

# Task-based Embedded Assessment of Functional Abilities

Matthew L. Lee  
Human-Computer Interaction Institute  
Carnegie Mellon University  
matthew.lee@cs.cmu.edu

July 13, 2012  
(last revision July 1, 2012)

Thesis Dissertation

## Thesis Committee

Anind K. Dey (chair, HCII)  
Scott Hudson (HCII)  
Sara Kiesler (HCII)  
Judith Matthews (Univ. of Pittsburgh)  
Elizabeth Mynatt (Georgia Tech)

# Abstract

Many older adults desire to maintain their quality of life by living and aging independently in their own homes. However, it is difficult for older adults to notice and track the subtle changes in their own abilities because they can change gradually over long period of time. Technology in the form of ubiquitous sensors embedded in objects in the home can play a role in keeping track of the functional abilities of individuals unobtrusively, objectively, and continuously over a long period of time. This work introduces a sensing technique called embedded assessment of wellness that uses the everyday objects in the home that individuals interact with to monitor how well specific tasks important for independence are carried out. After formative studies on the information needs of older adults and their caregivers, a sensing system called dwellSense was designed, built, and evaluated that can monitor, assess, and provide feedback about how well individuals take their medications, use the phone, and make coffee. Multiple long-term (over 10 months) field deployments of dwellSense were used to investigate how the data collected from the system were used to support greater self-awareness of abilities and intentions to improve in task performance. Presenting and reflecting on data from ubiquitous sensing systems such as dwellSense is challenging because it is both highly dimensional as well as large in volume, particularly if it is collected over a long period of time. Thus, this work also investigates the time dimension of reflection and has identified that real-time feedback is particularly useful for supporting behavior change and longer-term trended feedback is useful for greater awareness of abilities. Traditional forms of assessing the functional abilities of individuals tend to be either biased, lacking ecological validity, infrequent, or expensive to conduct. An automated sensor-based approach for assessment is compared to traditional performance testing by a trained clinician and found to be match well with clinician-generated ratings that are objective, frequent, and ecologically valid. The contributions from this thesis not only advance the state of the art for maintaining quality of life and care for older adults but also provide the foundations for the class designing personal sensing systems that aim to assess an individual's abilities and support behaviors through feedback of objective, timely sensed information.

## Table of Contents

<b>1</b>	<b>Introduction.....</b>	<b>6</b>
1.1	Motivation.....	6
1.1.1	Accurate Functional Assessments.....	7
1.1.2	Overcoming the Barrier of Information Overload.....	8
1.2	Thesis Approach .....	9
1.2.1	Identifying Information Needs and Uses.....	10
1.2.2	Designing a Sensing System for Task-based Assessment of Wellness in the Home 10	
1.2.3	Supporting Awareness and Behaviors with Objective Sensor Information .....	11
1.2.4	Comparing Sensor-based Assessment of Wellness with Traditional Assessments..	12
1.3	Thesis Statement .....	13
1.4	Expected Contributions and Research Aims .....	13
1.5	References .....	14
<b>2</b>	<b>Background and Related Work.....</b>	<b>17</b>
2.1	Embedded Assessment.....	17
2.1.1	Smart Home Systems – Living Laboratories.....	17
2.1.2	Smart Home Systems – Deployments.....	19
2.1.3	Task Based Assessment .....	19
2.2	Evaluations of Stakeholder Needs.....	21
2.3	References .....	22
<b>3</b>	<b>Investigating the Information Needs and Potential Uses of Embedded Assessment</b>	<b>27</b>
3.1	Concept Validation Method.....	27
3.1.1	Participants.....	28
3.1.2	Interviews & Analysis.....	28
3.1.3	Concepts for Validation.....	28
3.1.4	Task Completion vs. Task Performance.....	30
3.1.5	Long-term vs. Short-term .....	31
3.1.6	Process Details.....	31
3.2	The Potential to Support Awareness of Functional Abilities.....	32
3.3	Usefulness of Task-based Embedded Assessment Data Features .....	33
3.3.1	Task Completion vs. Task Performance.....	33
3.3.2	Long-term vs. Short-term .....	33
3.3.3	Process Details.....	34
3.4	Limitations of Embedded Assessment Data.....	34
3.4.1	The Why is Missing.....	35
3.4.2	Searching for Significance .....	35
3.4.3	Noisy Data from the User, Not from the Sensors .....	36
3.5	Summary of Findings from Concept Validation .....	37
3.6	References .....	37
<b>4</b>	<b>dwellSense: A Task-based Embedded Assessment System.....</b>	<b>39</b>
4.1	dwellSense Sensing Capabilities.....	39
4.1.1	Medication Monitor.....	39

4.1.2	Telephone Tracker .....	40
4.1.3	Coffee Chronicler .....	41
4.1.4	Wireless networking .....	41
4.2	dwellSense Data Infrastructure .....	41
4.3	dwellSense Data Presentation .....	42
4.4	User Reflective Design Process .....	43
4.4.1	Sensing Unobtrusively .....	44
4.4.2	Identify Meaning by Engaging Users with their own Data .....	44
4.4.3	Supporting goals with data .....	45
4.4.4	The User Reflection Design Process in Action .....	45
4.5	Pilot Deployment .....	45
<b>5</b>	<b>Supporting Self-Reflection and Awareness of Functional Abilities .....</b>	<b>48</b>
5.1	Background in Reflection and Models of Behavior Change .....	48
5.2	Case Study Methodology .....	50
5.3	Sensor Deployment and Data .....	51
5.3.1	Smart Pillbox .....	51
5.3.2	Phone Sensor .....	52
5.3.3	Instrumented Coffeemaker .....	52
5.3.4	Deployment Data .....	53
5.4	Data Visualizations .....	53
5.5	Interacting with the Data .....	56
5.5.1	Looking for Anomalies/Mistakes .....	57
5.5.2	Generating Explanations .....	57
5.5.3	Confirming with Details .....	58
5.6	Attitudinal Reactions to the Data .....	58
5.6.1	Supporting Accurate Awareness .....	58
5.6.2	Intention to be More Consistent .....	59
5.6.3	Desire to Share Data and Potential for Misinterpretation .....	59
5.7	Behavioral Reactions to the Data .....	59
5.7.1	Analysis of Medication Behaviors .....	60
5.7.2	Analysis of Phone Use Behavior .....	65
5.8	Design Recommendations .....	66
5.9	Pilot Study Summary .....	67
5.10	References .....	68
<b>6</b>	<b>The Time Dimension of Reflection .....</b>	<b>70</b>
6.1	Supporting Explanations of Data Events .....	70
6.2	Background .....	71
6.2.1	Feedback .....	71
6.2.2	Goal Setting .....	72
6.2.3	Medication Taking and Adherence .....	73
6.3	Study Design .....	74
6.3.1	dwellSense version 2.0 .....	74
6.3.2	Participants .....	76
6.3.3	Deployment Timeline .....	77
6.4	Hypotheses .....	80
6.5	Measures .....	81
6.5.1	Measures of behavior .....	81
6.5.2	Measures of accuracy of self-awareness .....	81

6.5.3	Measures of self-reported subjective ability .....	82
<b>6.6</b>	<b>Results .....</b>	<b>82</b>
6.6.1	Reflection and Behavior Change .....	82
6.6.2	Accuracy of Self-Awareness .....	93
6.6.3	Self-Ratings of Abilities .....	97
6.6.4	Removing Real-Time Feedback and Behavior Change.....	100
<b>6.7</b>	<b>Discussion .....</b>	<b>105</b>
6.7.1	Benefits of Real-time Feedback.....	105
6.7.2	Benefits of Long-Term Reflection .....	105
<b>6.8</b>	<b>Summary .....</b>	<b>106</b>
<b>6.9</b>	<b>References .....</b>	<b>106</b>
<b>7</b>	<b>Automatic Assessment with Sensors .....</b>	<b>108</b>
7.1	Performance Testing .....	108
7.2	Research Questions.....	110
7.3	Data Collection.....	111
7.4	Automated Performance Testing using Sensors .....	112
7.4.1	Rule-based Assessment with Sensor Data.....	112
7.5	Results for Comparing Automatic Assessment with Performance Testing.....	116
7.5.1	Agreement in ratings of phone use .....	117
7.5.2	Agreement in ratings of medication taking .....	117
7.5.3	Agreement in ratings of coffee making.....	118
7.5.4	Summary of Agreement between Sensor-based Ratings and Performance Testing Ratings.....	119
7.6	Representativeness of Behaviors During Performance Testing .....	120
7.6.1	Differences in Medication Taking Actions .....	121
7.6.2	Differences in Coffee Making Actions.....	123
7.6.3	Summary of Representativeness of Evaluated Tasks .....	125
7.7	Comparing Strengths of Sensor-based Assessment and Performance Testing.....	126
7.7.1	Sensor-based assessments capture critical steps .....	126
7.7.2	Performance testing has a wider scope of evaluation .....	126
7.7.3	Sensor-based assessment can capture typical behaviors over time .....	127
7.7.4	Sensor-based assessment can capture more precise measures than performance testing .....	128
7.8	Summary .....	128
7.9	References .....	129
<b>8</b>	<b>Conclusion.....</b>	<b>131</b>
8.1	Support for Thesis.....	131
8.2	Contributions .....	133
8.2.1	Contributions to HCI and Design .....	133
8.2.2	Contributions to Computer Science.....	135
8.2.3	Contributions to Health Sciences.....	136
8.3	Limitations and Future Work.....	137
8.4	References .....	139

# 1 Introduction

Many older adults desire to maintain their quality of life by living and aging independently in their own homes. Successful aging requires an awareness of the subtle but cumulative changes in cognitive and physical abilities that older adults experience (Fried et al., 1991). With an accurate awareness of their abilities, older adults can make the necessary adjustments (such as setting routines, relying on cognitive or physical aids, or undergo medical treatments) that will support them as they age. However, it is difficult for older adults to notice and track the subtle changes in their own abilities because they can change gradually over long period of time. Changes in everyday cognitive and physical abilities usually manifest themselves in changes in functional ability, that is, how well individuals are able to carry out Instrumental Activities of Daily Living (IADLs) (Lawton & Brody 1969), everyday tasks such as meal preparation, managing medication, and using the telephone. Tasks like these are performed as a part of the individual's routines, and as a result, older adults may not pay close attention to how they perform these tasks and the subtle errors they may make. Thus, there exists an opportunity for technology to monitor how older adults (and indeed, all types of individuals) carry out routine tasks and assess the individual's abilities over a long period of time. In particular, technology in the form of ubiquitous sensors embedded in objects in the home, using a technique called embedded assessment (Morris et al., 2003), can play a role in keeping track of the functional abilities of older adults unobtrusively, objectively, and continuously over a long period of time. In fact, many previous research efforts have recognized the potential of a smart home to assist older adults as they age (Abowd et al., 2002; Helal et al., 2005).

However, these smart home sensors can generate an overwhelmingly large amount of data, particularly if they monitor multiple tasks during the long time span (often years) over which these functional changes occur. Furthermore, it is unclear how the data from these systems can directly influence the individuals being monitored to maintain their independence as well as their caregivers to make more informed decisions about how to provide care. The goals of this thesis therefore is to understand how to collect, analyze, present, and use information about how an individuals carry out everyday tasks over time to 1) assess their functional abilities and 2) increase their self-awareness of their abilities as well as help them maintain their ability to age in place.

In particular, this thesis will develop a task-based embedded assessment system, identify how data from embedded assessment systems improve upon existing measures (from self-report and performance testing), demonstrate how stakeholders (clinicians and older adults themselves) make use of this new source of continuous and objective data to assess functional abilities and health, and provide guidelines for how to analyze and present data in a way that leverages both the abilities of the user as well as the power of computational analysis.

## 1.1 Motivation

Assessing an individual's functional abilities is critical for understanding how well an individual is able to remain independent. However, it is often difficult to obtain an accurate assessment based on self-reported and caregiver-

reported abilities. The promise of smart homes that use embedded sensors is to provide detailed information about the behaviors of residents. Using sensors to collect this information by itself will not necessarily be useful, but the information must be summarized and presented in a way that highlights the most relevant details to assist in understanding the wellbeing of the resident.

### 1.1.1 Accurate Functional Assessments

Much of the prior work in embedded sensing in the home for eldercare focuses mainly on detecting safety-critical incidents such as falls or hazardous conditions such as leaving the stove on or a door unlocked. Indeed, these are important events to detect, as they provide opportunities to provide assistance to the individual in a dangerous situation. In these cases, technology is merely reactive, only capable of intervening after the individual has injured him/herself or is already in danger. However, sensing technology has reached a level of sophistication such that it can monitor the environmental or health conditions that lead up to accidents and thus can play an important role in preventing accidents before they occur. For example, non-adherence to medications is one common contributor for increasing the risk for falling (Woolcott et al., 2009). Tracking an individual's ability to manage and adhere to a medication routine can highlight when falls may be more likely and can also provide insight into the cognitive and physical limitations that the individual is experiencing.

Information collected passively with sensors about how individuals perform tasks can act as a trigger and provide new opportunities for earlier intervention to maintain independence by avoiding unsafe conditions. Maintaining adequate independence to live at home can reduce the substantial financial and emotional costs of institutionalization in an assisted living facility or nursing home (Anashensel et al., 2000). The information about everyday performance can provide earlier indicators useful for diagnosing conditions common among the elderly such as Alzheimer's disease or Parkinson's disease. Particularly with neurological conditions such as Alzheimer's and Parkinson's disease, earlier intervention has been shown in some cases to delay the progression of the more severe stages of the disease (Sitzer et al., 2006; Clare 2003; Loewenstein et al., 2004; Valenzuela & Sachdev 2009), improve psychological symptoms of the disease (Schneider-Beeri et al., 2002), provide caregivers with more time to adjust to provide adequate care (Schulz & Martire, 2004), and to reduce the financial burdens (Leifer 2003; Findley 2007; Langa et al., 2001). Thus, enabling earlier intervention for a degenerative condition using sensitive information triggers can have substantial impacts on society.

However, earlier intervention is only feasible if there exist early indicators or detectable risk factors for subsequent changes. There exists a stage before an individual becomes disabled (that is, formally diagnosed with a disabling condition) called "preclinical disability" (Fried et al., 1991) in which the individual experiences a decline in abilities but is able to use compensatory strategies to remain functional. For example, an individual who is beginning to have difficulty balancing when reaching items from a tall cabinet can brace herself against a wall to be more stable as a compensatory strategy. An example of compensating for cognitive decline is an individual who is beginning to have memory problems and has difficulty remembering to complete all the steps for making a pot of coffee; the individual may slow down, focus, and be more deliberate in her actions to compensate for declining cognitive abilities. Using these compensatory strategies, the outcome of the task can be maintained at an acceptable level by changing the process used to perform the task. The concept of preclinical disability has mostly been studied in the domain of physical disability. Similar findings with respect to cognitive disability show that there also exists a "prodromal" phase before a formal diagnosis of Alzheimer's disease in which there are detectable declines in cognitive and functional performance (Amieva et al., 2005).

Individuals often focus on the outcome, rather the process of their tasks as a measure of their abilities. For example, when making a pot of coffee, as long as the coffee tastes good, the individual would consider that as a success, despite the inefficiencies and mistakes they made. The compensatory strategies employed by individuals to maintain their level

of functionality can hinder their overall awareness of the fundamental changes in their abilities. As a result, many older adults are not aware of the cognitive, physical, and functional changes they experience as they get older. Self-reported sensory abilities like vision and hearing often are underestimated (Ott et al., 1996; Holland & Rabbitt 1992; Barrett et al., 2005). Likewise, self-reported cognitive abilities like executive functioning and memory also have been shown to be often inaccurate in both individuals with and without cognitive impairment (Graham et al., 2005). Self-reports of the functional abilities to carry out IADLs have also been shown to be mediated by cognitive reserve (Suchy, Kraybill & Franchow, 2010). Moreover, even if the individual is aware of a functional limitation, it might be dismissed as simply a normal part of growing older even though the consequences of the functional loss may be non-trivial (Lorenz 2009). For example, disruptions in sleep patterns due to chronic pain, in particular, are easily dismissed even though poor sleep can result in (at least temporarily) falls and impaired cognitive function.

In addition to self-reports, reports from caregivers such as relatives or friends, often serve as other sources of information to understand the well-being of an older adult. However, caregiver reports from friends, relatives, or neighbors can also be inaccurate, particularly with caregivers who may have infrequent contact with the individual. Like self-reports, caregiver reports can be biased either to report either more or less impairment (Okonkwo et al, 2009; Kemp et al., 2002). Patient self-reports and caregiver-reports have been found to differ, even in the context of patients with formal diagnoses of Mild Cognitive Impairment and Alzheimer's disease where impairments are more apparent (Ready, Ott, & Grace, 2004).

In the clinical setting, doctors and occupational therapists can use performance-based testing instruments (such as Diehl et al., 2005; Holm & Rogers, 1999; Owsley et al., 2002) by having patients perform tasks in the presence of a trained observer either in the clinic or at home. Clinicians often evaluate how well an individual carries out Instrumental Activities of Daily Living (IADLs), a standard battery of tasks important for maintaining a high level of independence, which includes taking medication, using the telephone, managing finances, shopping, preparing a meal, and using transportation. Each IADL can be broken down into individual steps. The observer's goal is to detect in which low-level steps of the IADL the patient is struggling and to provide appropriate interventions. However, these assessments are expensive to conduct, as they require a trained clinician (usually an occupational therapist) to administer them. Moreover, testing in the clinic forces the individual to perform in an artificial and often unfamiliar setting, which can cast doubt on the validity of the performance assessment data. Alternatively, the therapist can travel to the individual's home for direct observations in a setting more familiar for the individuals but again, this is costly in both time and money. Consequently, these assessments are performed infrequently and usually only after a problem has noticeably impacted everyday functioning. Performance effects can also bias the accuracy of the results, where patients may act differently during the one-time assessment from how they normally function in their everyday lives. Thus, individual and doctors need more frequent, less expensive, and more objective measures of an individual's functional ability to carry out Instrumental Activities of Daily Living.

### **1.1.2 Overcoming the Barrier of Information Overload**

Sensing technology has the potential to monitor behaviors in the home objectively, continuously, longitudinally (possibly even over several years) and with great detail. The volume of data can culminate into a vast and detailed lifelog. To make sense of all this information, sensing systems can rely on computer algorithms to process and interpret the data and find specific events such as falls or safety hazards. However, interpreting complex behaviors such as IADLs is likely to require not only computer analysis but also the interpretation of a human to make sense of the information and to identify the patterns that indicate wellbeing. Indeed, the collected information itself, rather than any action initiated by the system, can be used as an intervention to support a better awareness of the individual's functional abilities. Older adults can reflect on data about how well they performed tasks important for independence so they can make the appropriate adaptations to remain functional. The data can be shared with caregivers and clinicians to provide them with a better idea of how the individual is doing and provide better care.



