity, often requires the individual to restructure her priorities as well as her daily and social routines. If a person wants to start biking to work, for example, she might need to arrange for her spouse to take over childcare duties, find a place where she can shower before work and lock up the bike, figure out how to deal with bad weather and so on. Similarly, making changes to one's diet can be made more difficult if the person habitually goes to lunch with her boss and coworkers who prefer restaurants with fatty foods or if she has regular pizza and movie nights with her spouse. Changes in such routines affect much more than what one eats; these routines play an important role in formation and maintenance of social relationships, negotiation of workplace politics and other aspects of a person's life that have very little to do with food per se. Attempts to alter such routines can negatively affect important relationships and often face pushbacks. Indeed, the clinical literature indicates that social influence is by far the most common cause of relapse during addiction treatment and that it is a key determinant of success of health behavior change in general [28,30,42].

Beyond social and logistical factors, behavior change is often also thwarted by changes in circumstances. Getting injured or sick, having a particularly busy period at work, going on a trip and similar changes in daily routine often disrupt fragile new habits. A disruption is particularly likely if the changes are unexpected (e.g., getting the flu), or if the person has not made explicit plans about how to deal with the upcoming changes.

Finally, internal factors are also significant. The outcomes that make health behavior change desirable—reduction in health risks, feeling better, looking better, avoiding adverse effects of illnesses such as diabetes—are often both temporarily distant and uncertain. Behavioral economics has shown that such outcomes are very deeply discounted and that the subjective value of more immediate rewards—a tasty pizza, a pleasant evening of watching TV—can easily outweigh the value of health-related outcomes at the moment of choice (e.g., when a person comes back from work tired and is deciding whether to go to the gym or have a frozen pizza and relax with a book) [25].

Similarly, an extensive literature shows that the human capacity for self-control is limited. The same finite self-regulatory resources that are needed to change one's habits are also used to navigate self-regulatory challenges of everyday life, such as trying to stay unemotional during a difficult conversation with one's boss, or dealing with a crying child. Behavior change setbacks commonly occur during times of high stress and in situations when one's self-regulatory resources are depleted by demands in other areas of one's life, such as work pressures or emotionally difficult situations like caring for a sick family member [32,33]. Similarly, the extent to which behavior change goals have been internalized and integrated (i.e., whether they are perceived as something that the person really *wants* to do or just something that she feels she *should* do) can significantly affect the success of a person's behavior change efforts [11]. Furthermore, falling off the wagon even a little—by having too much ice cream, for example—often triggers much bigger setbacks for the behavior change process. Marlatt & George [28] have termed this "*I already screwed up, so what's the point of trying*"-reasoning the "abstinence violation effect."

This brief discussion should provide at least a glimpse into the complexity of the behavior change process and why it is so difficult and protracted. What should be clear, we hope, is that few technologies, however clever, could hope to address all aspects of the behavior change process. How successful a person is in her efforts to change her habits during a particular period will depend on many factors beyond the technology that she is using to make behavior change easier.

Thus, assessing a technology's contribution to behavior change, among the myriad other factors that shape an individual's behavior at any given time, requires large studies and significant resources. Anyone who has tried to deploy a research prototype in the field with members of the general public knows just how challenging such deployments can be even with 10 to 20 participants over a few weeks. To unambiguously demonstrate the effect of a technology on behavior change while controlling for other factors, such as renewed commitment, social pressures, or the effects of participating in a study, requires a deployment to hundreds or even thousands of people and a matching control group. The error-prone nature of early research technologies—and the resulting need for technical support—make such deployments prohibitive for many HCI technologies, especially in early stages of development.

Efficacy trials: a technology may be effective, but *why*?

Studies that can conclusively demonstrate effectiveness of a technology for behavior change do exist, of course.

Randomized control trials (RCT), the gold standard of efficacy research in health sciences, are precisely the kind of evaluation that can show whether a technology helped bring about a change in behavior. Originally developed to test effectiveness of pharmacological substances, in recent years, RCTs are increasingly being used for technological interventions as well. For example, a recent systematic review by van der Berg and colleagues [5] identified 10 RCTs of internet-based interventions for promoting physical activity. RCTs of other types of health promotion technologies, such as mobile phone applications, are also becoming common.

Once a novel HCI technology becomes mature, an RCT, like those conducted in health sciences, becomes necessary if we really want to demonstrate that the new technology is effective. In fact, we later argue that HCI researchers should not shy away from such studies and that they can productively collaborate with colleagues in health sciences to conduct them. For an early technology, however, large controlled studies have other downsides beyond mere feasibility. In particular, RCTs often reveal little about *why* the technology under evaluation is or is not effective. Understanding why a technology worked or did not work, however, is precisely how HCI researchers can determine what to do to advance technology design.

For example, Hurling and colleagues [21] recently used a small RCT to evaluate a system for encouraging physical activity that combined a wrist-worn accelerometer, a mobile phone, and a web application that implemented several behavior change strategies: self-monitoring, identification of barriers to change, planning, problem-solving, public commitment, and customized feedback. The study assessed how this system performed in comparison with a control intervention. The control group used accelerometers, but did not get feedback about their activity nor did they have access to the website. All participants (N=77) took part in a three-week pretest period during which they used the accelerometer and the phone to establish their baseline level of physical activity. After this initial period, the participants were administered the International Physical Activity Questionnaire (IPAQ) and researchers measured participants' weight, height, body fat percentage, and resting blood pressure. The participants were then randomized into the control group (N=30) or experimental intervention (N=47) for nine weeks. The study used the IPAQ and accelerometer data as primary outcome measures, and the change in weight, body fat, and blood pressure, along with a set of cognitive variables as secondary measures.

Although the IPAQ data showed no significant difference between the groups in overall physical activity, it indicated that the intervention group increased their amount of leisure time activity significantly more than the control group. The accelerometer data also showed that the intervention group had a significantly higher reduction in the amount of time spent sitting. The study found that the intervention group also lost more body fat than the control group and had a higher increase in perceived control and intentions related to exercise.

In many respects, Hurling et al's results are compelling. The study was well thought out, it used a range of outcome measures, and the study design provided evidence that the observed differences were due to the differences between the two interventions—in this case, the access to the behavioral intervention website and the accelerometer data. Yet, at least four issues limit the usefulness of Hurling et al's evaluation and suggest why RCTs should not be seen as the only valid model for evaluating health-promotion technologies in HCI, especially in early stage research.

First, although the study results strongly suggested that the obtained differences in the two groups were due to the difference in the access to technology, Hurling et al's sample size—large by HCI standards but very small by RCT standards—was not big enough to control for many other factors that could have played a role in the observed differences, such as motivation, social pressure and so on. The results were strongly suggestive, but not conclusive. To truly control for a full range of potentially confounding factors, a much larger RCT would be needed. Such factors are the reason why RCTs in health sciences typically enroll hundreds or thousands of participants.

Second, because the intervention combined a variety of different behavior change strategies, the study could not determine which elements were the most effective or how different aspects of the system interacted to affect the overall effectiveness. To determine the relative efficacy of the different system components requires an RCT with several intervention arms—each of which has only certain aspects of an intervention system. Such RCTs require even larger numbers of participants. Furthermore, even such large multi-arm studies typically only show which components are effective, but not why they are effective.

Third, although the comments that their participants left on the intervention website's message board suggested that the graphs of the accelerometer data were a particularly helpful aspect of the intervention, Hurling et al did not collect qualitative data that would be needed to *"make a thorough analysis of how participants perceived the system"* (p. 9). Thus, Hurling et al's study found that their intervention was effective but not how or why it was effective. How the system was used by participants, how well it fit into their daily lives, which aspects of the system they found to be most helpful, what problems they faced, how different components of the system worked together, and other similar questions—the answers to which would be useful for designing future systems—could not be identified with this evaluation. Most RCTs of behavior change technologies suffer from this same weakness.

From the standpoint of HCI, this issue is significant. If we don't know *why* our systems work (or don't work), how can we design better systems in the future? Especially at early stages of development, then, interviews and other qualitative assessments should be an essential part of HCI evaluations. Although, in principle, such assessments can be added to an RCT for at least a subset of the study participants, smaller studies with a significant qualitative component can address many potentially serious design problems before resources are expanded on a large RCT.

Finally, due to their size and cost, efficacy trials typically evaluate complex systems that combine many intervention strategies to maximize effectiveness. Hurling et al's study is a good example of this trend: in terms of size, its 77 participants and three-month study duration are on the small and short end of what is typically done for an RCT. Even such a small RCT, however, was only done on a complex, multi-faceted system that had been under development for several years. The resources and effort required to run true efficacy trials make evaluations of innovative technologies that embody early-stage, high-risk ideas simply infeasible. Insofar as this type of innovation is precisely what HCI as a field tries to do [cf 1], efficacy studies like those done in health sciences are a poor fit for early HCI research.