Whereas computer systems may be good at interpreting and processing large amounts of information, older adults may have difficulties understanding complex data, particularly if they are experiencing age-related cognitive declines or are in the early stages of Alzheimer’s disease (Rizzo et al., 2000). The overwhelming amount of data combined with older adults’ cognitive limitations and unfamiliarity with sensing technology make it likely for such systems, if not well-designed, to overwhelm older adults with data, hinder adoption, and limit the insights into behavior. Thus, sensing systems that provide information-based interventions to support awareness must be designed so information is presented in a way that is compatible with the capabilities and needs of its users. Identifying the information needs and sensemaking processes of the various stakeholders is the first step in knowing how to present embedded assessment data to stakeholders. The lessons learned from identifying information needs and sensemaking processes can be used to inform the design of computational tools and analyses that aid the user’s interpretation of the sensor data.

In addition to self-reflection for an older adult, the behavioral data collected from the sensors has the potential to be a valuable data stream in the clinic for making better diagnoses. This potential, however, is moderated by the clinician’s ability to understand the data, interpret its significance, and find the relevant information for providing care to the patient. In order for the home sensor data to be integrated into the clinical workflow, it must be designed and presented in a way that highlights the most important, relevant, and salient details for the clinician. Health care trends in the United States show that the work demands placed on doctors are always increasing. The introduction of electronic health records (EHRs) makes for easy access to electronic data such as from home sensors, but along with EHRs come the potential for overloading the clinician with an overwhelming amount of patient information during visits (Berner & Moss, 2005). Studies have shown that the average time of a patient’s visit with a primary care physician in the United States has increased marginally from 1997 to 2005 by 16% (to an average of about 20.9 minutes), but more importantly, the number of topics and concerns discussed has increased 30% (to an average of 7.1 topics), leaving less time to devote to each topic (Abbo et al., 2008). Furthermore, behavioral data from the home can be a helpful source of information for screening tests and for practicing preventative medicine. However, a lack of time is a common reason that doctors do not practice preventative medicine (Kottke, Brekke, & Solberg, 1993). Introducing a new data stream without properly overcoming the challenge of information overload can be a burden that compounds the existing difficulties in devoting enough time for preventative medicine and providing care.

One of the main goals of this thesis is to understand how to design information systems that allow older adults, caregivers, and clinicians to understand, interpret, and use the large amount of data collected from home sensors while avoiding overloading them with more information than they need.

## 1.2 Thesis Approach

In order to understand how data from an embedded assessment system can be designed to be usable and useful, this thesis follows an approach that begins with understanding the information needs of older adults and their caregivers (Chapter 3). With these information needs and a good understanding of the homes and everyday lives of older adults as the target context, we introduce an approach called “embedded assessment of wellness” and develop a sensing system that assesses how well everyday tasks important for independence are performed (Chapter 4).

We use field deployments to collect real data and behaviors from real users to evaluate the effectiveness of the system for assessing functional abilities. The field deployment also acts as a platform to investigate how the data can be presented and used by older adults and their caregivers to increase self-awareness and help them achieve their goals of living independently. The first field deployment lasting 18 months serves as a pilot study with a small number of individuals (Chapter 5). With a case study approach for the pilot deployment, we evaluate the robustness of the sensing technology and as well as understand how people engage with data about their own behaviors.

Based on the results of the pilot field deployment, we identify opportunities to augment the system to support the ability of the individual to understand and use the data to maintain their independence. A larger deployment of the

sensing system (Chapter 6) for 10 months follows the pilot deployment to determine whether the results from the pilot deployment can be generalized to larger population. The larger deployment also allows us to present information to individuals in different ways and measure how different presentation methods affects awareness and behaviors.

During this larger field deployment, we use the structured questionnaires and functional assessments traditionally used in clinical studies to collect data about the cognitive, physical, and functional abilities of the individuals to compare with the automated assessments based on the home sensor data. We not only design the sensing and presentation layers of the system, but also develop methods to analyze an individual's task performance and calculate a score useful for clinical assessment. We compare these scores calculated from objectively sensed behaviors with the clinical measures also collected during the deployment (Chapter 7). We conclude a summary of contributions (Chapter 8).

### 1.2.1 Identifying Information Needs and Uses

The data collected from task-based sensing systems in the home is not only helpful for automated detection of anomalous events, but the information also can be useful for direct consumption by stakeholders (older adults, caregivers, and clinicians). However, the longitudinal task-based sensing approach can generate a large amount of data. As a first step in understanding how to make the large amount of embedded assessment data useful and usable for stakeholders, this thesis will identify the information needs of older adults, their caregivers, and their doctors with respect to the particular goal of feeling empowered to maintain their self-awareness and independence. Through a combination of formative user studies and evaluations with data collected from field deployments, this thesis will contribute an understanding of which tasks and behaviors stakeholders find helpful for measuring an older adult's functional abilities. Furthermore, it is unclear how stakeholders would use embedded assessment information if it were available, and thus this thesis also will investigate not only the information needs but also the usefulness of the embedded assessment data such as sharing with other stakeholders, maintaining awareness, making adaptations, or early diagnosis. This thesis will include a formative evaluation using scenario-based evaluation techniques that help stakeholders envision a reality in which the data are readily available will be used to answer the following research questions:

- RQ1 What are the information needs of stakeholders (older adults, family caregivers, and doctors)?
- RQ2 Is embedded assessment data potentially useful, and how? Would it support an awareness of abilities?

### 1.2.2 Designing a Sensing System for Task-based Assessment of Wellness in the Home

Approaches for applying sensing technology in the home generally fall into two categories: general activity monitoring and specific task monitoring. In general activity monitoring, easily deployed sensors such as motion detectors, video cameras, door sensors, microphones, wearable tags, and other environmental sensors capture gross movements and activities in the home. These systems can detect when an individual is moving around and can roughly estimate when they are engaging in behaviors in particular rooms (Demongeot, Virone, & Duchene, 2002). These systems also often aim to determine a baseline or "normal" pattern of activity and to find anomalies in the frequency or pattern of movements and activities in the home. To characterize a more specific activity, these systems can use machine learning to find particular sensor data patterns that correspond to particular activities performed in the home. This usually requires a fair amount of labeled ground truth data, which is often difficult to obtain from home settings.

The other approach for applying sensing technology, specific task monitoring, focuses on particular tasks that residents perform in the home, for example: sleep, appliance usage, walking in a predefined area, preparing a meal, or taking medication. Examples of sensing systems that focus on particular tasks include: a specialized bed sensor can detect restlessness and sleep patterns (Skubic et al., 2009), special load sensors embedded in the floor can detect the resident's gait (Helal *et al.*, 2005), a pressure mat that monitors whether the individual is in bed, a computer-vision system that

monitors hand washing for people with dementia (Mihailidis et al., 2007), and smart appliances that monitor a user's interactions (Helal et al., 2005).

In order to monitor the aspects of home life that may be most indicative of cognitive, physical, and functional decline, this thesis will introduce a sensing approach called “embedded assessment of wellness” that focuses on particular tasks, in particular, Instrumental Activities of Daily Living (IADLs) (Lawton & Brody, 1969). Performance on IADLs has been shown to be related to cognitive deficits (Tomaszweski et al., 2009; Cahn-Weiner et al., 2000). These tasks are commonly used in clinical practice to assess the functional abilities for patients (Diehl et al., 2005; Holm & Rogers, 1999; Owsley et al., 2002), and thus the sensor data about these tasks should also be representative of their functional abilities and may be more easily integrated into the clinical workflow.

Monitoring how often an individual performs IADLs can provide an indicator for a change in functional abilities because as individuals find the tasks more difficult or more dangerous given their abilities, they perform them less often. However, an even earlier or more sensitive indicator for declines in functional abilities is *how well* the task is performed (Owsley et al., 2002). Individuals are likely to make mistakes, slow down, or produce poorer task outcomes before they decide to decrease the frequency of the task or stop performing the task altogether. In addition to monitoring how often individuals perform IADLs, the sensing approach used in this thesis allows also attempts to monitor how well the individual performs the task, sometimes called *task adequacy* (Holm & Rogers, 1999) or *task performance*. For example, the individual may engage in taking their pills everyday and achieve an acceptable end goal (taking the correct pill at the right time), but the process used to achieve that end goal might vary considerably between episodes. The task process can be quantified as a measure of wellness and may include performance anomalies such as missing or mis-ordered steps, pauses that may indicate confusion or extra processing time, and recovered or unrecovered errors. These anomalies influence the process of the task and can provide good indicators for the cognitive, physical, and functional abilities of the individual. For example, an individual that manages to dial the phone correctly eventually to reach a pharmacy but has to make multiple attempts because of repeated misdialing is considered to be completing the task independently but be performing the task with a less than ideal level of adequacy or wellness. Unlike other sensing systems that focus only whether the individual completes a task or not, the sensing approach used in this thesis tracks task performance in addition to simply task completion. Thus, this thesis also addresses the following engineering and design question:

- RQ3 How do we build a sensing system that can assess how well individual's carry out Instrumental Activities of Daily Living in their own homes?

### 1.2.3 Supporting Awareness and Behaviors with Objective Sensor Information

The field deployments will involve embedding the sensing technologies in the homes of community-dwelling older adults for almost a year. The data collected and the informational interventions introduced during these deployments will provide an opportunity to identify how the data is used, what meaning individuals attach to the sensing data, and whether it supports their ability to be more self-aware of their functional abilities. In contrast to the formative evaluation study undertaken prior to the deployment where individuals described what they would do, the field deployment will allow us to answer the following research questions about actual actions:

- RQ4 How is embedded assessment data actually used? Does it support an awareness of abilities or change in behavior?

RQ4 is closely related to RQ2. RQ2 inquires about the *potential* uses of the embedded assessment data and is answered in this thesis via a formative scenario-based study, whereas RQ4 inquires about the *actual* uses of embedded assessment data and is addressed via a field study with real users interacting with their own data.

This thesis will also explore more specifically the time dimension of reflection to understand not only the impact of embedded assessment data but also the frequency and temporal pattern of reflecting on the data. Using a field study that compares the impact of near real-time feedback with feedback after longer periods, this thesis will address the following research question:

- RQ5 Does near real-time (when compared to delayed) feedback provide earlier opportunities for supporting a correct self-awareness of functional abilities?

#### 1.2.4 Comparing Sensor-based Assessment of Wellness with Traditional Assessments

The sensing approach used in this thesis also aims to address some of the drawbacks of existing methods of assessment [Table 1-1]. Self- and caregiver-reports of functional abilities can lack objectivity. Performance testing conducted in the lab produces objective data, but does not place contextually-appropriate demands on the individual and thus can lack ecological validity. Performance testing in the home can produce both objective and ecologically-valid data but can be expensive and, like self- and informant- reports, is typically performed infrequently and cannot identify new deficits in the period between evaluations. The sensors developed in this thesis aim to capture unobtrusive, objective, continuous, and ecologically-valid data. In particular, existing artifacts (*e.g.*, pillbox, coffee maker, telephone) commonly used by older adults are being augmented with sensing technology. The sensors are designed in such a way that they are minimally noticeable in the home and do not require the individual to change their routines, but are still capable of longitudinally and objectively collecting and interpreting information on the user’s task completion and task performance.

Features of Assessment Measures	Self Report and Informant Report	Performance Testing	Embedded Task-based Sensing in the Home
<b>Unobtrusive</b>	Yes Individuals do not have to “do” anything.	No Requires a special visit with a special observer.	Yes Individuals simply interact with the existing objects in their homes.
<b>Objective</b>	No Individuals and caregivers can be biased because of poor insight or relationships.	Yes Standardized tasks are used with an unbiased third-party assessor.	Yes Sensors passively record objective, quantitative data about task performance.
<b>Timely</b>	No Individuals and informants are not asked very often to assess their abilities, and when they are, it often is too late.	No Performed infrequently because it is expensive.	Yes Sensors installed in the home before any signs of decline can provide the early signs of changes in performance.
<b>Ecologically-valid</b>	Yes Reports are based on behaviors observed in daily life.	Yes, but often No Sometimes performed in the individual’s home with actual tasks, but the individual may exhibit a testing effect and perform differently than normal.	Yes Sensor-based assessments are based on behaviors observed in daily life.
<b>Detailed</b>	No Individuals usually report on general performance rather than on the specific errors in task	Yes The trained assessor observes and analyzes how well each step of a task is performed.	Yes Sensors can give precise timings and sequencing of steps in a task.

	performance.		
--	--------------	--	--

**Table 1-1. How different types of assessment methods differ in terms of desirable features for assessment measures. This thesis explores whether task-based sensing for embedded assessment have these features.**

To better understand how to implement task-based home sensing and its benefits over conventional assessment methods, this thesis aims to address the following research questions:

- RQ6 Can automatic sensor-based assessment match the ratings of task adequacy from traditional performance testing?
- RQ7 What aspects of task performance is sensor-based assessment better suited for? What aspects of task performance is performance testing better suited for?

This task-based sensing system will be deployed in the homes of older adults who are living on their own in their own homes and data about how often and how well they perform IADLs will be collected longitudinally over a period of 10 months. During the monitoring period, older adults will perform their everyday activities as they normally would and produce real, organic data about their own functional ability to carry out IADLs. A trained occupational therapist will visit the individuals at various times during the monitoring period and administer performance testing. Ratings from the therapist's assessments will be compared with the ratings generated from the automatic sensor-based system in order to identify the relative strengths and weaknesses of sensor-based assessment and traditional performance testing.

### 1.3 Thesis Statement

This thesis will prove the following statement:

Embedded assessment of wellness can provide ecologically valid assessments of task performance and reflecting on the generated data supports new opportunities for timely assessment of functional abilities for older adults.

### 1.4 Expected Contributions and Research Aims

In this thesis, I make the following contributions in the fields of Human-Computer Interaction & Design, Computer Science/Engineering, and Health Theory & Practice:

- HCI / Design
  - Identified the information needs of older adults, their caregivers, and clinicians for understanding functional changes associated with aging.
  - Described the sense-making process that people use to understand their own behaviors based on sensor data and identified breakdowns in this process as opportunities for computational support.
  - Demonstrated that reflecting on embedded assessment data leads to greater self-awareness and provides opportunities to make changes necessary to remain independent.
  - Demonstrated that real-time and long-term presentations of data can have different effects on self-awareness and behavior change.
  - "User Reflective Design Framework" that leverages human insights for designing intelligent personal sensing systems
- Computer Science / Engineering

- Designed, built, and evaluated a task-based sensing system comprising a suite of intelligent sensors that monitor the key steps in common Instrumental Activities of Daily Living.
- Demonstrated that ratings from a sensing system that uses rule-based assessment can match the ratings by a trained clinician.
- Health Theory and Practice
  - Demonstrated how real-time feedback can support individuals in carrying out Instrumental Activities of Daily Living with greater adequacy.
  - Developed a system that can sense how an individual behaves typically at home in the absence of a human observer.
  - Identified and quantified the testing effect associated with in-home performance testing by showing how tested behaviors differ from typical behaviors.

## 1.5 References

- Abbo, E.D., Zhang, Q., Zelder, M., and Huang, E.S. The increasing number of clinical items addressed during the time of adult primary care visits. *Journal of General Internal Medicine* 23, 12 (2008), 2058–2065.
- Abowd, G. D., Bobick, A. F., Essa, I. A., Mynatt, E. D., & Rogers, W. A. (2002). The aware home: A living laboratory for technologies for successful aging. *Proceedings of the AAAI-02 Workshop "Automation as Caregiver"* (pp. 1–7). Retrieved from <https://www.aaai.org/Papers/Workshops/2002/WS-02-02/WS02-02-001.pdf>
- Amieva, H., Jacqmin-Gadda, H., Orgogozo, J.M., et al. The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* 128, 5 (2005), 1093.
- Aneshensel, C.S., Pearlin, L.I., Levy-Storms, L., and Schuler, R.H. The transition from home to nursing home mortality among people with dementia. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 55, 3 (2000), S152.
- Barrett, A.M., Eslinger, P.J., Ballentine, N.H., and Heilman, K.M. Unawareness of cognitive deficit (cognitive anosognosia) in probable AD and control subjects. *Neurology* 64, 4 (2005), 693.
- Cahn-Weiner, D.A., Malloy, P.F., Boyle, P.A., Marran, M., and Salloway, S. Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. *The Clinical Neuropsychologist* 14, 2 (2000), 187–195.
- Clare, L. Cognitive training and cognitive rehabilitation for people with early-stage dementia. *Reviews in Clinical Gerontology* 13, 01 (2003), 75–83.
- Demongeot, J., Virone, G., Duchêne, F., et al. Multi-sensors acquisition, data fusion, knowledge mining and alarm triggering in health smart homes for elderly people. *Comptes Rendus Biologies* 325, 6 (2002), 673–682.
- Diehl, M., Marsiske, M., Horgas, A.L., Rosenberg, A., Saczynski, J.S., and Willis, S.L. The revised observed tasks of daily living: A performance-based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology* 24, 3 (2005), 211.
- Findley, L.J. The economic impact of Parkinson's disease. *Parkinsonism & Related Disorders* 13, (2007), S8–S12.
- Fried, L.P., Herdman, S.J., Kuhn, K.E., Rubin, G., and Turano, K. Preclinical disability: hypotheses about the bottom of the iceberg. *Journal of Aging and Health* 3, 2 (1991), 285.
- Graham, D.P., Kunik, M.E., Doody, R., and Snow, A.L. Self-reported awareness of performance in dementia. *Cognitive Brain Research* 25, 1 (2005), 144–152.
- Helal, S., Mann, W., El-Zabadani, H., King, J., Kaddoura, Y., and Jansen, E. The gator tech smart house: A programmable pervasive space. *Computer*, (2005), 50–60.
- Holland, C.A. and Rabbitt, P. People's awareness of their age-related sensory and cognitive deficits and the implications for road safety. *Applied Cognitive Psychology* 6, 3 (1992), 217–231.
- Holm, M. and Rogers, J. Performance assessment of self-care skills. *Assessment in occupational therapy mental health:*