To summarize, in this section we argued that due to the complexity and long-term nature of the behavior change process, demonstrating efficacy in the traditional clinical sense is infeasible for early HCI research. Efficacy studies like RCTs are extremely valuable for determining whether a specific, complete system can bring about clinically relevant changes in people's behavior. For mature systems, such evaluations are appropriate and important. However, for novel technologies at early stages of development, other evaluation goals, such as trying to gain a deeper understanding of how and why a system is used, are more valuable. In the rest of this paper, we propose alternative types of evaluations for novel HCI technologies.

RETHINKING EFFICACY

The first point that we have to address is whether HCI researchers need to test the efficacy of their technology, given the difficulties that we described earlier. We argue that efficacy is important, even in early stages of technology development, but that a narrow, constrained notion of efficacy can help HCI researchers determine whether their systems are working as intended. In addition, this constrained notion of efficacy has other benefits, including the ability to compare different implementations of a behavior change strategy.

Tying evaluation to behavior change strategies

As we have argued, behavior change is a complex process; however, technological interventions support that process only by implementing specific *strategies*. In his book *Persuasive Technology*, Fogg [15] identifies seven types of persuasive strategies: reduction, tunneling, tailoring, suggesting at the right time, self-monitoring, surveillance, and conditioning. Of these, self-monitoring—tracking one's own activities—is the most prevalent, and most systems for health behavior change include this component. Houston [7], Fish'n'Steps [24], Laura [6], UbiFit [9], the Mobile Lifestyle Coach [17], and PmEB [40] are among recent systems that employ self-monitoring as a central intervention component. Conditioning—usually by means of positive reinforcement, but sometimes also with punishment—is another common strategy. UbiFit, for example, rewards users for meeting their weekly activity goals through the appearance of a butterfly on the background screen of their mobile phone. Fish'n'Steps rewards or punishes its users by changing the facial expression of the user's virtual pet fish based on how close the user came to meeting her daily activity goal. Additionally, health-promotion technologies often implement strategies that are not on Fogg's list, such as social learning (e.g., MAHI [26]), social influence (e.g., teamwork and competition implemented in Houston, Fish'n'Steps, and the Mobile Lifestyle Coach), priming (e.g., UbiFit), goal negotiation and coaching (e.g., Laura [6]), and leveraging intrinsic motivation (ViTo [34]).

These various strategies support the behavior change process in different ways. Self-monitoring, for example, makes the monitored activities more salient to the person. As such, self-monitoring interventions typically work only while the intervention is going on but tend to fade away soon after the intervention is discontinued. Other strategies have a long ramp-up time, but are expected to produce lasting effects (e.g., social learning). Such specific effects are a direct outcome of the psychological processes on which a particular intervention strategy is based.

Taking into account how a particular intervention strategy works can enable us to test whether the system is doing what it is supposed to do without needing to show behavior change in the traditional clinical sense we discussed above. Thus, we propose that a tight link between outcome measures and the specific intervention embodied in a system should be a key factor in designing studies that evaluate HCI technologies for health behavior change.

Consider self-monitoring. Over 40 years of research has shown that simply keeping track of a behavior changes the frequency of that behavior in a desired direction [22,35]. However, this effect is not permanent. Unless other changes and strategies are put in place, once self-monitoring is discontinued, the target behavior tends to slowly return to its pre-intervention levels. Self-monitoring works, in other words, only as long as the individual self monitors. Although an individual can achieve a long-term change in behavior even after she discontinues self-monitoring, the resulting changes are due to other structures that she has put in place to maintain the desired behavior, rather than to self-monitoring itself.

To test a novel self-monitoring intervention, then, an evaluation should assess whether during the period when

participants are actively using the system, the rates of the target behavior increase from their baseline levels prior to the intervention. In addition, it should test whether after the intervention is stopped, the rates of behavior begin to go down again. Such a pattern could be seen in even a few weeks, obviating the need for a long-term study. If this pattern is demonstrated, the researcher could have some confidence that at the most basic level the intervention is doing what it was designed to do. The evaluation would not demonstrate conclusively that the effect was purely due to self-monitoring—in particular, novelty could play a role—but if the results showed a very different pattern than what was expected (e.g., no frequency increase during the intervention or the frequency of behavior continues to increase even after the intervention stops), the researcher would know that something else is going on in addition to or in place of the intended intervention.

Similar reasoning applies to other intervention strategies as well. If the intervention is primarily an educational one, such as the educator-assisted reflection for diabetes management implemented in MAHI (Mamykina, et al., 2008), we would expect that the disease management skills would gradually increase over time, as would the level of the internal locus of control, disease management self-efficacy and other related measures. We would also expect that these effects would persist beyond the end of the intervention. A good evaluation of such an intervention would try to assess the increase of knowledge, self-efficacy, etc. over the course of the intervention, and would then test how these gains in knowledge and the corresponding psychosocial factors held over time. Mamykina et al's study provides a good example of matching the evaluation to the educational intervention strategy. They took all measurements both before and after the intervention (and the changes were compared to those of patients participating in the same diabetes education class but who did not use MAHI). In addition, Mamykina et al write that they intend to conduct 3- and 6-month follow-ups to assess the stability of the changed measures. Some type of a follow-up like the one proposed by Mamykina et al, even if it were done only after a month or so, could be an important component of efficacy evaluations for educational interventions. Mamykina et al's study itself, of course, makes a number of other valuable contributions that are not tied to such a follow-up. We discuss the nature of such contributions and their importance shortly.

One other benefit of targeted evaluations of this sort is that they let researchers explore risky ideas that simply might not work. For example, DeShazo et al [13] developed a set of simple mobile phone games that are intended to teach nutritional skills to patients with diabetes. Although a promising idea, it is not clear whether playing such games during brief microbreaks could really increase patients' knowledge or if patients would want to play the games. An evaluation that tests participants' nutritional knowledge before and after even over a few weeks could begin to answer these questions without the need to test whether participants' eating habits have also begun to change as a result of the intervention.

Finally, we should briefly address the question of how researchers can determine what types of evaluations are appropriate for a particular intervention strategy. Initially, ideas for evaluation will need to be drawn from the literature from which the intervention strategies themselves were drawn (e.g., social psychological literature on priming [e.g., 19], literature on the effects of schedules of reinforcement, etc.). We hope that over time, as more researchers find good ways of testing their interventions, evaluations for common intervention strategies will be systematized and customized to the needs of HCI. Such systematization would make the work of both researchers and reviewers much easier.

Benefits of tailored efficacy evaluations

The conception of efficacy that we have been advocating has an important benefit: it makes it possible to standardize, to some extent, how specific intervention strategies are evaluated, which, in turn can enable comparisons of different implementations of the same intervention strategy.

Consider the example of priming. Priming refers to the psychological phenomenon that elements of the environment—images, words, smells, places and so on—can activate cognitive representations, such as goals and concepts, associated with those environmental stimuli [3,19]. One important characteristic of priming is that a goal activated through contextual cues has the same kind of effects on cognition and behavior as a goal that is activated through conscious pursuit. An activated goal makes it more likely for the person to engage in activities directed toward that goal, notice opportunities for such activities, and be sensitized to information related to the goal. A common example is when people are researching a big purchase, such as a new car. All of a sudden, they see cars they are considering everywhere they turn, they notice and remember information about gas mileage, etc. The activation of the get-a-new-car goal makes the typically unnoticed aspects of their world having to do with cars become much more salient than usual.

In the context of health behavior change, priming has been used by Consolvo et al's UbiFit Garden system [9] to keep individuals engaged with their commitment to physical activity. UbiFit uses the background screen of a mobile phone to display stylized representations of both how physically active the user has been that week and whether she has reached her weekly physical activity goal. Every time the individual uses the phone to make a phone call, check her calendar, or answer a text message, she also sees this representation of her level of physical activity and is reminded of her goal to be physically active.

Although data from interviews in UbiFit studies indicated that seeing the background screen of the mobile phone kept

physical activity in the front of participants' minds, UbiFit's priming component—and that of future systems that implement priming—could be tested in a direct way too. Some possibilities for such direct assessments include: a standard set of experience sampling [10] questions that assess participants' awareness of and confidence in their level of physical activity during the previous week; tasks that test salience of physical activity concepts (e.g., ask participants to generate ideas about benefits of the decision not to drive to work); asking participants to list opportunities in their lives when they can be physically active, etc. If a priming intervention is successful, individuals using the intervention would be expected to do better on such tasks than when they do not have the priming intervention, or when they are in a non-priming control group.