- an integrative approach. In *Assessments in occupational therapy mental health: an integrative approach*. B. Hemphill-Pearson, Thorofare, NJ, 1999, 117-124.  
<http://www.elitecare.com>.
- Kemp, N.M., Brodaty, H., Pond, D., and Luscombe, G. Diagnosing dementia in primary care: the accuracy of informant reports. *Alzheimer Disease & Associated Disorders* 16, 3 (2002), 171.
- Kottke, T.E., Brekke, M.L., and Solberg, L.I. Making "time" for preventive services. *Mayo Clinic proceedings*. Mayo Clinic, (1993), 785.
- Langa, K.M., Chernew, M.E., Kabeto, M.U., et al. National Estimates of the Quantity and Cost of Informal Caregiving for the Elderly with Dementia\*. *Journal of General Internal Medicine* 16, 11 (2001), 770-778.
- Lawton, M.P. and Brody, E.M. Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist* 9, 3 Part 1 (1969), 179.
- Leifer, B.P. Early diagnosis of Alzheimer's disease: clinical and economic benefits. *Journal of the American Geriatrics Society* 51, 5s2 (2003), S281-S288.
- Loewenstein, D.A., Acevedo, A., Czaja, S.J., and Duara, R. Cognitive rehabilitation of mildly impaired Alzheimer disease patients on cholinesterase inhibitors. *American Journal of Geriatric Psych* 12, 4 (2004), 395.
- Lorenz, A. Indicators of preclinical disability: Women's experiences of an aging body. *Journal of women & aging* 21, 2 (2009), 138-151.
- Mihailidis, A., Boger, J., Canido, M., and Hoey, J. The use of an intelligent prompting system for people with dementia. *interactions* 14, 4 (2007), 34-37.
- Morris, M., Lundell, J., Dishman, E., and Needham, B. New perspectives on ubiquitous computing from ethnographic study of elders with cognitive decline. *Ubicomp 2003: Ubiquitous Computing*, (2003), 227-242.
- Okonkwo, O.C., Griffith, H.R., Vance, D.E., Marson, D.C., Ball, K.K., and Wadley, V.G. Awareness of Functional Difficulties in Mild Cognitive Impairment: A Multidomain Assessment Approach. *Journal of the American Geriatrics Society* 57, 6 (2009), 978-984.
- Ott, B.R., Lafleche, G., Whelihan, W.M., Buongiorno, G.W., Albert, M.S., and Fogel, B.S. Impaired awareness of deficits in Alzheimer disease. *Alzheimer Disease & Associated Disorders* 10, 2 (1996), 68.
- Owsley, C., Sloane, M., McGwin, G., and Ball, K. Timed instrumental activities of daily living tasks: Relationship to cognitive function and everyday performance assessments in older adults. *Gerontology* 48, (2002), 254-265.
- Ready, R.E., Ott, B.R., and Grace, J. Patient versus informant perspectives of Quality of Life in Mild Cognitive Impairment and Alzheimer's disease. *International journal of geriatric psychiatry* 19, 3 (2004), 256-265.
- Rizzo, M., Anderson, S.W., Dawson, J., Myers, R., and Ball, K. Visual attention impairments in Alzheimer's disease. *Neurology* 54, 10 (2000), 1954.
- Schnaider Beerli, M., Werner, P., Davidson, M., and Noy, S. The cost of behavioral and psychological symptoms of dementia (BPSD) in community dwelling Alzheimer's disease patients. *International journal of geriatric psychiatry* 17, 5 (2002), 403-408.
- Schulz, R. and Martire, L.M. Family caregiving of persons with dementia: prevalence, health effects, and support strategies. *American Journal of Geriatric Psych* 12, 3 (2004), 240.
- Sitzer, D.I., Twamley, E.W., and Jeste, D.V. Cognitive training in Alzheimer's disease: a meta-analysis of the literature. *Acta Psychiatrica Scandinavica* 114, 2 (2006), 75-90.
- Skubic, M., Alexander, G., Popescu, M., Rantz, M., and Keller, J. A smart home application to eldercare: Current status and lessons learned. *Technology and Health Care* 17, 3 (2009), 183-201.
- Suchy, Y., Kraybill, M.L., and Franchow, E. Instrumental activities of daily living among community-dwelling older adults: Discrepancies between self-report and performance are mediated by cognitive reserve. *Journal of Clinical and Experimental Neuropsychology* Jul, 1 (2010), 1-9.
- Tomaszewski Farias, S., Cahn-Weiner, D.A., Harvey, D.J., et al. Longitudinal changes in memory and executive functioning are associated with longitudinal change in instrumental activities of daily living in older adults. *The Clinical Neuropsychologist* 23, 3 (2009), 446-461.
- Valenzuela, M. and Sachdev, P. Can cognitive exercise prevent the onset of dementia? Systematic review of randomized clinical trials with longitudinal follow-up. *American Journal of Geriatric Psych* 17, 3 (2009), 179.
- Woolcott, J. C., Richardson, K. J., Wiens, M. O., Patel, B., Marin, J., Khan, K. M., & Marra, C. A. (2009). Meta-analysis of the Impact of 9 Medication Classes on Falls in Elderly Persons. *Archives of Internal Medicine*, 169(21), 1952-1960. doi:10.1001/archinternmed.2009.357
- Yarnall, K.S., Pollak, K.I., Ostbye, T., Krause, K.M., and Michener, J.L. Primary care: is there enough time for

prevention? *American Journal of Public Health* 93, 4 (2003), 635.



# 2 Background and Related Work

## 2.1 Embedded Assessment

Embedded assessment, the concept of using embedded sensors in the home to monitor the functional abilities of older adults, was introduced by Morris *et al.* [2][39]. Dishman [40] also envisioned sensing systems that continuously collect data on functional abilities to promote healthy behaviors, detect diseases earlier by finding disease signatures, and facilitate informal caregiving. They envisioned systems that could automatically collect data to assess wellbeing, detect disease earlier, and facilitate informal caregiving. Embedded assessment systems were envisioned to include three components: monitoring, compensation, and prevention. This thesis focuses on the first component of monitoring because it is a necessary for enabling the second and third components, compensation and prevention. This thesis also investigates how the data collected from the monitoring phase can be used for compensating for decline and also for developing preventative strategies to proactively maintain independence. Morris *et al.* [39] propose that monitoring the amount of external assistance needed to complete a task can be a measure of ability and overall health. The approach used in this thesis takes a slightly more ambitious approach to monitor the small errors in task performance that occur earlier *before* external assistance is required. This approach follows the guideline suggested by Morris *et al.* [2] that embedded assessment technology would be most effective if it is used before the onset of the disease or disability. Morris *et al.* also suggested that customizing the sensing and presentation of the sensor data to each individual, what they call “extreme personalization” is important for ensuring that the data has significance. The sensing approach we use in this thesis follows a similar technique for customizing the sensors for each individual’s method of carrying out particular IADLs. Furthermore, Morris *et al.* highlight the need to provide direct value to the individual who is monitored as one of the main barriers to adoption. However, embedded sensing can often provide only indirect value to the monitored individual by sharing the information with caregivers and clinicians. Thus this thesis will evaluate the value that older adults can directly receive from using embedded assessment data for self-reflection and self-awareness of functional abilities. One of the main challenges of embedded assessment is in understanding how to support individuals who want to manage their own health with the data collected from these systems. In the next section, a brief survey is presented of relevant embedded systems that perform functional assessments of an individual’s ability to live independently.

### 2.1.1 Smart Home Systems – Living Laboratories

The concept of a smart home has been part of a ubiquitous computing vision ever since Mark Weiser’s vision [41]. A number of research groups around the world have explored the potential of a smart home and the sensors that make a home intelligent by building laboratories where new types of sensing technologies can be developed and tested in a relatively controlled environment. Many smart home projects focus on monitoring of physiological parameters and environmental conditions to provide assistance in the form of automation. In this section, we discuss a selection of the relevant smart home projects that focus on monitoring the health and wellness of residents. For a broader survey of smart home projects, both in the United States and abroad, see [42][43].

Many smart home living laboratories contain technology to monitor when residents are performing various activities around the home and when unsafe conditions such as stove left on or a fall might have occurred. For example, the GatorTech Smart House [35] from the University of Florida was designed to monitor the general safety and operational conditions of the home and provide warnings and automated assistance when necessary. One aspect of the

house is to track how individuals interact with various appliances around the home such as the washing machine, stove, and microwave and to provide guidance with its operation for those who have difficulty with them. The GatorTech Smart House can also track the mobility of the residents using a smart floor and ultrasonic beacons. It also tracks sleep patterns using a specialized bed sensor. Similar to the GatorTech Smart House, the AwareHome [44] at Georgia Tech also tracks the movements and gait of the residents using a smart floor. The AwareHome also helps residents find lost objects with the help of RFID tags and also provides assistance with completing tasks such as preparing a meal [45]. Specialized sensing on the electrical system of the home can also provide information about what objects residents are interacting with. The AwareHome has also served as a testbed for various applications aimed to connect residents (presumably older adults) with remote caregivers [46], to capture and access an audio buffer [47], and to characterize overall activity or mood in the home [44]. The Ubiquitous Home project in Japan has instrumented a real apartment with cameras to track activities and movement as well as special vibration sensors in the floor to track how the resident is walking. Residents carry with them an RFID tag to allow the system to track their location in the apartment and provide context-dependent services and assistance. At the NTT DoCoMo lab [48], RFID tags are placed on objects and carried by the resident so that the resident's interactions with objects can be reconstructed at any given time to recognize their behavior according to activity models. Intel Research has been looking at using techniques using RFID tags, video, and common sense knowledge to bootstrap activity recognition in noisy, less structured real-world environments [49][50][51][52].

Living laboratories not only can be environments to develop and refine new sensing technologies but they can also be used to collect short-term data on the activities of a temporary resident. The PlaceLab [53] at MIT instrumented an apartment with various sensors and had a 30 year-old and 80 year-old individual each live in the apartment for 14 days. Simple state change sensors were placed on cabinets, doors, and objects around the home to recognize different the different activities of the residents. Their sensing approach was intended to be general-purpose, relying on first collecting information from many objects and spaces in the home and then using supervised machine learning to identify and recognize when individuals are performing different activities. To collect class labels for the data, the PlaceLab project used experience sampling. However, even with experience sampling, it was difficult to generate enough labeled data to recognize the fine-grained activities. The classification algorithms used Bayesian models and were able to recognize basic ADLs such as bathing, toileting, grooming, and preparing lunch with greatest confidence. The CASAS project [54] at Washington State also uses machine learning to classify and recognize activities based on sensors placed in a three-bedroom on-campus apartment to monitor the state of various appliances such as water usage, stove usage, and power consumption. Contact sensors on other objects such as the phone book, cooking pot, and medicine container also help contribute to providing information for activity recognition and classification. Based on data from a student who lived in the apartment for one month, the CASAS project was able to classify activities such as cooking, watching television, grooming, sleep, night wandering, and taking medications [55][56]. They were also able to use unsupervised learning to discover what activities individuals engaged in most frequently and see when the frequency of these activities changed [57]. Both the CASAS and PlaceLab projects use a "dense" sensing approach where sensors are scattered in the environment and activities are dynamically recognized with machine learning. The approach of this thesis uses an intentionally more constrained sensing approach that relies less on machine learning and more on simpler heuristics applied to particular tasks common across many individuals. With an understanding of existing task routines and sensors that can detect object manipulations at each fine-grained step, a heuristic-based model can be generated and used not only for recognizing the activities and tracking their frequency or pattern but also for evaluating task performance, that is, how well individuals carry out a particular task. Before we discuss related work on systems that focus on evaluating task performance, we will first describe some of the field deployments of embedded assessment technology that collects real data from real people.

### 2.1.2 Smart Home Systems – Deployments

Smart home technologies usually begin in the incubating environment of the living laboratory. Evaluations of these technologies require an individual to live in these labs to produce test data. Typically only short-term data is collected from the lab, and thus smart home projects often migrate their technologies out of the lab and into deployments out in the homes of real individuals. With real data from real individuals, researchers can test the robustness of their sensing systems and to verify whether they are capable of handling noisy real world behavior. Researchers can also explore whether and how embedded assessment data can be predictive (or at least retrospectively predictive) of changes in health.

TigerPlace at the University of Missouri-Columbia is a specialized independent-living facility that has been instrumented with various sensors to monitor the wellbeing of its 34 elderly residents. Residents range from 70 to 90 years old, 90% of whom have a chronic illness. The suite of sensors, called the In-Home Monitoring System, include motion sensors, a temperature sensor for the stove, a bed sensor that can track restlessness, and a privacy-preserving video system that monitors for falls. Case studies of the data collected over approximately two years at TigerPlace show that the certain behaviors captured by the sensors (such as bed restlessness) change near a health event such as having surgery [58]. Furthermore, changes in the overall activity and mobility of the resident as measured by motion sensors have been associated with health events [33].

An early adopter of smart home technologies is EliteCare at Oatfield Estates, a continuing care facility in Oregon [32]. The locations of residents are tracked using wireless beacons, and their sleep patterns are tracked with load cells on their beds. The information is shared with family members and health care providers through an internet portal. EliteCare has partnered with the Orcatech group at Oregon Health and Science University (OHSU) as a testbed site. Based on data from EliteCare residents, the Orcatech group found that bed load cells can be useful for detecting sleep patterns. The Orcatech group has also instrumented the homes of fourteen community-dwelling older adults with door sensors to monitor individuals entering and exiting rooms and motion sensors to track movement and walking speed. With this combination of sensors capturing data for at least six months, researchers were able to track the overall activity of healthy individuals and individuals with a diagnosis of Mild Cognitive Impairment, a precursor to Alzheimer's disease. They found that the overall activities and walking speed of the individuals with Mild Cognitive Impairment were more variable than cognitively healthy individuals [59]. Researchers at OHSU have also been using data from field deployments of embedded assessment technology to investigate how to establish a baseline pattern of activity performance and to find anomalies in the individual's routines. Considering data on bedtime, wake time, and sleep duration, they were able to determine both acute and gradual changes in the sleep routines [60]. The Orcatech group also has other ongoing deployments including a nine-home deployment with their standard suite of door and motion sensors to track overall activity in the home [61].

The field deployments undertaken in this thesis will follow a similar longitudinal approach but focus on not only how often and when the tasks for independence are completed but also will measure how well the tasks are performed, in an attempt to find earlier indicators of changes in functional, physical, or cognitive health.

### 2.1.3 Task Based Assessment

The ability to carry out everyday tasks such as Instrumental Activities of Daily Living (IADLs) [36] is important for maintaining independence. IADLs such as preparing a meal, using the telephone, taking medications, managing finances, doing laundry, and taking transportation are performed fairly frequently and require a high level of cognitive and functional ability to perform. Thus assessing how well older adults carry out these tasks can be a good indicator for any changes in their abilities. Traditional methods of assessment are typically based on standardized questionnaires [62][63] for older adults and their caregivers/informants to report their functional abilities. These self- and informant-reported can be biased and thus inaccurate [25][21]. Standardized performance testing in the clinic or home is usually

administered when more detailed information about the individual's functional abilities are required. Standardized performance tests [64][65][66][67] use a range of techniques from very standardized setups in the laboratory to open-ended activity analyses (often used by occupational therapists) to observe and test individuals in their own homes.

One performance test, the Performance Assessment of Self-Care Skills (PASS) test evaluates how well an individual is able to carry out tasks such as preparing a meal, paying with a check, balancing a checkbook, using the telephone, using household tools, obtaining information from the media, playing bingo, and using the stove. The PASS evaluates task performance along three dimensions: safety, independence, and adequacy. An individual performs a task safely if they do not place themselves in a dangerous situation while performing the task. An individual performs a task independently if they do not require external assistance to complete the task. A task is performed adequately if the task process and outcome are acceptable for the given task. Similarly, Gill *et al.* [68] define two components of disability for older community-dwelling adults: dependence and difficulty. An individual may be able to complete a task independently but experience great difficulty during the task process. The constructs of task difficulty and adequacy can provide a framework for understanding how task performance, in addition to task completion or frequency, can provide sensitive measures of functional abilities for older adults, particularly if assessments of task adequacy can be done frequently, objectively, and inexpensively in a naturalistic setting of the home.

Embedded assessment systems in the home can play an important role in the assessing the task performance of individuals frequently, objectively, and inexpensively. Even though smart home systems tend to follow an approach that recognizes high-level activities, some systems have been designed to monitor how well individuals carry out specific tasks. Specific task assessment is often coupled with specific task assistance. For example, Mihailidis *et al.* [34] have developed a system that monitors how an individual with dementia carries out the task of washing their hands. The system uses computer vision to detect when the individual is (or is not) carrying out a particular step such as turning on the water or using the soap and can provide specific assistance to the individual as to which step to perform next. The system provides more information than simply whether the individual has completed or not completed the hand washing task successfully. Whereas the main goal of the application is to provide assistance for hand washing, monitoring of the task process can provide valuable information for assessing how well the individual is performing the task with and without prompts.

In addition to Basic ADLs such as hand washing, other research has focused on more complex tasks such as using a coffee maker. Researchers at the University of Michigan investigated whether measures of performance during the multi-step task of making coffee with a coffee maker was correlated with cognitive abilities. In a study involving multiple individuals of varying cognitive abilities following instructions to use a coffee maker, Hodges *et al.* [69] found that task performance measure such as edit distance (a mathematic measure of how far the individual deviates from an ideal path for completing the task) is correlated with standardized measures of general neuropsychological integrity. That is, individuals with more compromised cognitive abilities tended to make more mistakes and take extra steps to complete the task. Other task performance measures such as task duration, action gaps, and object misuse also had suggestive correlations with other psychological factors. Hodges *et al.* [70] used machine learning to explore how combinations of factors or measures of task performance can distinguish between healthy and cognitively impaired individuals.

Anomaly detection is another method for assessing the quality, adequacy, or difficulty of a task. Cook & Schmitter-Edgecombe [71] at Washington State University found that by applying hidden Markov models to sensor data about task performance conducted under controlled (by introducing specific errors) and uncontrolled (by introducing naturalistic errors at non-prescribed times) laboratory settings, they were able to identify changes in consistency and identify errors in the task performance for predefined subset of tasks such as preparing a meal, telephone use, hand washing, eating and medication use, and cleaning around the home. The setup in the instrumented apartment includes motion and temperature sensors as well as contact sensors that detect whether or not the individual is interacting with

the stove, cooking pot, phone, phone book, sink, and medicine box. The Markov models are used to calculate how different a given task performance is from a model of “normal” or expected task performance. Normal performance was modeled after 20 undergraduate participants who performed the task without errors. The non-normal performance data was generated an additional 20 participants who inserted either errors specified by the researchers or a non-specific error similar to what a person with dementia would do somewhere in the task process. Sequences of events that were sufficiently different from the normal sequence of events as modeled by the Markov model were considered anomalous or inconsistent. Also they found that the length of time it takes to complete a task to be indicative of the presence of an error in the task.

Thus prior work has found potential in task-based embedded assessment to measure the quality and adequacy of task performance. In addition to providing a way to quantify functional abilities, data from embedded assessment systems also has the potential to provide direct value to older adults, their caregivers, and doctors. In the follow section, we describe prior attempts to explore the information needs of stakeholders and how embedded assessment may be able to meet those needs.

## 2.2 Evaluations of Stakeholder Needs

The stakeholders that can benefit from embedded assessment systems include the older adults who are being monitored, their family and professional caregivers who look after them, and their health care providers. Prior research has investigated the factors that influence adoption and acceptance of long-term monitoring technology.