As the field matures, such standard study procedures could be developed for a variety of intervention strategies, enabling researchers to begin to understand how the design and implementation details affect the effectiveness of an intervention strategy. For example, comparisons of different priming interventions for physical activity would enable us to answer important HCI research questions, such as how granular a priming representation needs to be (i.e., is there a need for a separate representation of each category of physical activity as in UbiFit or is a single summary representation enough?), whether a representation needs to provide feedback on the level of physical activity or if an image that a user strongly associates with her physical activity goals is enough, and how the effectiveness of priming for promoting physical activity is affected if other important goals are represented as well (e.g., the goal to spend more time with one's family). Similarly, standard measures, such as consistency of logging, could enable us to compare diet logs that require each food to be journaled separately (e.g., PmEB [40]) with those that use some form of shorthand logging, such as the point system of the Mobile Lifestyle Coach [17] or the food journal in the Wellness Diary [29].

Similarly, the use of common measures could help researchers to understand how culture and context in which users live and work affect the use of their health applications. For example, if evaluations of diet logs consistently assessed the percentage of consumed food items that users logged, researchers could begin to gather quantitative evidence for the importance of having the foods that a particular cultural group eats in the food database or for the importance of the match between how food groups are listed in the system and how users conceptualize them in their day-to-day life. Knowledge gained in this way would be extremely valuable for informing the design of future systems, both those developed by HCI researchers and by researchers in other fields.

BEYOND EFFICACY: OPENING THE BLACK BOX

Although tailored study procedures can help test whether a system is doing what it's supposed to be doing without a need to demonstrate long-term behavior change, in HCI research, efficacy should not be seen as the only valid goal of evaluation—or even as a primary goal. Especially in early stages of development, a deep understanding of the *how* and *why* of the system use by its target users should be a central goal for evaluations of systems for health behavior change. In fact, it is this opening of the black box of the system use—and the design knowledge that results from it—that is arguably the biggest contribution that HCI can make to the development of effective systems in this domain. We argue this point below.

Uncovering potential problems

Complex technical systems rarely come into the world fully formed and bug-free. The initial versions often don't work precisely as the designers intended, both due to problems and bugs and to the unanticipated effects of specific design choices. Thus, in situ or field evaluations of systems for health behavior change should be of the utmost importance in the early stages of development. Rogers et al [38] have argued that even short in situ evaluations that contain a significant qualitative component often uncover a range of problems with a system that traditional lab-based usability evaluations are unlikely to uncover. The literature on systems for health behavior change provides ample evidence for this claim. In the field study of the Mobile Lifestyle Coach, for example, Gasser et al [17] found that participants were extremely frustrated by the inability to edit or delete entries they made in the application's food and diet diary. Although this issue could slip by in a lab usability test and would not show up at all in an RCT, the qualitative component of the field study highlighted this element as being particularly problematic. Similarly, Tsai et al's [40] evaluation of PmEB highlighted issues with entry of food items into their caloric balance tracking application that did not come up during a previous lab usability test.

In addition to identifying bugs and missed features, field studies can uncover unintended consequences of the design. Both Consolvo et al [7] and Lin et al [24] found that some intentional aspects of their designs backfired when their systems were tested in situ. During post-study interviews, Lin et al found that a number of their participants reacted negatively to the competitive aspect of the social version of the Fish'n'Steps application and to the punishment (sad fish) when participants were not being active. Similarly, Consolvo et al found that Houston's focus on step count and lack of facilities for logging other types of physical activity discouraged some participants from engaging in more intense physical activities. Because they could not journal when they went for a bike ride or rock climbing, sometimes the participants decided not to do those activities at all—an effect that Consolvo et al have openly said that they did not intend.

Such issues could have a significant impact on the system's effectiveness. Discovering them allows researchers to make changes to their design to increase the probability that the system will be effective. Just as importantly, the discovery of such potential problems constitutes a contribution in its own right to the body of knowledge on how to design effective systems. For these reasons, smaller studies with a significant qualitative component lasting a few weeks should be conducted even when larger evaluations are a possibility. Jumping directly from a lab usability evaluation to a large efficacy study, as is commonly done in health sciences, means that many design problems are never uncovered, potentially seriously affecting the system's effectiveness and impeding new research.

Understanding use

Beyond uncovering problematic design elements, we need evaluations that help us gain a deep understanding of user perceptions and patterns of system use. We need answers to questions such as when people choose to use or not use a system, whether and how they share their data with other people, what aspects of the system they find most helpful or frustrating and why, and what other things they wish the system could do. Answers to these kinds of questions can help us design technology that fits into people's lives and that is likely to be effective for helping them change their habits. Similarly, field evaluations can uncover a great deal about both how users understand a technology and how their mental models affect what they might do with the technology (e.g., putting a pedometer in a purse instead of wearing it on the waist). Such evaluations can also assess users' need for privacy, a particularly important issue in regards to health information. Although the investigation of such issues has long been a part of HCI evaluations, we emphasize its importance in this domain. Behavior change is a difficult and fragile process. A deep understanding of how technology interacts with other important factors that affect behavior change—people's attitudes and preferences, their relationships, the context in which they live and work, etc.—is critical for the development of effective tools.

For example, Houston, one of Consolvo et al's first tools for supporting health behavior change, was intended to aid individuals in becoming more physically active by enabling them to share their daily step counts with a group of friends. One of the findings that emerged from the interviews during Houston's evaluation was how important it was to the participants to be able to annotate the step counts that they were sharing. Although the participants enjoyed sharing when they were being active, when they had inactive days, sharing became uncomfortable. Being able to annotate their step counts with messages, such as that they had the flu, made it possible for them to continue using the system instead of pulling back due to their discomfort.

Similarly, qualitative studies could help uncover how gender or cultural attitudes affect the use of a behavior change tool. For example, interviews in the 3-month UbiFit trial [8] indicated that some participants—mostly men—had a negative attitude toward the garden motif of the glanceable display. Similarly, a study of an educational cell phone game, such as those developed by DeShazo et al [13], might find that female players are less engaged by a space invaders game than male players are, making it less effective for them than a quiz game with the same educational content.

Finding out about these types of issues makes it possible to design technologies that accommodate subtle factors that shape both the behavior change process and the use of tools aimed at supporting it. This is particularly important since technologies are not simple sums of their various elements. A negative attitude toward the representation used to show feedback on physical activity in a system like UbiFit could offset any positive effect that the self-monitoring or priming components of the intervention would otherwise have. Although some such issues can be envisioned in advance, many will need to be uncovered as individuals begin to use a deployed version of the system in the context of their daily lives. Thus, during the initial evaluations of a novel system, investigation of patterns of use and users' experiences with the system should be seen as a primary evaluation goal.

Uncovering issues of this type is one of the most important contributions that HCI can make to the design of technologies for health behavior change. Cataloging and classifying these issues will help create a body of knowledge about which design elements, in what circumstances, can effectively encourage healthy behavior. This knowledge will help both HCI researchers to build better systems as well as health sciences researchers to understand the importance of design and the influence that specific implementation choices have on the effectiveness of technologies that they are developing and evaluating.

Finally, to maximize the impact of their work, once a research technology becomes mature, HCI researchers should pursue formal validation of the technology's ability to encourage healthy behavior change. Such validations are most likely to occur through collaborations between HCI and health science researchers since the technical and health-assessment demands of such evaluations require the expertise of both fields. We encourage HCI researchers to seek such collaborations and shepherd their technologies to a point where they can become truly useful to the general population. But possibilities for collaboration do not stop here. HCI researchers can also make significant contributions to health science research. For example, the tools developed by HCI researchers, such as MyExperience [16], can enable health science studies that were previously not possible. Similarly, HCI expertise in designing techniques for collection and analysis of rich data, such as real-time physical activity data or the data about the person's environment, could greatly increase what can be learned in large RCTs. The availability of such dense

datasets would make it possible for RCTs to tell us not only whether a technology works, but also about the circumstances when it does or does not work. The knowledge gained with this shift could be great indeed.

CONCLUSION

In this paper, we proposed a way to think about evaluations of technologies for health behavior change in the context of HCI research. In particular, we argued that to truly demonstrate that a technology brought about its ultimate goal of behavior change requires large studies with control groups that are typically not feasible for HCI technologies in early stages of development. At the same time, we proposed that tailoring evaluations to their intervention strategies can enable HCI researchers to show that their systems are doing what they are supposed to be doing, without requiring a full-blown demonstration of behavior change. An added benefit of such tailored evaluations is that they can enable researchers to directly compare different implementations of the same intervention strategy and to learn how different aspects of design or implementation affect the efficacy of that intervention strategy. Finally, we suggested that a critical contribution of evaluations in this domain, even beyond efficacy, should be to deeply understand how the design of a technology for behavior change affects the technology's use by its target audience in situ. It is this knowledge, we argue, that will most readily advance our ability to develop systems that effectively help individuals in the important and challenging task of changing their routines in order to become healthier.