A common theme among formative evaluations of smart home health monitoring technologies is that they are necessary only when health changes are already apparent. Kang *et al.* [72] discuss the potential benefits of in-situ monitoring in the home, which includes detecting adverse events, providing information for better diagnoses of conditions, and capturing the dynamic nature of the progression of a disease. Kang *et al.* discuss that the largest barriers to adoption include user friendliness, the possibility of reducing human contact, and the specialized training necessary to learn to use a new type technology. They also highlight the need for technology to be employed before it may be deemed necessary in order to ensure safety and even prevent disability. Thus, embedded assessment systems must not unduly stigmatize users as disabled when they are not. Kang *et al.* call for more active participation from health care providers and older adults in designing embedded assessment systems that will help them meet their needs. Similar to findings from Kang *et al.*, Kentta *et al.* [73] found that acceptance of services for independent living were mediated by their credibility, usefulness, ethicality, ease of use, and desirability. They used a scenario-based evaluation method to explore how stakeholders viewed different types of services for older adults. Home health monitoring was one of the services evaluated and it was considered a useful service but was only considered necessary when there was a clear threat to health and independence. Health care professionals found scenarios related to ensuring safety and tele-rehabilitation as most useful. Likewise, Courtney *et al.* [74] also found that an important factor in the willingness of residents at an assisted living facility to accept smart home technology is not the concern for privacy but rather their self-perception of need for the technology. Some of the sub-factors that contribute to a self-perception of need include their self-perception of their health, physical condition, mental and emotional condition, influence of family and friends, influence of health care professionals, the physical environment, the type of technology, and the perceived redundancy of the technology. They found that the individuals that most need home health-monitoring technology (because they are not aware of their own health changes) are those individuals that are least likely to adopt it. Like Kang *et al.*, Courtney *et al.* recommend that primary care providers play a more active role in encouraging individuals to adopt embedded assessment technology to ensure their long-term health. Beach *et al.* [75] also investigated the privacy tradeoffs with home monitoring technology. They found that individuals were most concerned with monitoring sensitive or personal activities such as toileting and sharing personal information with the government or insurance companies. They also found that individuals currently with disability were most accepting of monitoring technology and sharing that information with other stakeholders. Beaudin *et al.* [76] also performed formative evaluations to

investigate which health domains individuals wanted to track. Using a number of displays that showed hypothetical data about weight, chronic conditions, headaches, activity around the home (such as watching television), and nutrition, they found that there was general interest in collecting information for personalized, longitudinal collection and self-investigation of health.

Demiris *et al.* [77] conducted focus groups with clinical and non-clinical stakeholders to evaluate different smart home technologies as formative research for University of Missouri-Columbia's TigerPlace facility. They found that non-intrusive, user friendly, accurate, reliable and inconspicuous sensing such as pressure pads and infrared motion sensors were acceptable for sensing a variety of activities in the home, when compared with more invasive sensors such as iris recognition. Similar to previous studies, stakeholders expressed a requirement for technology to avoid generating new hazards, place burden on residents, limit the range of acceptable activities, increase anxiety, or promote stigmas. In another set of focus groups, Demiris *et al.* [78] investigated attitudes about adopting sensor-based smart home technologies. Similar to the previously mentioned research, they found that stakeholders expressed that monitoring technologies are better for those who are more frail and may need the assistance. Perceived need and frailty were the two factors that influenced acceptance the most. Discussions in the focus group also centered around home monitoring technologies acting as assistive technologies that helped during an emergency situation rather than as sources of health information useful for preventative health. Stakeholders valued technologies used as safety monitors such as keeping track of whether the stove is left on and motion sensors to track when an individual may have fallen. Simple safety-monitoring technologies provided stakeholders with peace of mind. As part of the TigerPlace project, Demiris *et al.* [79] also conducted evaluations of the technology with actual users of smart home technologies. Study participants who had the In Home Monitoring system installed (motion detectors, stove sensor, contact/door sensors, and video for monitoring falls) were involved in participatory design sessions to encourage them to discuss their feedback on the usefulness of sensor technologies. From these sessions, they discovered the three-phase process by which residents adjusted and adopted the sensor systems installed into their homes: familiarization, curiosity, and integration. Individuals first familiarized themselves with how the sensor may be intrusive. The familiarization phase is followed by the curiosity phase where they see how their own behaviors affect the operation of the sensors. After the first two interactive phases, residents settle into the integration phase where they generally ignore the sensors and carry on with their routines. Evidence from this study shows that individuals may accept the home sensing if it is adequately unobtrusive and is relatively easy to ignore on an everyday basis.

In summary, prior research has found that perceived need or value is a critical part of acceptance of smart home sensing technologies, however, if sensing is unobtrusive enough, the potential value from the system can provide peace of mind. One of the strengths of embedded assessment systems is that it can monitor an individual over a long period of time and provide information that may be very helpful in the long-term future for understanding the progression of decline. In this thesis, we aim to understand the short-term (in addition to the long-term) information needs of stakeholders in order to provide value in the short-term, which can aid in acceptance and adoption of embedded assessment technology. Moreover, we will investigate how the specific information collected from task-based embedded assessment can help meet those information needs. The sensing approach used in this thesis aims to minimize the costs and barriers to adoption (privacy, intrusiveness, learning to use new technologies, etc.,) and to maximize the value of the system by making the information understandable and presented in a way that allows it to be useful.

## 2.3 References

1. Fried, L.P., Herdman, S.J., Kuhn, K.E., Rubin, G., and Turano, K. Preclinical disability: hypotheses about the bottom of the iceberg. *Journal of Aging and Health* 3, 2 (1991), 285.
2. Morris, M., Lundell, J., Dishman, E., and Needham, B. New perspectives on ubiquitous computing from

- ethnographic study of elders with cognitive decline. *Ubicomp 2003: Ubiquitous Computing*, (2003), 227–242.
3. Aneshensel, C.S., Pearlin, L.I., Levy-Storms, L., and Schuler, R.H. The transition from home to nursing home mortality among people with dementia. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 55, 3 (2000), S152.
  4. Sitzer, D.I., Twamley, E.W., and Jeste, D.V. Cognitive training in Alzheimer's disease: a meta-analysis of the literature. *Acta Psychiatrica Scandinavica* 114, 2 (2006), 75–90.
  5. Clare, L. Cognitive training and cognitive rehabilitation for people with early-stage dementia. *Reviews in Clinical Gerontology* 13, 01 (2003), 75–83.
  6. Loewenstein, D.A., Acevedo, A., Czaja, S.J., and Duara, R. Cognitive rehabilitation of mildly impaired Alzheimer disease patients on cholinesterase inhibitors. *American Journal of Geriatric Psych* 12, 4 (2004), 395.
  7. Valenzuela, M. and Sachdev, P. Can cognitive exercise prevent the onset of dementia? Systematic review of randomized clinical trials with longitudinal follow-up. *American Journal of Geriatric Psych* 17, 3 (2009), 179.
  8. Schnaider Beerli, M., Werner, P., Davidson, M., and Noy, S. The cost of behavioral and psychological symptoms of dementia (BPSD) in community dwelling Alzheimer's disease patients. *International journal of geriatric psychiatry* 17, 5 (2002), 403–408.
  9. Schulz, R. and Martire, L.M. Family caregiving of persons with dementia: prevalence, health effects, and support strategies. *American Journal of Geriatric Psych* 12, 3 (2004), 240.
  10. Leifer, B.P. Early diagnosis of Alzheimer's disease: clinical and economic benefits. *Journal of the American Geriatrics Society* 51, 5s2 (2003), S281–S288.
  11. Findley, L.J. The economic impact of Parkinson's disease. *Parkinsonism & Related Disorders* 13, (2007), S8–S12.
  12. Langa, K.M., Chernew, M.E., Kabeto, M.U., et al. National Estimates of the Quantity and Cost of Informal Caregiving for the Elderly with Dementia\*. *Journal of General Internal Medicine* 16, 11 (2001), 770–778.
  13. Amieva, H., Jacqmin-Gadda, H., Orgogozo, J.M., et al. The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* 128, 5 (2005), 1093.
  14. Ott, B.R., Lafleche, G., Whelihan, W.M., Buongiorno, G.W., Albert, M.S., and Fogel, B.S. Impaired awareness of deficits in Alzheimer disease. *Alzheimer Disease & Associated Disorders* 10, 2 (1996), 68.
  15. Holland, C.A. and Rabbitt, P. People's awareness of their age-related sensory and cognitive deficits and the implications for road safety. *Applied Cognitive Psychology* 6, 3 (1992), 217–231.
  16. Barrett, A.M., Eslinger, P.J., Ballentine, N.H., and Heilman, K.M. Unawareness of cognitive deficit (cognitive anosognosia) in probable AD and control subjects. *Neurology* 64, 4 (2005), 693.
  17. Graham, D.P., Kunik, M.E., Doody, R., and Snow, A.L. Self-reported awareness of performance in dementia. *Cognitive Brain Research* 25, 1 (2005), 144–152.
  18. Suchy, Y., Kraybill, M.L., and Franchow, E. Instrumental activities of daily living among community-dwelling older adults: Discrepancies between self-report and performance are mediated by cognitive reserve. *Journal of Clinical and Experimental Neuropsychology* Jul, 1 (2010), 1–9.
  19. Lorenz, A. Indicators of preclinical disability: Women's experiences of an aging body. *Journal of women & aging* 21, 2 (2009), 138–151.
  20. Okonkwo, O.C., Griffith, H.R., Vance, D.E., Marson, D.C., Ball, K.K., and Wadley, V.G. Awareness of Functional Difficulties in Mild Cognitive Impairment: A Multidomain Assessment Approach. *Journal of the American Geriatrics Society* 57, 6 (2009), 978–984.
  21. Kemp, N.M., Brodaty, H., Pond, D., and Luscombe, G. Diagnosing dementia in primary care: the accuracy of informant reports. *Alzheimer Disease & Associated Disorders* 16, 3 (2002), 171.
  22. Ready, R.E., Ott, B.R., and Grace, J. Patient versus informant perspectives of Quality of Life in Mild Cognitive Impairment and Alzheimer's disease. *International journal of geriatric psychiatry* 19, 3 (2004), 256–265.
  23. Diehl, M., Marsiske, M., Horgas, A.L., Rosenberg, A., Saczynski, J.S., and Willis, S.L. The revised observed tasks of daily living: A performance-based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology* 24, 3 (2005), 211.
  24. Holm, M. and Rogers, J. Performance assessment of self-care skills. Assessment in occupational therapy mental health: an integrative approach. In *Assessments in occupational therapy mental health: an integrative approach*. B. Hemphill-Pearson, Thorofare, NJ, 1999, 117–124.
  25. Owsley, C., Sloane, M., McGwin, G., and Ball, K. Timed instrumental activities of daily living tasks: Relationship to cognitive function and everyday performance assessments in older adults. *Gerontology* 48, (2002), 254–265.

26. Rizzo, M., Anderson, S.W., Dawson, J., Myers, R., and Ball, K. Visual attention impairments in Alzheimer's disease. *Neurology* 54, 10 (2000), 1954.
27. Berner, E.S. and Moss, J. Informatics challenges for the impending patient information explosion. *Journal of the American Medical Informatics Association* 12, 6 (2005), 614–617.
28. Abbo, E.D., Zhang, Q., Zelder, M., and Huang, E.S. The increasing number of clinical items addressed during the time of adult primary care visits. *Journal of General Internal Medicine* 23, 12 (2008), 2058–2065.
29. Kottke, T.E., Brekke, M.L., and Solberg, L.I. Making "time" for preventive services. *Mayo Clinic proceedings. Mayo Clinic*, (1993), 785.
30. Yarnall, K.S., Pollak, K.I., Ostbye, T., Krause, K.M., and Michener, J.L. Primary care: is there enough time for prevention? *American Journal of Public Health* 93, 4 (2003), 635.
31. Demongeot, J., Virone, G., Duchêne, F., et al. Multi-sensors acquisition, data fusion, knowledge mining and alarm triggering in health smart homes for elderly people. *Comptes Rendus Biologies* 325, 6 (2002), 673–682.
32. <http://www.elitecare.com>.
33. Skubic, M., Alexander, G., Popescu, M., Rantz, M., and Keller, J. A smart home application to eldercare: Current status and lessons learned. *Technology and Health Care* 17, 3 (2009), 183–201.
34. Mihailidis, A., Boger, J., Canido, M., and Hoey, J. The use of an intelligent prompting system for people with dementia. *interactions* 14, 4 (2007), 34–37.
35. Helal, S., Mann, W., El-Zabadani, H., King, J., Kaddoura, Y., and Jansen, E. The gator tech smart house: A programmable pervasive space. *Computer*, (2005), 50–60.
36. Lawton, M.P. and Brody, E.M. Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist* 9, 3 Part 1 (1969), 179.
37. Tomaszewski Farias, S., Cahn-Weiner, D.A., Harvey, D.J., et al. Longitudinal changes in memory and executive functioning are associated with longitudinal change in instrumental activities of daily living in older adults. *The Clinical Neuropsychologist* 23, 3 (2009), 446–461.
38. Cahn-Weiner, D.A., Malloy, P.F., Boyle, P.A., Marran, M., and Salloway, S. Prediction of functional status from neuropsychological tests in community-dwelling elderly individuals. *The Clinical Neuropsychologist* 14, 2 (2000), 187–195.
39. Morris, M., Intille, S.S., and Beaudin, J.S. Embedded assessment: Overcoming barriers to early detection with pervasive computing. *Pervasive Computing*, (2005), 333–346.
40. Dishman, E. Inventing wellness systems for aging in place. *Computer*, (2004), 34–41.
41. Weiser, M. The computer for the 21st century. *Scientific American* 272, 3 (1995), 78–89.
42. Chan, M., Estève, D., Escriba, C., and Campo, E. A review of smart homes-present state and future challenges. *Computer methods and programs in biomedicine* 91, 1 (2008), 55–81.
43. Demiris, G. and Hensel, B.K. Technologies for an aging society: a systematic review of "smart home" applications. *Yearbook of Medical Informatics*, (2008), 33–40.
44. Kientz, J.A., Patel, S.N., Jones, B., Price, E., Mynatt, E.D., and Abowd, G.D. The georgia tech aware home. *CHI'08 extended abstracts on Human factors in computing systems*, (2008), 3675–3680.
45. Tran, Q.T., Calcaterra, G., and Mynatt, E.D. Cook's collage: Deja vu display for a home kitchen. *Proceedings of HOIT*, (2005), 1–16.
46. Rowan, J. and Mynatt, E.D. Digital family portrait field trial: Support for aging in place. *Proceedings of the SIGCHI conference on Human factors in computing systems, April*, (2005), 02–07.
47. Hayes, G.R., Truong, K.N., Abowd, G.D., and Pering, T. Experience buffers: a socially appropriate, selective archiving tool for evidence-based care. *CHI'05 extended abstracts on Human factors in computing systems*, (2005), 1435–1438.
48. Isoda, Y., Kurakake, S., and Nakano, H. Ubiquitous sensors based human behavior modeling and recognition using a spatio-temporal representation of user states. (2004).
49. Philipose, M., Fishkin, K.P., Perkowitz, M., et al. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, (2004), 50–57.
50. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., and Rehg, J.M. A scalable approach to activity recognition based on object use. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, (2007), 1–8.
51. van Kasteren, T., Noulas, A., Englebienne, G., and Kröse, B. Accurate activity recognition in a home setting. *Proceedings of the 10th international conference on Ubiquitous computing*, (2008), 1–9.



52. Buettner, M., Prasad, R., Philipose, M., and Wetherall, D. Recognizing daily activities with rfid-based sensors. *Proceedings of the 11th international conference on Ubiquitous computing*, (2009), 51–60.
53. Intille, S., Larson, K., Tapia, E., et al. Using a live-in laboratory for ubiquitous computing research. *Pervasive Computing*, (2006), 349–365.
54. Kim, E., Helal, S., and Cook, D. Human activity recognition and pattern discovery. *Pervasive Computing, IEEE* 9, 1 (2009), 48–53.
55. Chen, C., Das, B., and Cook, D.J. A Data Mining Framework for Activity Recognition In Smart Environments. *Proceedings of the International Conference on Intelligent Environments 2010*, (2010).
56. Nazerfard, E., Das, B., Holder, L.B., and Cook, D.J. Conditional Random Fields for Activity Recognition in Smart Environments. *Proceedings of IHI 2010*, (2010).
57. Rashidi, P. and Cook, D.J. Keeping the resident in the loop: Adapting the smart home to the user. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 39, 5 (2009), 949–959.
58. Tyrer, H.W., Aud, M.A., Alexander, G., Skubic, M., and Rantz, M. Early Detection of Health Changes In Older Adults. *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, (2007), 4045–4048.
59. Hayes, T., Abendroth, F., Adami, A., Pavel, M., Zitzelberger, T.A., and Kaye, J.A. Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimer's and Dementia* 4, 6 (2008), 395–405.
60. Hayes, T., Pavel, M., and Kaye, J. An Approach for Deriving Continuous Health Assessment Indicators From In-home Sensor Data. *Technology and Aging: Selected Papers from the 2007 International Conference on Technology and Aging*, (2008), 130–137.
61. Kaye, J., Hayes, T., Zitzelberger, T., et al. Deploying wide-scale in-home assessment technology. *Technology and Aging: Selected papers from the 2007 International Conference on Technology and Aging*, (2007), 19–26.
62. Pfeffer, R.I., Kurosaki, T.T., Harrah, C.H., Chance, J.M., and Filos, S. Measurement of functional activities in older adults in the community. *Journal of gerontology* 37, 3 (1982), 323.
63. Cullum, C.M., Saine, K., Chan, L.D., Martin-Cook, K., Gray, K.F., and Weiner, M.F. Performance-based instrument to assess functional capacity in dementia: The Texas Functional Living Scale. *Cognitive and Behavioral Neurology* 14, 2 (2001), 103.
64. Burns, T., Mortimer, J.A., and Merchak, P. Cognitive Performance Test: a new approach to functional assessment in Alzheimer's disease. *Journal of geriatric psychiatry and neurology* 7, 1 (1994), 46.
65. Griffith, H.R., Belue, K., Sicola, A., et al. Impaired financial abilities in mild cognitive impairment: a direct assessment approach. *Neurology* 60, 3 (2003), 449.
66. Zanetti, O., Frisoni, G.B., Rozzini, L., Bianchetti, A., and Trabucchi, M. Validity of direct assessment of functional status as a tool for measuring Alzheimer's disease severity. *Age and ageing* 27, 5 (1998), 615.
67. Schwartz, M.F., Segal, M., Veramonti, T., Ferraro, M., and Buxbaum, L.J. The Naturalistic Action Test: A standardised assessment for everyday action impairment. *Neuropsychological Rehabilitation* 12, 4 (2002), 311–339.
68. Gill, T.M., Robison, J.T., and Tinetti, M.E. Difficulty and dependence: two components of the disability continuum among community-living older persons. *Annals of Internal Medicine* 128, 2 (1998), 96.
69. Hodges, M.R., Newman, M.W., and Pollack, M.E. Object-Use Activity Monitoring: Feasibility for People with Cognitive Impairments. (2009).
70. Hodges, M., Kirsch, N., Newman, M., and Pollack, M. Automatic Assessment of Cognitive Impairment Through Electronic Observation of Object Usage. *Pervasive Computing*, (2010), 192–209.
71. Cook, D.J. and Schmitter-Edgecombe, M. Assessing the quality of activities in a smart environment. *Methods of information in medicine* 48, 5 (2009), 480.
72. Kang, H.G., Mahoney, D.F., Hoenig, H., et al. In Situ Monitoring of Health in Older Adults: Technologies and Issues. *Journal of the American Geriatrics Society* 58, 8 (2010), 1579–1586.
73. Kentta, O., Merilahti, J., Petakoski-Hult, T., Ikonen, V., and Korhonen, I. Evaluation of technology-based service scenarios for supporting independent living. *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, (2007), 4041–4044.
74. Courtney, K.L., Demiris, G., Rantz, M., and Skubic, M. Needing smart home technologies: the perspectives of older adults in continuing care retirement communities. *Informatics in Primary Care* 16, 3 (2008), 195–201.
75. Beach, S., Schulz, R., Downs, J., Matthews, J., Barron, B., and Seelman, K. Disability, age, and informational privacy attitudes in quality of life technology applications: Results from a national Web survey. *ACM Transactions on Accessible Computing (TACCESS)* 2, 1 (2009), 1–21.