ACKNOWLEDGMENTS

We'd like to thank the CHI reviewers for their thoughtful comments. Their feedback made this paper much stronger.

REFERENCES

1. Abowd, G.D., Mynatt, E.D., and Rodden, T. The human experience [of ubiquitous computing]. *IEEE pervasive computing 1*, 1 (2002), 48–57.
2. Anderson, I., Maitland, J., Sherwood, S., et al. Shakra: Tracking and Sharing Daily Activity Levels with Unaugmented Mobile Phones. *Mobile Networks and Applications 12*, 2-3 (2007), 185-199.
3. Bargh, J.A., Gollwitzer, P.M., Lee-Chai, A., Barndollar, K., and Trötschel, R. The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology 81*, 6 (2001), 1014-1027.
4. Barkhuus, L. and Rode, J. From mice to men: 24 years of evaluation in CHI. *ACM CHI '07 - Alt.CHI*, (2007).
5. van der Berg, M.H., Schoones, J.W., and Vliet Vlieland, T.P.M. Internet-based physical activity interventions: A systematic review of the literature. *Journal of Medical Internet Research 9*, 3 (2007).
6. Bickmore, T.W., Caruso, L., and Clough-Gorr, K. Acceptance and usability of a relational agent interface by older urban results. *Proceedings of CHI 2005*, ACM Press (2005).
7. Consolvo, S., Everitt, K.M., Smith, I., and Landay, J.A. Design requirements for technologies that encourage physical activity. *Proceedings of CHI 2006*, ACM Press (2006), 457-466.
8. Consolvo, S., Klasnja, P., McDonald, D.W., et al. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. *Proceedings of the 10th international conference on Ubiquitous computing*, ACM (2008), 54-63.
9. Consolvo, S., McDonald, D.W., Toscos, T., et al. Activity sensing in the wild: a field trial of ubifit garden. *Proceeding of CHI 2008*, ACM (2008), 1797-1806.
10. Csikszentmihalyi, M. and Larson, R. Validity and reliability of the Experience Sampling Method. *Journal of Nervous and Mental Disease 175*, 9 (1987), 526-536.
11. Deci, E.L. and Ryan, R.M. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory 11*, 4 (2000), 227-268.
12. Denning, T., Andrew, A., Chaudhri, R., et al. BALANCE: towards a usable pervasive wellness application with accurate activity inference. *Proceedings of the 10th workshop on Mobile Computing Systems & Applications*, ACM (2009), 1-6.
13. DeShazo, J., Harris, L., Turner, A., and Pratt, W. Grounded in theory, user-centered, and someplace else: Designing and remotely testing mobile diabetes video games. *Journal of Telemedicine and Telecare 16*, 7 (2010), 378-382.
14. Dixon, M. and Fogarty, J. Prefab: implementing advanced behaviors using pixel-based reverse engineering of interface structure. *Proceedings of CHI 2010*, ACM (2010), 1525-1534.
15. Fogg, B.J. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, New York, 2003.
16. Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B.L., and Landay, J.A. MyExperience: A system for in situ tracing and capturing of user feedback on mobile phones. *Proceedings of the 5th international conference on Mobile systems, applications and services (MobiSys '07)*, ACM Press (2007), 57-70.
17. Gasser, R., Brodbeck, D., Degen, M., Luthiger, J., Wyss, R., and Reichlin, S. Persuasiveness of a Mobile Lifestyle Coaching Application Using Social Facilitation. *Proceedings of the First International Conference on Persuasive Technology*, Springer-Verlag (2006), 27-38.
18. Greenberg, S. and Buxton, B. Usability evaluation considered harmful (some of the time). *Proceeding of CHI 2008*, ACM (2008), 111-120.
19. Higgins, E.T. Knowledge activation: Accessibility, applicability, and salience. In E.T. Higgins and A.W. Kruglanski, eds., *Social psychology: Handbook of*