76. Beaudin, J.S., Intille, S.S., and Morris, M.E. To track or not to track: user reactions to concepts in longitudinal health monitoring. *Journal of medical Internet research* 8, 4 (2006).
77. Demiris, G., Skubic, M., Rantz, M.J., et al. Facilitating interdisciplinary design specification of "smart" homes for aging in place. *Studies in Health Technology and Informatics* 124, (2006), 45.
78. Demiris, G., Hensel, B.K., Skubic, M., and Rantz, M. Senior residents' perceived need of and preferences for "smart home" sensor technologies. *International journal of technology assessment in health care* 24, 01 (2008), 120–124.
79. Demiris, G., Oliver, D.P., Dickey, G., Skubic, M., and Rantz, M. Findings from a participatory evaluation of a smart home application for older adults. *Technology and Health Care* 16, 2 (2008), 111–118.
80. Lee, M.L. and Dey, A.K. Embedded assessment of aging adults: A concept validation with stakeholders. *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, (2010), 1–8.
81. Glaser, B.G. and Strauss, A.L. *The discovery of grounded theory: Strategies for qualitative research*. Aldine, 1967.
82. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (Revised 4th ed.)*. Washington, DC, 2000.
83. Lee, M. and Dey, A. Reflecting on Pills and Phone Use: Supporting Awareness of Functional Abilities for Older Adults. *Proceedings of CHI 2011*, (2011), in press.
84. Birnholtz, J. and Jones-Rounds, M.K. Independence and interaction: understanding seniors' privacy and awareness needs for aging in place. *Proceedings of the 28th international conference on Human factors in computing systems*, (2010), 143–152.
85. Mamykina, L., Mynatt, E., Davidson, P., and Greenblatt, D. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, (2008), 477–486.
86. Mirowsky, J., & Ross, C. E.. Eliminating defense and agreement bias from measures of the sense of control: A 2 x 2 Index. *Social Psychology Quarterly*, 54, (1991)127–145.
87. Patterson, D.J., Fox, D., Kautz, H., and Philipose, M. "Fine-grained activity recognition by aggregating abstract object usage," *Proceedings. Ninth IEEE International Symposium on Wearable Computers*, (2005)44- 51.
88. Feki MA, Biswas J, Tolstikov A. Model and algorithmic framework for detection and correction of cognitive errors *Technololgy Health Care*. 17, 3, (2009)203-19.
89. Phua, C.; Foo, V.S.-F.; Biswas, J.; Tolstikov, A.; Aung-Phyo-Wai Aung; Maniyeri, J.; Weimin Huang; Mon-Htwe That; Duangui Xu; Chu, A.K.-W.; , "2-layer Erroneous-Plan Recognition for dementia patients in smart homes," *e-Health Networking, Applications and Services, 2009. Healthcom 2009. 11th International Conference on* , vol., no., pp.21-28,

# 3

## Investigating the Information Needs and Potential Uses of Embedded Assessment

In order to design a sensing system that can monitor the functional abilities of individuals in their home, it is important to identify what tasks to focus on and what information about task performance would stakeholders find useful. The approach in this thesis is to start with gathering design requirements from potential users of embedded assessment systems. To address research questions RQ1 and RQ2 about the information needs and potential uses of embedded assessment data, we have conducted a formative concept validation study (Lee & Dey, 2010) to understand how older adults, their caregivers, and clinicians would use task-based embedded assessment data. In this section, we discuss the results of a concept validation study of potential embedded sensing systems designed to monitor how well elders perform everyday activities. In this qualitative study with stakeholders (elders, family caregivers, and medical clinicians), we proposed concepts for home sensing systems and investigated how these systems and the data they collect can be used to improve recognition of changes associated with functional and cognitive decline. We identified the information needs of stakeholders as well as what value they would gain from embedded assessment data about IADLs, including improving elders' awareness of their abilities and empowering caregivers, doctors, and occupational therapists (OTs) to make better-informed decisions for treatment. We also discuss a number of issues that need to be addressed to obtain the most value from an embedded assessment approach and provide recommendations for designers of embedded assessment systems.

As discussed in Chapter 2, many prior works have looked at the value of monitoring the frequency or pattern of activities an individual performs in the home. An important question left unaddressed in previous research is whether information about *how well* IADLs are performed would actually provide value to elders, caregivers, and clinicians as earlier indicators for changes in functional abilities. If so, it is necessary to understand how to present performance information to each stakeholder. Embedded assessment technologies, like many sensor-based systems, can collect an overwhelming amount of data. This raises the following questions: How can the data in low-level sensor streams be presented as salient summaries for use by stakeholders? How do the information needs of elders differ from those of their caregivers and clinicians?

### 3.1 Concept Validation Method

We investigated these questions using a concept validation technique using concrete scenario-based descriptions of embedded home sensing concepts (described below) and various representations of the data that these systems could produce.

### 3.1.1 Participants

Our concept validation study was conducted with sixteen participants: four fully-functioning elders (age range 67-86), six family caregivers, three geriatricians, and three occupational therapists (OTs). We focused on independent, fully-functioning elders because they would likely benefit most from early detection. We recruited them from a social club for retired employees of a corporation. Because the caregivers of these elders lived out of town, we recruited (from Craigslist) other caregivers who looked after the health of a parent. The geriatricians and OTs worked at a large local university hospital.

### 3.1.2 Interviews & Analysis

The concept validation session with elders started out by asking them to assess their own functional abilities and to identify any declines in health. We discussed with them how they become aware of declines and what they do when they become aware. We then introduced three embedded assessment concepts (described in the next paragraph) as probes for discussion to get their impressions about whether they wanted these technologies in their home. Then we showed them representations of data hypothetically generated from having these sensing concepts in their home, to probe their impressions about the usefulness of these data. We began with representations that showed the least amount of information (*e.g.*, short-term, task completion only, no process detail) and asked whether this information was useful, in what way was it useful, what action (if any) they would take, and what additional information they wanted. In response to their request for more information, we would show them other representations that had more features (longer-term views, process steps, *etc.*). Sessions with the clinicians (geriatricians and OTs) followed the same procedure but began with a discussion about how they currently collect functional data about a patient and also included a discussion about how embedded assessment systems can fit into their practice. Likewise, sessions with caregivers began with asking them how they currently keep track of their parents' health. Based on transcribed audio recordings of the concept validation sessions, we used an axial coding technique to code each transcribed comment from our participants and generate themes common across our stakeholders. Stakeholder comments about each data dimension were identified and grouped.


### 3.1.3 Concepts for Validation

The following sensing concepts for embedded assessment of specific IADLs were evaluated in this study: Medication Monitor (Figure 3-1), Coffee Chronicler (Figure 3-2), and Telephone Tracker (Figure 3-3). Medicine taking, coffee making and telephone use were chosen based on a number of factors. We considered the entire canonical list of Instrumental Activities of Daily Living because they are commonly used in existing self-report, informant report, and expert assessment instruments (Zanetti et al., 1998; Holm & Rogers, 1999; Owsley et al., 2002). We also considered the current state of sensing technology so that our concepts would be feasible for implementation. In an earlier field study with eight community-dwelling older adults, we also observed how elders perform these tasks in their everyday routines to identify the low-level steps and to understand how existing simple sensors could detect the individual steps of the tasks.

The Medication Monitor consists of a smart pillbox, a vision-enabled kitchen table, and an augmented water glass. The smart pillbox knows its location, when the user is grasping it, which doors are opened, and how much time the individual takes to decide which door to open. Once the pills are placed on the table, the vision-enabled kitchen table uses a ceiling-mounted camera to identify which pills are on the table and to monitor the pill-sorting task. The intelligent water glass senses its position on the counter, when it is filled, when it is grasped, and when it is tilted while drinking. The combination of these various devices can be used to sense when each step is started or finished, how long she spends in each step the occurrence of errors such as opening the wrong door or leaving the pills on the table.


## Medication Monitor

Smart Pillbox




- knows where you put it
- which doors were opened
- whether pills were taken out
- knows how long/hard it took for you to open a door

Kitchen table that "sees" pills



- knows what pills are on it
- knows what pills were removed
- small camera mounted on the ceiling that only sees the table top

Intelligent Water Glass



- knows whether it's been filled
- knows if it's being held
- knows when you are drinking (orientation)
- knows where you place it

**Figure 3-1. The Medication Monitor concept that tracks the process of taking medications.**

## Coffee Chronicler

Smart Coffee Maker



knows:

- when coffee filter & coffee grounds are in machine
- when enough water is in the machine
- when the carafe runs empty and can burn
- how often it's used and at what time of day
- all the steps in making coffee

Steps for making coffee (example):

1. Walk to coffee maker
2. Take out old filter/grounds and throw away
3. Put in new filter
4. Measure out coffee grounds and put in machine
5. Pour out old coffee
6. Fill carafe with correct amt of water
7. Pour water into machine.
8. Turn on machine
9. Turn off machine when done with coffee

**Figure 3-2. The Coffee Chronicle concept that uses an instrumented coffeemaker to monitor the multiple steps in making a pot of coffee.**

The Coffee Chronicler concept consists of an augmented coffee maker that can detect if the carafe is empty, the quantity of coffee grounds in the machine, how much water is the machine, and whether the ratio of coffee to water is reasonable. The Coffee Chronicler can detect when steps are missed, repeated, or performed not as well as they should be (for example, measuring out too many scoops of coffee grounds).

The Telephone Tracker monitors the frequency of incoming and outgoing calls, which may provide an indicator of social connectedness. The Telephone Tracker is not only able to track if calls were made successfully but also can detect the errors in the task process such as when the user misdials the telephone.



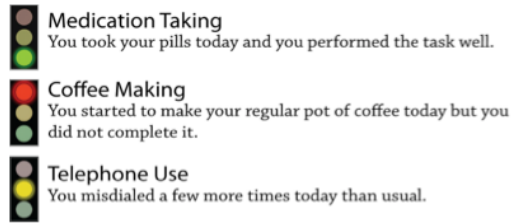
**Figure 3-3. The Telephone Tracker concept that monitors how an individual dials, answers, and spends time on the phone.**

Based on our sensing concepts, we generated simulated data that would be collected if these systems were deployed for a year in an individual's home. Our data representations of IADL task behavior showed three features made possible by the low-level sensing of the tasks (Figure 3-4, Figure 3-5, Figure 3-6):

- 1) task performance (instead of only task completion)
- 2) long-term view
- 3) process details about individual steps of the task.

#### 3.1.4 Task Completion vs. Task Performance

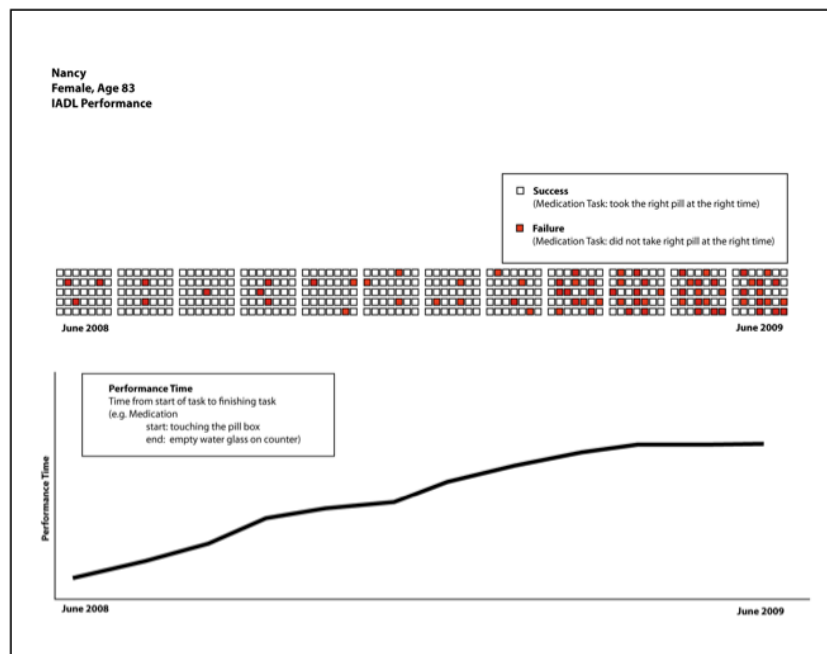
Like other related systems, our embedded assessment concepts can sense whether an individual has completed the task or not. However, our concepts were also designed to track how well the user performs these tasks. Included in the measure of task performance are: the amount of time spent on the task, how accurately they performed the task (*e.g.*, measuring out coffee), and the number of recovered errors. One of our simulated data examples (Figure 3-5) showed nearly perfect task completion early on (*e.g.*, no missing pills) but, at the same time, also showed inefficiencies involved in the task (*e.g.*, taking longer than usual to sort the pills).



**Figure 3-4. High-level data representation that shows short-term task performance. A green light indicates normal performance, a yellow light indicates a decrease in task performance for the current day, and a red light shows a failure in task completion.**

### 3.1.5 Long-term vs. Short-term

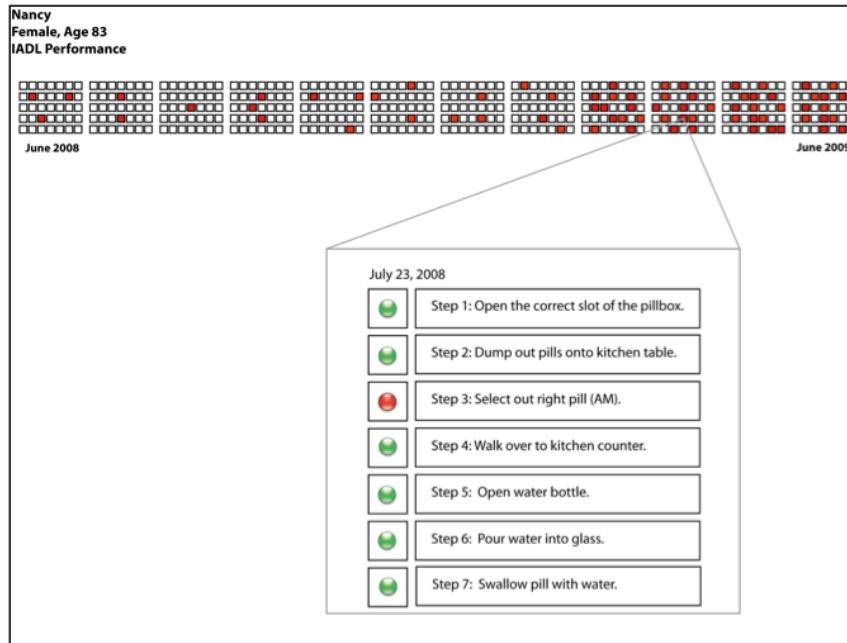
Our representations either showed a longitudinal range of data (a year's worth of aggregated or sampled short-term data, *e.g.*, Figure 3-5) or short-term data (*e.g.*, (Figure 3-4) that shows task status for a single day or week. Many home sensing systems emphasize intervention based on short-term data about task completion, so we wanted to assess the value in viewing long-term data about task performance.



**Figure 3-5. Long-term representation of task completion (top, calendar-style) and task performance (bottom, average time to complete medication-taking task).**

### 3.1.6 Process Details

One of the fortunate side effects of designing a system that monitors the task performance in addition to merely task completion is that the system has intimate knowledge of each atomic step in the process of carrying out the IADL. We provided information (*e.g.*, Figure 3-6) about which steps were completed, attempted but not completed, not initiated at all, or completed out of order. We investigated whether this highly-detailed information would be useful for understanding the precise nature of any breakdowns observed and for developing appropriate interventions.



**Figure 3-6. Long-term representation of task completion (top, calendar-style) and task performance (bottom, average time to complete medication-taking task).**

### 3.2 The Potential to Support Awareness of Functional Abilities

The results of the concept validation showed how embedded assessment data provide stakeholders with a greater awareness of changes in functional abilities and which specific features of the data were valuable to different stakeholders and how these features would support their goals.

During our interviews, the elders without any significant functional deficits in our study experienced a conflict between their current sense of awareness of their own abilities and their concerns about losing awareness in the future. In their current state, a monitoring system is redundant because they feel they know and can stay aware of their own capabilities, breakdowns, and inefficiencies. Many said they saw the need for monitoring only after they start to have a problem with these particular tasks. For example, Elder #1 (E1) said, *“If it got to the point where it was essential ... where I was making mistakes, then some sort of a system like that would be useful.”*

However, the same elders also recognized that they may lose their ability to stay aware of changes in their abilities. Many reflected on the experiences of their own parents or older friends as they struggled with decline in the last stages of their lives. E3 said, *“We know when we get older we will lose our hearing and memory.”* E2 said *“I want the information in my house when data indicates that people around my age or a little younger start to get problems that are serious...I don’t want to wait until I know there is a problem.”* As a result, these elders expressed a desire to have embedded assessment in their homes right now so they can maintain awareness and remain functional longer. Even though we did not include the perspectives of more impaired individuals in this study, these perspectives came through the vicarious experience of the healthy individuals.

Elders had no problems and even suggested sharing detailed information about private IADLs, such as medication taking, with their family, doctor, and close friends. E3 said, *“We talk about our health problems all the time with our friends anyway.”* Clinicians were more concerned about privacy and made it clear that patients must know that they are being monitored before the data could be used in the clinic.



The geriatricians in our study indicated that embedded assessment data provides them with information they do not normally have access to, especially due to the limited amount of time (a few minutes) they can spend asking patients about the details of their abilities. Occupational therapists are accustomed to dealing with functional assessment data but said embedded assessment data could provide them with a larger time window into a patient's abilities rather than infrequent snapshots of functioning. Caregivers also found the embedded assessment process data to be useful for showing details about their loved one's abilities that they would not normally know because they do not normally talk about these types of "mundane" tasks. Caregiver #6 (CG6) said, "*I don't know if she's screwing up and not telling me. She could be screwing up everyday making her coffee. It would show that she was slipping more than she would let on.*" In the following section, we describe which specific features of the data each stakeholder wanted, which can help inform the design of representations of embedded assessment data.