basic principles. Guilford Press, New York, 1996, 133-168.
20. Hoffman, C., Rice, D., and Sung, H.Y. Persons with chronic conditions. Their prevalence and costs. *JAMA: The Journal of the American Medical Association 276*, 18 (1996), 1473-1479.
21. Hurling, R., Catt, M., De Boni, M., et al. Using internet and mobile phone technology to deliver an automated physical activity program: Randomized controlled trial. *Journal of Medical Internet Research 9*, (2007), e7.
22. Kopp, J. Self-monitoring: A literature review of research and practice. *Social Work Research & Abstracts 24*, (1988), 8-20.
23. Lee, G., Tsai, C., Griswold, W.G., Raab, F., and Patrick, K. PmEB: a mobile phone application for monitoring caloric balance. *CHI '06 extended abstracts on Human factors in computing systems*, ACM (2006), 1013-1018.
24. Lin, J.L., Mamykina, L., Lindtner, S., Delajoux, G., and Strub, H.B. Fish'n'Steps: Encouraging physical activity with an interactive computer game. *Proceedings of Ubicomp 2006*, Springer (2006), 261-278.
25. Logue, A.W. Self-control and health behavior. In W.K. Bickel and R.E. Vuchinich, eds., *Reframing health behavior change with behavioral economics*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2000, 167-192.
26. Mamykina, L., Mynatt, E., Davidson, P., and Greenblatt, D. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. *Proceeding of CHI 2008*, ACM (2008), 477-486.
27. Marcus, B.H., Dubbert, P.M., Forsyth, L.H., et al. Physical activity behavior change: issues in adoption and maintenance. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association 19*, 1 Suppl (2000), 32-41.
28. Marlatt, G.A. and George, W.H. Relapse prevention and the maintenance of optimal health. In S.A. Shumaker, E.B. Schron, J.K. Ockene and W.L. McBee, eds., *The handbook of health behavior change*. Springer, New York, 1998, 33-58.
29. Mattila, E., Pärkkä, J., Hermersdorf, M., et al. Mobile diary for wellness management--Results on usage and usability in two user studies. *IEEE Transactions on Information Technology in Biomedicine 12*, 4 (2008), 501-512.
30. Moore, S.M., Dolansky, M.A., Ruland, C.M., Pashkow, F.J., and Blackburn, G.G. Predictors of women's exercise maintenance after cardiac rehabilitation. *Journal of Cardiopulmonary Rehabilitation 23*, 1 (2003), 40-49.
31. Morris, M., Kathawala, Q., Leen, T.K., et al. Mobile therapy: Case study evaluations of a cell phone application for emotional self-awareness. *Journal of Medical Internet Research 12*, 2 (2010), e10.
32. Muraven, M. and Baumeister, R.F. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin 126*, 2 (2000), 247-259.
33. Muraven, M., Tice, D.M., and Baumeister, R.F. Self-control as limited resource: Regulatory depletion patterns. *Journal of Personality and Social Psychology 74*, 3 (1998), 774-789.
34. Nawyn, J., Intille, S.S., and Larson, K. Embedding behavior modification strategies into a consumer electronic device: a case study. *Proceedings of UbiComp 2006*, Springer (2006), 297–314.
35. Nelson, R.O. Assessment and therapeutic functions of self-monitoring. In M. Hersen, R.M. Eisler and P.M. Miller, eds., *Progress in behavior modification*. Academic Press, New York, 1977.
36. Prochaska, J.O., Johnson, S., and Lee, P. The transtheoretical model of behavior change. In S.A. Shumaker, E.B. Schron, J.K. Ockene and W.L. McBee, eds., *The handbook of health behavior change*. Springer, New York, 1998, 59-84.
37. Radtke, K.L. Exercise compliance in cardiac rehabilitation. *Rehabilitation Nursing: The Official Journal of the Association of Rehabilitation Nurses 14*, 4 (1989), 182-186, 195.
38. Rogers, Y., Connelly, K., Tedesco, L., et al. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *UbiComp 2007: Ubiquitous Computing*. 2007, 336-353.
39. Smith, B.K., Frost, J., Albayrak, M., and Sudhakar, R. Integrating glucometers and digital photography as experience capture tools to enhance patient understanding and communication of diabetes self-management practices. *Personal Ubiquitous Computing 11*, 4 (2007), 273-286.
40. Tsai, C.C., Lee, G., Raab, F., et al. Usability and feasability of PmEB: A mobile phone application for monitoring real time caloric balance. *Mobile Networks and Applications 12*, 2-3 (2007), 173-184.
41. Twardella, D., Merx, H., Hahmann, H., Wüsten, B., Rothenbacher, D., and Brenner, H. Long term adherence to dietary recommendations after inpatient rehabilitation: prospective follow up study of patients with coronary heart disease. *Heart (British Cardiac Society) 92*, 5 (2006), 635-640.
42. Witkiewitz, K. and Marlatt, G.A. Relapse prevention for alcohol and drug problems: that was Zen, this is Tao. *The American Psychologist 59*, 4 (2004), 224-235.