### 3.3 Usefulness of Task-based Embedded Assessment Data Features

The results were analyzed to identify the preferences of each stakeholder along the three data dimensions presented. Different stakeholders needed different combinations of data features to assist them with achieving their goals.

#### 3.3.1 Task Completion vs. Task Performance

All stakeholders found the task completion information (whether the task was completed with an acceptable outcome) useful because it showed how often the individual did not or was unable to complete the task. When an important task such as medication taking was missed consistently, all stakeholders recognized the need for intervention. In addition and more importantly, all stakeholders also found task performance (the quality of the outcome, the amount of effort or time spent, the number of errors encountered during the task) to be helpful.

Occupational therapists found performance time and the number of errors to be valuable in their practice because it gives them a measure of *adequacy*, a measure they normally look for in functional assessments. To apply the appropriate adaptation, OTs evaluate functional abilities along three different criteria: *independence* (the ability for the patient to carry out the task on their own), *adequacy* (the ability of the patient to perform the task with precision and economy of effort), and *safety* (the ability of the patient to avoid potentially dangerous situations) (Holm & Rogers 1991).

Elders also said that task performance data would provide them with early indicators for problems. Regarding Medication Monitor data, E2 said, "*[Task completion] is more useful for telling you that you're not taking your pills, [task performance time] is more useful for telling you what the problem is.*" Elders used these early indicators as triggers for adaptations to ensure task completion. For example, when seeing an increase in misdials (while still completing the call eventually) in the Telephone Tracker data, E4 said, "*It helps me understand what the problem is and how I can correct it. Gee, I'd better take a little more time in dialing, or make the names bigger.*"

Geriatricians said that performance data provided them with more information to understand the patient's abilities from a qualitative standpoint than they could get from a clinical test or observation. Geriatrician G3 said, "*Absolutely, it's a trigger for clinical investigation, to bring the patient in, sitting down and talk and figuring out what's going on.*" Caregivers said they would use decreases in task performance as a trigger to keep a closer eye on their loved one and to start a conversation if the changes are getting worse. CG3 said, "*I think it's because [it] initiates a conversation between me and my mother. I would never know that she was misdialing more often unless she complaining about it.*"

#### 3.3.2 Long-term vs. Short-term

All interviewees found the long-term view of the data to be useful for understanding the trajectory of decline. Geriatricians said that the long-term view provided them with information about the evolution of the disability. A sudden onset of a problem can indicate an acute (or even temporary) change due to some trauma or change in the patient's life. A gradual onset of a problem can indicate a pattern more consistent with a progressive disease such as

dementia. OTs found the long-term view useful for understanding the nature of the particular disability and identifying the variation in people’s abilities over time. OT2 said, “*An assessment visit is a one-shot deal. It’s almost impossible to observe people’s habits.*”

Elders also considered the long-term representation useful for understanding how their abilities change. In fact E4 said that the long-term view “*sells the entire idea,*” meaning that the longitudinal data showing a pattern of decline provides a compelling reason to have an embedded assessment system in his home. Unlike geriatricians and occupational therapists, elders also said they would also like short-term views of the data, particularly for giving them an extra sense of security for the memory-intensive task of taking medications, which is consistent with findings from (Hayes et al., 2006). E2 said he wanted to have a more immediate indicator for missed pills or pills taken twice than his system of writing the date on the pill bottle. For the coffee making and telephone task, elders said they would be able to remember and notice if they had problems because they have more noticeable, though less critical, consequences, so short-term information about these tasks would be unnecessary.

### 3.3.3 Process Details

Elders, caregivers, and occupational therapists were interested in the breakdown of tasks at the process level because these stakeholders have the responsibility to identify and fix problems. In contrast, geriatricians found the process steps information to be too detailed. Geriatricians pointed to social workers, OTs, or geriatric nurses as professionals better suited for acting on process information. G2 said, “*I don’t think it’s appropriate for me to do all the things that other team members do.*” In their regular practice, OTs decompose tasks into individual steps so that they can identify the exact disability and provide the most appropriate adaptation to compensate for that disability. Caregivers wanted initiate a dialog to find the causes of the problem and a solution to help their loved ones to remain independent. CG2 said process details were helpful “*because you can know what pills are giving her the problem. The problem is to figure out the right solution [and] how to make it easier for them.*” Most elders also felt the process details were helpful in finding the exact problem to address. However, one elder, E4, said that this information was overkill because reading all the steps and seeing which were good or bad was too cognitively demanding and perhaps emotionally alarming. He said his ability and patience to look through these reports would be even less as his abilities decline. E4 said, “*I like the data but I don’t think I can make sense of it. Having a little voice come out and say ‘It’s taking you a lot of time to do step one’ that would be very helpful.*” He suggested the system highlight only the salient deviations instead of every step.

## 3.4 Limitations of Embedded Assessment Data

In our validation sessions, stakeholders also brought up some limitations in being able to interpret the data captured by the sensing concepts. These limitations discussed below lead us to design recommendations (Table 1) that will help inform the design of embedded assessment systems that can assist stakeholders in making sense of task performance data. The following sections detail each of these findings and limitations to using the data.

Findings	Design Recommendations
Embedded assessment increases awareness and is useful in clinical judgment.	Provide appropriate representations of data to different stakeholders to support awareness for elders and provide ecologically-valid, longitudinal data to clinicians.
The “Why” is missing from the data.	Embedded assessment data are merely triggers for further explorations of underlying health issues. Support data-driven inquiry with the user.
Lack of validation for critical values of significance.	Codify embedded assessment data into scales with critical values to quantify significance in the data. Correlate embedded assessment data with standard psychometric instruments.
The underlying behavior is noisy, not just the sensor data.	The system should include a rich model of the user’s actions to accommodate acceptable deviations from established routines. The system can highlight patterns and make simple suggestions to facilitate data exploration but should ultimately allow clinicians to use their own experience and intuition.

**Table 2. Summary of Findings and Design Recommendations**

### 3.4.1 The Why is Missing

Embedded assessment holds the promise that technology will be able to collect real life data from users to provide the answers to many of the questions doctors have about the deficits that people encounter in their everyday lives. However, our investigation revealed that the data collected from our sensing concepts are merely observed behaviors that require further explanation. An observed behavior can have any number of bio-psycho-social causes. For example, when looking at a chart showing that a patient has been increasingly skipping his daily medication, G2 remarked that while it shows an important pattern, the chart does not show why this behavior occurred. She said she would “*think about whether the [data] is providing information in a way that makes you think about different reasons for non-adherence.*” These reasons can include cognitive problems (e.g., forgetting to take their pills), medical problems (e.g., avoidance due to an unpleasant side effect), psychological problems (e.g., depression), or financial problems (e.g., can no longer afford to purchase medication). Doctors would also like to be made aware of a common deficit that manifests across different tasks. For example, a memory deficit might show up as beginning the pill taking routine but not completing it, forgetting to turn off the coffee maker, and dialing out-of-date phone numbers. The data should be shown in a way that makes these associations easy to spot.

Geriatricians said that they would engage the patient and their relatives in an extended interview and ask about their awareness of specific trends found in the embedded assessment data, to identify the possible reasons for the trends and provide the appropriate treatment. Likewise, when presented with the data about task inefficiencies or errors, caregivers would call up their loved one to find out the causes of the behaviors and try to assist them. CG5 said, “*Id probably talk to them about it. Try to troubleshoot and see if their routine had changed.*”

Occupational therapists, with their perspective of restoring functional abilities by intervening with compensatory techniques, need to know both the problem and its causes to apply the right adaptation. For example, consider two causes for skipping medication: forgetting to take the pills (cognitive) and not being able to reach all the pills in the pillbox (dexterity). An OT would apply two different adaptations to support the task: moving the pillbox to a more noticeable position or replacing the pillbox with a larger pillbox that is more easily grasped, respectively.

Elders expressed a need to understand the reasons for changes in their health as they get older. E3 remarked, “*As an individual, we like to figure out why.*” Embedded assessment data triggers them to investigate the causes and take proactive steps to control problems before they become bigger problems. For example, in reaction to data showing an increasing number of telephone misdials, E1 said he would like to know which numbers he was misdialing so that he could figure out what might be causing this behavior and stop misdialing.

We observed that embedded assessment data are best used as a trigger to explore and address the underlying causes of the problematic behaviors, rather than providing conclusive answers about the exact causes of the behavior. Thus designers working with this data should enable the user to investigate why the problems occurred. For example, systems can arm clinicians or caregivers with specific questions driven by the data to ask during a visit, or visualizations integrating multiple streams of data may help reveal the same underlying deficit.

### 3.4.2 Searching for Significance

Because the stakeholders have never been presented with the fine-grained and frequent data points provided by embedded assessment technology, they had difficulty determining when the illustrated changes in performance were significant enough to warrant concern and further action. Geriatrician G2 said about the data from the Telephone Tracker concept, “*At this level of subtlety, I probably don't know how much misdialing the patient would have to do before I would... do my cognitive impairment screening.*”

Geriatricians wanted to use these data in their clinical practice but expressed the concern that they needed a way to standardize the interpretation of the data. A normal part of their methodology is quantifying people's abilities and disabilities so that they can consistently and confidently apply heuristics (such as the DSM-IV (1994) criteria for dementia) for medical diagnoses of decline. Traditionally, doctors would rely on the subjective self-reports and reports from relatives about the functional abilities of the patient to provide them with evidence about how the impairment is interfering with everyday life. Now equipped with objective embedded assessment data, doctors want to operate on this objective data in a quantitative manner similar to how they operate on objective cognitive testing data such as from the Mini Mental Status Exam (Folstein, Folstein, & McHugh 1975). Doctors wanted embedded assessment data to be validated by their community so they can apply a heuristic (such as "*If the patient misses their medication 25% of the time, then it's time to be concerned and figure out what's going on.*"- G1). Occupational therapists also called for these task-specific "critical values" in the embedded assessment data to signal when these failures are interfering with the life of the patient. OT1 said, "*There would have to be critical red flags. How many times do they forget a certain step that's critical...by the tenth time or the sixth?*" The elders expressed the same need to understand when a change in observed functioning is sufficiently severe as to warrant either a minor reaction such as extra vigilance or a major reaction such as scheduling an appointment with a doctor or considering moving into an assisted living facility.

Caregivers on the other hand were able to decide on what data values would trigger them to initiate a conversation or provide assistance. The threshold values varied across different caregivers and was mostly determined by the caregiver's relationship with the individual. Some caregivers who keep in close contact with their loved one would ask about any small change, whereas others wanted to minimize their own intrusiveness into their loved one's life and would react only when they saw a dramatic decline in abilities. Even though some caregivers were hesitant to react to small changes, they still emphasized that they wanted to know about the small changes including the task performance information to understand how their loved one is doing on a regular basis. CG6 said, "*Seeing this line [Figure 3-5], it would indicate to me to keep a better eye on it.*"

Another factor that contributes to searching for significance was that some tasks are easier to determine critical values for than other tasks. For instance, all stakeholders easily set critical values for medication taking to be very low such that almost any change in performance warrants some investigation. In contrast, the coffee and telephone tasks were less critical for safety or health, so the critical values for the number of errors, missteps, or misdials are higher and less well-defined.

Embedded assessment systems should initially monitor tasks that have easily-defined critical values and should also closely align their approach with standardized functional assessments. Data from embedded assessment systems need to be correlated with other well-established outcome measures such as psychometric tests or diagnoses of dementia. Evaluations of embedded assessment systems should include measures for clinical outcomes. Evaluations such as in Kaye *et al.* 2007 provide good examples to follow.

### 3.4.3 Noisy Data from the User, Not from the Sensors

Even in the world of perfect sensors that can accurately detect people's actual behaviors, people's performance of tasks can be (and will likely be) highly variable. Unlike many applications of sensing technologies, embedded assessment not only has to deal with the noise generated from the sensors themselves but also the variability in the underlying behavior being sensed. Geriatrician G2 noted that many individuals do not follow a smooth, predictable stage of preclinical decline in functioning before the onset of a disease or dementia. People may experience a decline, recover momentarily, and revert back to a pattern of decline *or not*. The fact that embedded assessment technologies can capture performance data frequently at a high level of detail makes the temporary changes (potentially noise) in performance more apparent in the data. G2 remarked, "*There's just phenomenal variability in performance. People don't decline in a steady fashion.*"

*They're waxing and waning all over the place.*" Although the high-resolution data may be noisy, it can still enable clinicians to see a change from consistency to volatility in performance which is predictive of future decline.

Even if their abilities are relatively stable, individuals may still occasionally deviate from their routine when it is convenient to do so. For example E1 said, "I might not take all my pills all at one time, I might get up early, take a glucosamine [pill] if my knees hurt and then sleep until 11 and then take the rest then." No stakeholders wanted these small deviations to be flagged as errors because they are considered as "acceptable" noise. Some recommendations for designing sensing systems to accommodate noisy behaviors include: 1) Building rich models of users' actions including edge cases, and 2) Making sensing systems easy to update to accommodate acceptable deviations from established patterns.

The promise that embedded assessment will automatically provide early detection of disability based on clear, steady trends in the data may be more difficult to achieve than previously thought due to large variability in the actual behaviors being sensed. Geriatricians said even they have problems identifying meaningful patterns from the noisy data, so it would be difficult to automate this. Clinicians felt that the system should refrain from making a medical interpretation of the collected data, but rather allow clinicians to use their own experience and insights to figure out what problem(s) exists and exactly what caused it. Clinicians were comfortable with having systems take the role of identifying clear statistical patterns within variations and even suggesting particular avenues of inquiry. Embedded assessment systems can present information, highlight relationships, and even suggest causes but they should not aim to replace the clinical judgment.

### 3.5 Summary of Findings from Concept Validation

Embedded assessment technology can be used to monitor *how well* elders perform IADLs, not merely whether they were completed or not. Information about task effort, accuracy, and errors provide early indicators of decline before actual failures in task completion occur. Our subjects found this information from embedded assessment systems to be valuable, in providing increased awareness for elders and their family caregivers, and facilitating clinical judgment for geriatricians and occupational therapists. They found great potential in our concept sensing systems and the idea of embedded assessment.

However, we also discovered a number of issues that will impact how we will actually construct and deploy instances of embedded assessment in the next phase of our work. We found that different people carried out IADLs very differently from others, and thus embedded assessment systems that track task effort must be customized to the particular individual's method of carrying out the task. Our concept validation revealed three important issues that limit the usefulness of these systems. The data they produce do not explicitly explain why particular behaviors were observed, so sensing systems should either highlight the abilities underlying these behaviors or provide its data as triggers for further investigation. There is currently a lack of any standardized metrics by which to identify what frequency or severity of problems behaviors is significant enough to require action. Therefore, future evaluations of embedded assessment technology need to be correlated with functional clinical outcomes. Embedded assessment technology must not only deal with the noise in the sensing devices but also the large variations in the performance of IADLs. Systems can perform the statistics and highlight trends but should rely on the expertise of the user to make sense of the data.

### 3.6 References

- Diagnostic and statistical manual of mental disorders (DSM-IV). (1994). Washington, D. C. A. P. A.
- Diehl, M., Marsiske, M., Horgas, A.L., Rosenberg, A., Saczynski, J.S., and Willis, S.L. The revised observed tasks of daily living: A performance-based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology* 24, 3 (2005), 211.
- Folstein, M., Folstein, S. and McHugh, P. Mini-mental state. A practical method for grading the cognitive state of

patients for the clinician. *Journal of Psychiatric Research*, 12, (1975), 189–198.

- Hayes, T., Hunt, J., Adami, A. and Kaye, J. An Electronic Pillbox for Continuous Monitoring of Medication Adherence. In Proc. IEEE Engineering Medicine and Biological Society, (2006), 6400-6403.
- Holm, M. and Rogers, J. Performance assessment of self-care skills. Assessment in occupational therapy mental health: an integrative approach. In *Assessments in occupational therapy mental health: an integrative approach*. B. Hemphill-Pearson, Thorofare, NJ, 1999, 117-124.
- Lawton, M.P. and Brody, E.M. Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist* 9, 3 Part 1 (1969), 179.
- Lee, M.L. and Dey, A.K. Embedded assessment of aging adults: A concept validation with stakeholders. *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, (2010), 1–8.
- Kaye, J., Hayes, T., Zitzelberger, T., et al. Deploying wide-scale in-home assessment technology. *Technology and Aging: Selected papers from the 2007 International Conference on Technology and Aging*, (2007), 19-26.
- Owsley, C., Sloane, M., McGwin, G., and Ball, K. Timed instrumental activities of daily living tasks: Relationship to cognitive function and everyday performance assessments in older adults. *Gerontology* 48, (2002), 254–265.
- Zanetti, O., Frisoni, G.B., Rozzini, L., Bianchetti, A., and Trabucchi, M. Validity of direct assessment of functional status as a tool for measuring Alzheimer's disease severity. *Age and ageing* 27, 5 (1998), 615.

# 4

## dwellSense: A Task-based Embedded Assessment System

Based on the reactions in the concept validation study that data from task-based sensing has the potential to provide value to stakeholders, we have developed dwellSense, a system consisting of a suite of task-based sensors to monitor how well individuals carry out specific Instrumental Activities of Daily Living. The tasks selected for monitoring are the same that were evaluated in the concept validation and they include: taking medications, using the telephone, and using a coffeemaker. The making coffee task was the least well-received sensing concept because it was neither safety critical nor were there well-validated levels of performance by which to make judgments. Nonetheless, we decided that this task mimicked the multi-step process of the canonical meal preparation task commonly found in standardized IADL assessments. Furthermore, it is a common task that older adults in the United States perform frequently, making it easier to recruit and develop a generalizable sensing instrument for many individuals.

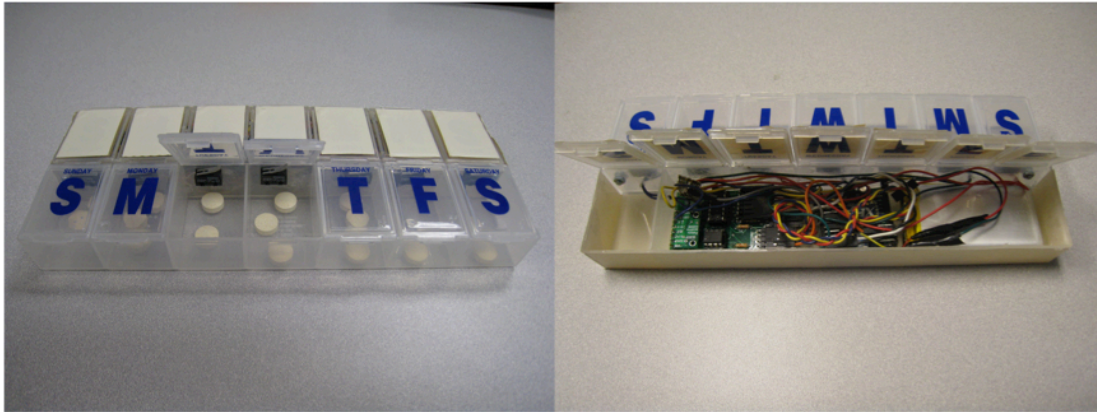
dwellSense consists of three main elements: sensing, infrastructure, and presentation, with the flow of information roughly proceeding through these elements in this order. dwellSense includes a suite of intelligent sensors that are designed to be embedded in the home environment to monitor how well individual carry out particular Instrumental Activities of Daily Living. dwellSense also includes the wireless and networking infrastructure to transfer data from the sensors to both local and remote repositories for processing and analysis. dwellSense also provides views of the sensor data through a web interface as well as through an in-home display. The following sections describe the sensing, infrastructure, and presentation components in more detail.

### 4.1 dwellSense Sensing Capabilities

dwellSense was designed to be able to monitor how well individuals perform everyday tasks in their own home. The concepts identified and validated in the concept validation (Chapter 3) were instantiated and engineered into into prototypes for testing. The approach used was to augment the existing artifacts that were already common in the household for mainly two reasons: 1) the familiarity of the objects themselves would minimize the concerns about adopting a “new” piece of technology and 2) simply adding sensing to existing objects can minimize the disruptions to the individual’s existing patterns of use. This augmentative approach was applied to a pillbox found in any drugstore for tracking medication taking, a landline phone for tracking telephone calls, and an off-the-shelf coffeemaker for tracking the multiple steps in making coffee.

#### 4.1.1 Medication Monitor

To monitor medication taking, a specialized pillbox (Figure 4-1) was designed to track which doors are opened and to track how the box is manipulated, held, shaken, or inverted. The aesthetic and functional design of the pillbox was deliberately made such that it was almost identical to consumer pillboxes common found in drugstores. In fact, the body of the pillbox consists of two standard, extra large sized, seven-day pillboxes. The pillboxes are attached back to back, with one pillbox hollowed out and used to contain the electronics, while the other pillbox functions identically to

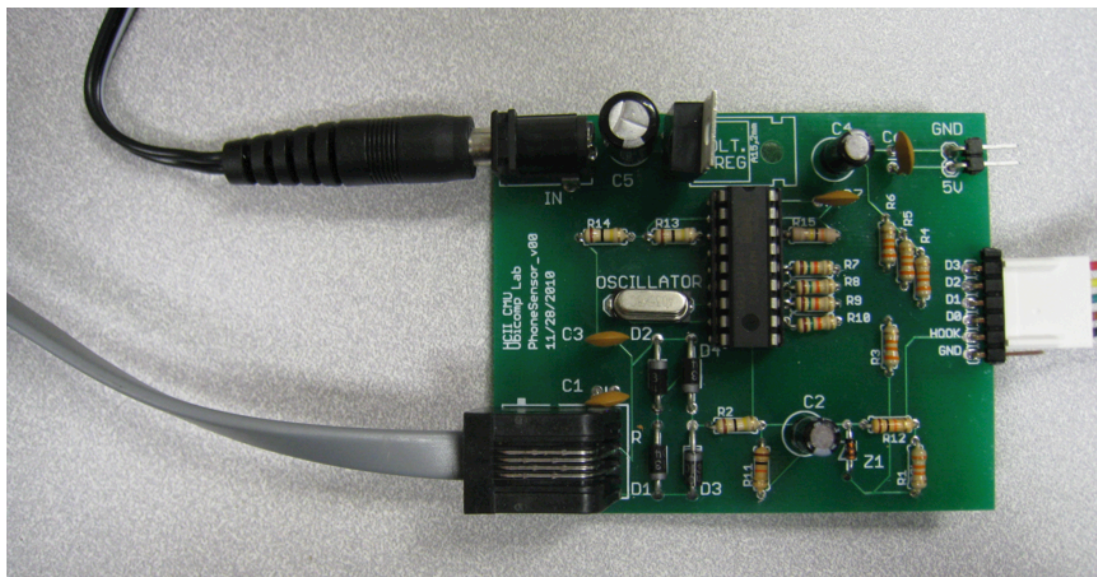


**Figure 4-1. The Smart Pillbox is part of the Medication Monitor system and can track when doors are opened and uses an accelerometer to track when the box is moved, shaken, and inverted.**

a non-augmented pillbox. The augmented pillbox is equipped with snap action sensors for each pillbox door to track when each door is opened. The augmented pillbox also contains a three-axis accelerometer that can track when the box is picked up, shaken, or inverted. Inverting the pillbox is a common gesture used by older adults to pour out the pills into their hands before taking them. The pillbox is simply the first step in taking medications. For each individual, we will consider their routines and find appropriate sensors to track their pill-taking task. For example, if an individual normally first retrieves their medications from the pillbox and then goes to the kitchen faucet to get water used to swallow their pills, a sensor that senses faucet use can provide one extra data point in their pill taking routine. Based on these sensors, the recorded data can be interpreted to identify task errors and inefficiencies such as choosing the wrong pillbox door.

#### 4.1.2 Telephone Tracker

The phone use task is monitored with the help of a custom-made electronic circuit (Figure 4-2) that plugs into the phone line and can keep track of the numbers dialed (or misdialed), incoming and outgoing calls, number of rings before answering the phone, and the duration of calls. The phone sensor does not record any of portion of the voice signal. The phone sensor has the benefit of not interfering with the normal operation of the telephone, making it nearly



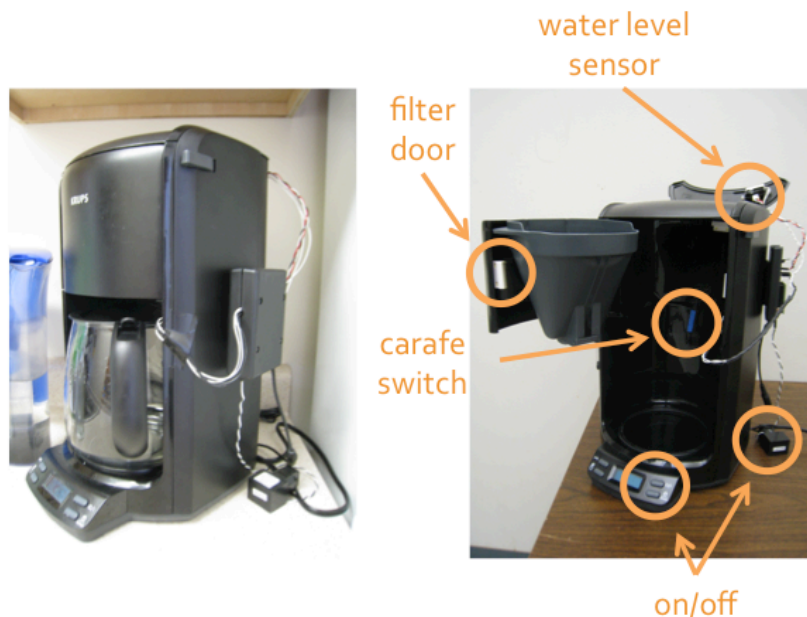
**Figure 4-2. Circuit for the Telephone Tracker sensor that plugs into the phone line and monitors when the phone is on/off the hook and which numbers are dialed.**



completely unobtrusive aside from the circuit discreetly placed out of sight in the home.

#### 4.1.3 Coffee Chronicler

To monitor the multi-step task of making a pot of coffee, a custom instrumented coffee maker (**Error! Reference source not found.**) was designed to track various steps for making coffee. The sensors can track when the water reservoir or filter door is opened and closed, whether the carafe is in place, the amount of water used, and whether the machine is turned on or off. Other auxiliary sensors in the environment can detect other steps in making coffee such as opening the cabinet where the coffee filters are placed, measuring a reasonable amount of coffee, or turning on the faucet to get water. Even though there are many acceptable action sequences to make a pot of coffee, there are still constraints in the order of steps that can be useful for identifying errors or inefficiencies. For all three tasks, the sensors are designed to monitor the individual steps of the tasks and can be used to identify recovered and non-recovered errors, measure the effort (in terms of time) it takes to perform the tasks. The detailed process data generated from the sensors is similar to the type of step-by-step data collected by standardized performance testing often used by occupational therapists (Holm & Rogers 1999).



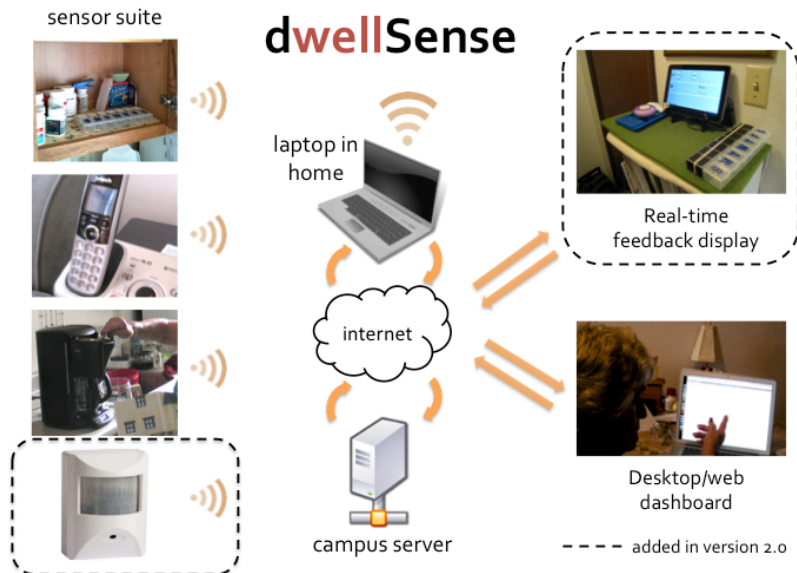
**Figure 4-3. The Coffee Chronicler uses sensors on an off-the-shelf coffeemaker to track the different steps in coffeemaking.**

#### 4.1.4 Wireless networking

The sensors also have been equipped with a microcontroller and a wireless radio that implements the Zigbee protocol to transmit their data in real time to a computer placed in the apartment. The microcontroller determines when the sensor changes state and then wakes up the radio momentarily to transmit the information and then puts the wireless radio back to sleep to conserve power. The instrumented pillbox and contact sensors are powered with batteries, but whenever possible, mains power is used for sensors (such as the phone sensor and instrumented coffee maker) that are not mobile and near a power outlet. The microprocessors and wireless radio are configured to poll the state of the sensor every 100 milliseconds and send any state changes over the network to be logged by the computer.

## 4.2 dwellSense Data Infrastructure

The dwellSense system includes an infrastructure (Figure 4-4) that allows data to flow from the sensors to a centralized repository for processing and distribution to the presentation elements. The sensors (described in the previous section) sense and transmit their information in real time via a Zigbee wireless protocol to a laptop placed in each individual's home where it is stored. Attached to the laptop is a Zigbee-to-USB-to-virtual-serial-data interface so that the laptop can receive the incoming data packets from the sensors distributed in the home. The laptop runs a custom Python script that parses the Zigbee data packets from the sensors and stores the data in its local file system. Periodically, the



**Figure 4-4. dwellSense system architecture. The dotted lines denote the components (motion sensing and real-time feedback display) that were added in dwellSense version 2.0 after the pilot deployment.**

laptop will connect to a remote server (located on the CMU campus) and upload the data collected for the past day via HTTP through a broadband connection or using its modem to dial up over the phone line. If the modem was used, the call was in the middle of the night while the individual was sleeping so that it would not interfere with the individual's use of the phone during the day.

The campus server runs the software to process the sensor data into human-level actions useful for assessing how a task is performed. The software consists of multiple Python modules to process data from each type of sensor (pillbox, phone, and coffeemaker). Each module utilizes a common custom library for interpreting sensor events and actions, and each module implements a different set of rules for interpreting the sequences of actions as user-initiated episodes and rating them based on rules determined by task analyses used in occupational therapy.

To ensure the integrity of the data and to check whether sensors are working correctly, a data integrity program is run on the latest day's worth of data for each participant. Sensors were designed to ping the laptop periodically to signal they are still powered. The program checks for missing sensor pings as an indication of a sensor needing its battery replaced. The schema for the data from each sensor also well-defined and the program checks whether the data from the sensor conforms within the schema. The program also checks looks for patterns of potential sensor failures such as switches that never close or missing DTMF codes. The program generates a report for each sensor and whether it appears to be working correctly or not and emails that report everyday to a researcher so that the researcher can schedule a visit to replace batteries or debug a malfunctioning sensor.

### 4.3 dwellSense Data Presentation

In addition to sensing, logging, and analyzing task performance, dwellSense also presents the sensor data and the results of its analysis to provide feedback to stakeholders about how the individual is functioning in their everyday lives. The sensor events are processed and episodes of medication taking, phone use, and coffee making are extracted and visualized via a web interface implemented in Javascript, CSS and HTML. The web interface shows graphs that show a short-term, detailed view of particular days (Figure 4-6) and graphs that show either a long-term view of the data over weeks or month (Figure 4-5).

dwellSense version 2.0 also includes a tablet-based display that provides real-time feedback. For details of this display,

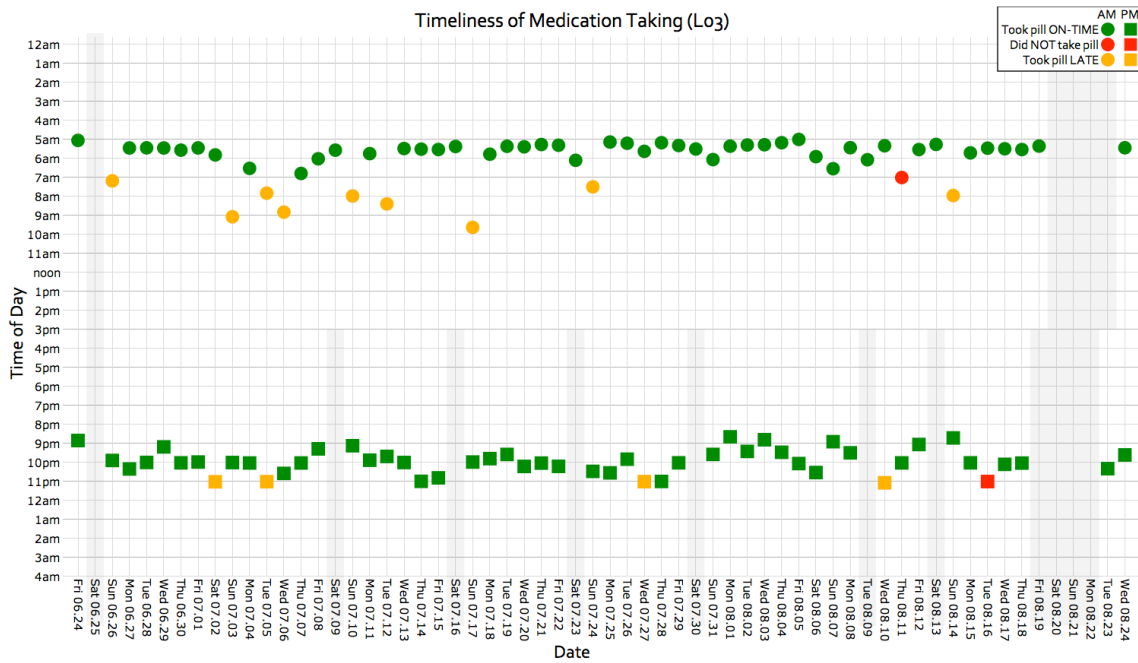


Figure 4-5. A long-term view of medication taking data.

see Chapter 6 about its design and deployment.

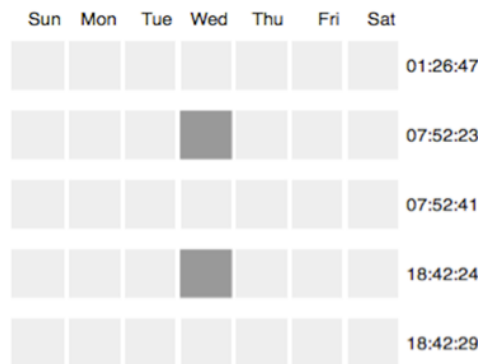
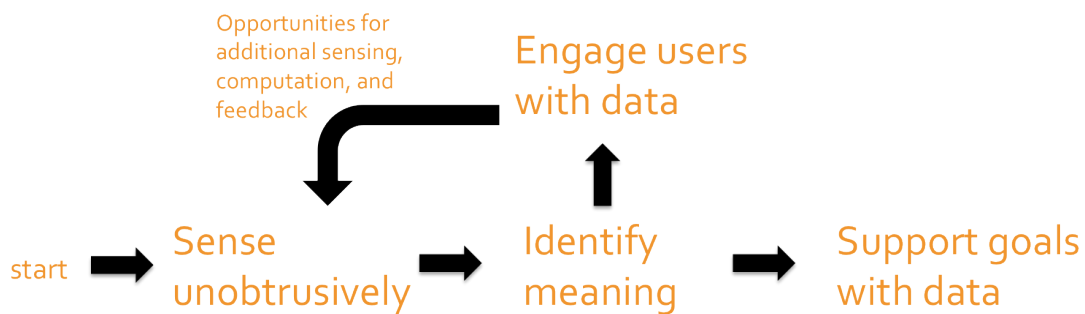


Figure 4-6. A short term view of medication taking data for one day. Each row represents the state of the pillbox doors. A darker shaded door indicates it is opened. This shows the user taking her morning pill at 7:52am and taking her evening pill at 6:42pm.

#### 4.4 User Reflective Design Process

dwellSense was developed using an iterative user-centered design process. A typical iterative user-centered design process used in human-computer interaction involves end users actively engaging with a new technology and providing use data as well as feedback on the benefits that it might provide to them. However, personal sensing systems like dwellSense, that passively record an individual's actions may not involve any direct engagement with the user; for example, with dwellSense, individuals simply carry on with their daily routines. Furthermore, the data collected by these systems often is less structured and thus can be more challenging to identify the meaningful information from the rich sensor streams. To design intelligent systems that not only record the actions that people perform but can also interpret these actions and highlight what actions the individual (or other stakeholders) may be interested in reviewing, designers need a method to understand the meaning users attach to the recorded content, reveal what limitations exist in interpreting the content, and identify opportunities for computation to support the user.

To address this need, this thesis introduces a specific type of user-centered process for personal sensing systems, called the User Reflective Design Process. The key step of this User Reflective Design Process is for the designer to engage users with their own data in order to understanding how they make sense of the data, what additional meaning they attach to the data that sensors may not be capturing, how they use the data to help them achieve their goals, and what limits their ability to make sense of the data.



**Figure 4-7 User Reflective Design Process for designing personal sensing systems.**

#### 4.4.1 Sensing Unobtrusively

The User Reflective Design Process for personal sensing system begins with selecting a set of parameters of the user's experience to sense unobtrusively using sensors, as most sensing systems do not want to unduly disturb the user in order to record her actions or her context. The data that the sensors collect can be quite large in volume and likely to be multidimensional. Exactly what data is important for understanding what the user is doing or what the user finds interesting to learn about is not always known beforehand; after all, a newly designed or new invented system is likely collecting a novel stream of data that the user has never before considered. Thus, the meaning of these newly collected data about the user's actions and context is likely unclear.

#### 4.4.2 Identify Meaning by Engaging Users with their own Data

In order to identify the meaning in the data, the key step in this design process is to engage users with their own data, in other words, to present users with a new aspect of themselves through the data. The designer can then observe how the user explores and investigates the data and what the user finds interesting. The additional meaning retrieved from the individual's memory of the original experience is triggered and can be expressed when the user is presented with these new data about themselves. Furthermore, the designer can also notice when individuals have difficult interpreting the data and when additional data about the context of particular episodes recorded in the data might be useful. The

demand for additional data provides an opportunity for additional forms of sensing to provide the additional explanatory data stream for the user. For example, consider the scenario where a designer is designing a system to track a user's step count as a measure of physical activity and as means to motivate individuals to be more physically active. After developing a prototype to record steps, the designer engages the user with data about her step counts over time. The user explores her step count data and wonders why on certain days she was more active than others. She realizes that the critical distinction is her location. On certain days, she works from home and is not physically active but when she is at work, she is very physically active. Thus, sensing the user's location would help not only make this relationship explicit for the user but it also allows the system to predict when the user might be more or less physically active and provide the appropriate prompts to encourage more activity. Engaging user with their data provides a way for designers to reveal the meaning attached to the content as well as identify new opportunities for sensing.

Engaging users with their data reveals the factors that lead to different types of behaviors. In a sense, the step of identifying meaning and factors for behaviors is akin to the concept of features selection in machine learning. In traditional feature selection, the features that are sensed by the system are ranked in how much they each alone predict the outcome state. Feature selection is limited to only the features that have already been sensed or recorded. However, by engaging the user with their data, a designer can consider the factors that exist in the mind of the user that have not been sensed by the system. Essentially, engaging users with their data is a first pass at feature selection to refine and iterate the system to include the most important or predictive features that valuable in subsequent machine learning.

#### **4.4.3 Supporting goals with data**

The final step in the User Reflective Design Process is to support the goals of the user by presenting them with the data or interventions inspired from the sensing system and the meaning that is attached to the data. Goals can vary from one application to another, but a common thread is that the system needs to be able to first understand the meaning of the data in order to know how to interpret it and provide the right type of feedback to the user.

#### **4.4.4 The User Reflection Design Process in Action**

We used the User Reflective Design Process first to develop another personal sensing system, MemExerciser (Lee & Dey, 2008), in which we identified the meaning in personal life logs of experiences to enhance the memory abilities of individuals with mild Alzheimer's disease. This same process was used to design dwellSense. The approach for dwellSense was to begin with a pilot deployment with a small number of participants and engage them with their data so that we could understand what they found interesting, what factors were critical for them to make sense of the data, and what limited their ability to make sense of the data. From these case studies, we discovered that individuals had difficulty knowing whether they were home or not home to take their medications, so in the next version of dwellSense (version 2.0), we added motion sensing to help disambiguate missed medications from being away from home. We also identified that they had difficulty remember the context of events from more than a week or two in the past, and thus in the next version of dwellSense (version 2.0), we added a real-time display to give users a more continuous awareness of their behaviors and eliminating the need to reflect far into the distant past. More details about dwellSense 2.0 and how it was deployed can be found in Chapter 6. In the next sections, the details of the pilot deployment of the dwellSense (version 1.0) are described as well as the case studies in which we engaged the pilot participants with their own data to gain insight into how they made sense of the data and how to refine the dwellSense system.

## **4.5 Pilot Deployment**

In order to evaluate the effectiveness and robustness of the sensors, we deployed the sensors in the homes of two community-dwelling older adults (age 76 and 82) who were living on their own and were recruited through a partnering organization specializing in caring for seniors (). Both participants were female and lived in their own apartment in a low-income senior high-rise building. Both individuals used pillboxes to manage their medications, regularly used their landline telephones, and made coffee with a coffee maker. Participant #1 (P1) is a 82 year-old retired nurse whose largest health issue is pain in the joints (knees and hands) resulting from arthritis. Participant #2 (P2) is a 76 year-old retired homemaker who has been diagnosed with Parkinson's disease. Most of P2's more debilitating symptoms of Parkinson's disease such as tremors and nerve problems are minimized by medication. However, P2 still suffers from difficulties in balancing when walking or standing as well as difficulty with her short-term memory. Participants' cognitive abilities were screened/tested using the Computer Assessment of Mild Cognitive Impairment (CAMCI) and the Digit Symbol Substitution Test. The CAMCI is typically used as a screening tool for Mild Cognitive Impairment (MCI). Based on CAMCI scores, P1 has a very low risk of MCI, whereas P2 has a moderate risk of MCI, though with a non-typical presentation of deficits. These results match the observations of researchers and also their self-reported medical conditions.

The pilot deployment began with a period of observation where researchers scheduled a visit with the participants to observe how they carried out different IADLs in their home. Both P1 and P2 were observed during one particular morning, and their routine included taking their morning pills, making breakfast, answering or making a phone call, and making a pot of coffee. These tasks were recorded and formed the basis for customizing the sensors to the



**Figure 4-8. Scenes from the pilot deployment. Top left: view of living room, with pillbox in foreground and laptop logging data behind the television. Top right: smart pillbox kept in the bed stand. Bottom left: instrumented coffee maker in the kitchen. Bottom right: sensor placed inconspicuously under the cushion of the easy chair.**

particular routines of each participant. For example, P1 keeps her coffee grounds in her refrigerator, so sensor was added to her refrigerator door to see if she opened it during the process of making coffee.

The sensors (pillbox, phone, and coffee) were installed gradually throughout the first three months of the deployment to allow time for researchers to focus on debugging each sensor before rolling out the next one. (Figure 10) The phased rollout also gave participants a chance to get accustomed to each new sensor in their home. For the most part, P1 and P2 reported that they did not find the sensors intrusive nor did it cause them to alter their normal routines. In the first month of the deployment, P1 and P2 were given a version of the augmented pillbox with four slots per day of the week. After approximately one and a half months, both P1 and P2 were given a version of the pillbox with only one slot per day. P1 and P2 were much more familiar and more comfortable with this type of pillbox. Moreover, the revised pillbox was more robust and reliable during everyday use. Also installed in each home was a laptop computer that logged the data transmitted wirelessly from the remote sensors. Each apartment was assigned a different network so as to prevent data from travelling from a sensor in one apartment to a laptop in another. The laptops were configured to upload their data securely to a server on campus every night using the modem and landline. Approximately every two weeks, a researcher scheduled a visit with the participants to check on the status of the sensors, replace batteries as necessary, and to retrieve a backup of the sensor logs stored on the laptop. Participants were offered a \$30 supermarket gift card every month for their participation. The pilot deployment began in March 2010 and is projected to last at least 12 months.

At roughly six months of data from the deployment, we conducted two case studies of how the two participants reflected on the data to support a correct awareness of their ability to use the phone and take their medications. This case study is described in the following section.



# 5

## Supporting Self-Reflection and Awareness of Functional Abilities

Individuals need to be aware of their own behaviors in order to know whether they need to fix a problem to improve their performance. Instrumental Activities of Daily Living (IADLs) performed everyday are difficult to keep track of because they have been routinized and the changes in performance may be subtle and vary from day to day (Kuriansky et al., 1976) (Little et al., 1986). Thus, sensors that can passively capture and create a log of how people carry out these mundane tasks can enable individuals to reflect on their own behaviors in a way not previously possible. In fact, in the concept validation study (Chapter 3), one of the main findings was that older adults (as well as their caregivers and doctors) found that viewing information about task performance over time to be potentially useful for understanding how an individual's abilities change over time. In this chapter, we investigate whether the data from a prototype embedded assessment system actually is able to capture and provide the information helpful for supporting self-reflection and self-awareness of one's abilities.

In Chapter 3, the overall concept of embedded assessment of wellness was validated through a formative concept validation study. With the concept having been initially validated, Chapter 4 describes the design of *dwellSense*, a prototype home sensing system to monitor how Instrumental Activities of Daily Living are carried out. The current chapter uses a real-world pilot deployment of *dwellSense* into the homes of two older adults to address research question RQ4, how is embedded assessment data *actually* used and whether the data can support a greater self-awareness of abilities. This chapter describes qualitative case studies (Lee & Dey, 2011) of two older adults and how they used the sensor data collected about their own behaviors to investigate and reflect on their abilities to maintain independence. From these case studies, we provide design recommendations on how to present personal data from home sensing systems to support reflection and sensemaking for older adults to increase awareness of their functional abilities as they age. We also describe the sensemaking process used by the participants to understand and interpret their data. We highlight opportunities for computational support to aid in the sensemaking process of personal sensor data. In the following sections, we first describe related work about reflection and health behavior, then we describe the specific details of the sensor deployments.

### 5.1 Background in Reflection and Models of Behavior Change

We investigate how older adults use sensor data to support reflection on their own health. Reflection can naturally lead to a new sense of self-awareness of one's abilities and then to an intention to act to change their current health state to a more desirable state. For example, by reflecting on how often she takes her medications on time, an individual can become more self-aware of how often she takes them too late, which can motivate her to take actions to take them more on time, such simply paying more attention to her habits or to set a reminder to take her medications on time. Thus, reflecting on information can be one of the first steps towards increased self-awareness and positive behavior change.



Self-tracking and reflection have long been used as components for health behavior change. A number of health behavior change models have been developed to explain how individuals make decisions about the actions they take to improve their health. We focus on following three particular models that include features that can be influenced by self-reflection: Theory of Planned Behavior (Ajzen, 1991), Health Belief Model (Janz & Becker, 1974), and Transtheoretical Model of Behavior Change (Prochaska & DiClemente, 1983) because these models.

The Theory of Planned Behavior (Ajzen 1991) is based on the Theory of Reasoned Action (Fishbein 1979). The Theory of Reasoned Action makes the rather straightforward claim that the most important determinant of adopting a behavior is the individual's intentions towards that behavior. An individual's intention towards a behavior is determined by two factors 1) the attitude towards that behavior and 2) the subjective norms associated with that behavior. An individual has a positive attitude towards the behavior if she thinks that the behavior will bring about a beneficial outcome. Reflection on feedback from embedded assessment systems is not likely to directly influence the main factors in the Theory of Reasoned Action (the individual's attitude towards an action or influence the subjective norms associated with the behavior). However, the Theory of Reasoned Action assumes that the individual is under a condition of relatively high volitional control, that is, the individual can perform actions without any obstacles. On the other hand, the Theory of Planned Behavior adds an additional component, the concept of perceived behavioral control, to accommodate conditions of low volitional control. An individual will work harder at performing a behavior if the perceived behavioral control is high. Perceived behavioral control is determined by the perceived existence of facilitating factors or barriers to performing the behavior. Reflecting on feedback from embedded assessment systems can improve self-awareness of behaviors, which can lead to a greater sense of control over their own behaviors. Thus, the Theory of Planned Behavior predicts that embedded assessment systems can play a role in highlighting mistakes or inefficiencies in how individuals perform IADLs so that individuals can be a better sense of control to address the mistakes and maintain a high level of adequacy in how they perform IADLs.

The Health Belief Model (Janz & Becker, 1974) has its roots in public health research and was developed to explain why certain health practices (such as tuberculosis screenings) were more easily adopted than others. When applied to the specific concern of healthy aging, the model claims that individuals will take actions to improve their functional abilities if they believe that 1) they are likely to experience declines (perceived susceptibility), 2) the consequences of declines is sufficiently severe (perceived severity), 3) the actions to improve their functional abilities are adequately beneficial (perceived benefits), 4) the barriers to taking that action is low, 5) there are adequate indicators of when/how to take the action (cues to action), and 6) whether they feel confident in their ability to take the action (self-efficacy). Reflecting on data from embedded assessment systems can potentially influence three of the six factors in the model: perceive susceptibility, cues to action, and self-efficacy. For example, in the case when individuals overestimate their medication adherence, seeing an objective record of how often they actually miss taking their medications can raise their awareness that their level of adherence is lower than as they had expected. Reflection on the sensor data not only shows that they are "susceptible" to forgetting their medications but also provides the individual with a "cue to action" to actually do something to improve. Reflecting on sensor data may also support an individual's self-efficacy, that is, to make the individual feel more confident that they can carry out IADLs even if the individual faces barriers or doubt. For example, if the data shows that the individual is doing as well as she had thought, the data would reinforce her perception of her abilities to take her medications correctly. Thus, the Health Belief Model provides a framework in which reflection on embedded assessment data can be used to affect how individuals think about their health and health actions, particularly in increasing perceiving susceptibility, cues to action, and self-efficacy.

Rather than simply focusing on the factors that lead to behavior change, the Transtheoretical Model (TTM) of Behavior Change (Prochaska & DiClemente, 1983) is a stage-based model that explains the precursors to action, how actions might be carried out, and the stages following adopting a particular behavior. The stages include: pre-contemplation, contemplation, preparation, action, maintenance, and termination. Personalized feedback from embedded assessment systems can help individuals across all these stages. In the specific case studies covered in this

chapter, the role of reflecting on data from embedded assessments systems will be explored for moving individuals from a pre-contemplation stage to the contemplation and preparation stages. In other words, reflecting on sensor data about medication adherence will raise awareness of an issue for those not yet contemplating making any changes in their medication-taking routines and move them into the phases in which they consider making changes to their routine and preparing to change their routines by stating intentions and making personal commitments to change. In fact, the TTM also includes a number of processes that explain why individuals adopt changes in their routines. Some of the processes at work in this study include consciousness raising (akin to perceived susceptibility in the Health Belief Model), dramatic relief (contrasting their own self-perception of their actions with the account of their actions captured in the objective sensor log), self-reevaluation (helping individuals realize that performing IADLs correctly is important to them), and stimulus control (adding appropriate cues to the environment to facilitate improved performance). The TTM provides a framework that can explain how reflecting on sensor data can help move individuals from not considering any changes to becoming aware of an issue to eventually making a change to improve their IADL performance.

The benefits of reflection have also been explored in the human-computer interaction research community. Li et al. (2010) discuss a framework for how users deal with information collected about themselves (such as how they carry out everyday activities). We analyze the reflection and action stages of that framework from the perspective of the older adult for this paper. Reflection has shown to be effective for people with diabetes, for improving their sense of control and improved diet outcomes (Mamykina et al., 2008). Reflection and the ensuing awareness of healthy and unhealthy behaviors for cardiac rehabilitation patients were found to be important for successful recovery (Maitland & Chalmers 2010). Tracking and feedback on physical activity supports people's ability to be physically active (Consolvo et al., 2008). Reflecting on behaviors has also been found to be useful for behavior change in other domains such as helping individuals reduce their water consumption (Erickson et al., 2012).

With the potential benefits of reflection understood from health theory and related HCI research, we investigate whether (and how) reflection specifically on sensor data about how everyday tasks are performed affects an individual's self-awareness of their abilities as well as their intentions to change and the actual changes they were able to implement. In the next section, we describe the case study approach we used to engage users with their data and to understand how they were able to reflect on the data to support an accurate self-awareness.

## 5.2 Case Study Methodology

A case study methodology was used to analyze the impact of sensing and reflection on the data collected by task-based sensors on these two individuals. The goal of using a case study methodology was to dive deeply into the rich details of the experiences of these two older adults, instead of the larger user testing approach that relies heavily on inferential statistics to draw conclusions that could apply more widely to a large audience. Following the User Reflective Design Framework, after sensing the user's actions unobtrusively, it is important to engage the user with their own data in order to understand what meaning they attach to it and identify the sensemaking process they use when interpreting the data, and thus, a case study approach was an effective way of engaging deeply with individuals. The case study methods we employed include in-depth semi-structured interviews, structured coding of qualitative data from interview sessions, and quantitative analyses that focus on change within a single individual's behaviors over time.

We recruited two older women who lived alone in their apartments through a professional connection with the management of a low-income senior apartment building. These two individuals represent a population that may benefit the most from monitoring technologies as they lack care support from a spouse or a daily caregiver, however the individuals differ in their overall cognitive and functional abilities.

Participant #1 (P1) (age 81) is a retired nurse, who is aging successfully. P1 has set in place the routines that ensure her safety in her medication taking. She prides herself in keeping up to date with the latest news and politics and overall

has a generally accurate impression of her own abilities. She has mobility issues that make it difficult for her to walk up and down stairs and for long distances. Based on psychometric testing (Computer Assessment of Mild Cognitive Impairment) (Saxton et al., 2009), P1 exhibits a slightly higher level of ability in attention, memory, and executive functioning compared to her peers of similar age and education level.

Participant #2 (P2) (age 77) is an retired homemaker who once struggled with moderate symptoms of Parkinson's disease. The disease once took away P2's ability to concentrate and maintain control her limbs without severe tremors. However, she has recently started to take medication that was effective at eliminating most of her Parkinson's symptoms, resulting in her being able to walk, write, and concentrate moderately well. She admits to having a memory problem due to the lingering symptoms of the disease, but generally believes that she is pretty aware and aging well. She also does not consider herself to be an organized person, preferring clutter to putting things away. Based on psychometric testing with the CAMCI, P2 exhibits a lower level of cognitive ability across attention, memory, and executive functioning compared to her peers of similar age and education level.

In order to understand the impact of sensing and reflecting on the sensor data, we engaged the two pilot participants (individually) with four months of their own data using a retrospective reflection session. During these sessions, conducted qualitative semi-structured interview sessions with each participant in which we showed them data about their pill taking and phone use tasks to allow them to reflect on their own abilities to stay independent. The interview consisted of a researcher-guided training phase (to ensure the participants could understand the visualizations) followed by a participant-guided exploration phase. In the training phase, the researcher first showed the participant visualizations of data from a short time frame (for example, from the day or week immediately preceding the interview) and then explained what the marks, axes, and dimensions represented. The researcher refrained from making any interpretations of the data (*e.g.*, "you made a mistake here" or "you missed your pills a lot in the past month"). The researcher then tested the participant's understanding by having her describe a visualization of another day's data.

After adequately demonstrating their comprehension, the participant was allowed to guide what level of detail of the data they wanted to see. We used a think-aloud study protocol to allow the participant to express her thoughts and reflections during the interview. To understand any change in awareness, we asked the participant to assess her own pill taking and phone use abilities before and after looking at the data. To understand the participant's intent for future actions, we also asked questions such as "Would you do anything differently because of what you are seeing, or not?" The interviews were video recorded. The video was segmented into units of analysis that consisted of a participant's single thought or stream of related thoughts. These segments were analyzed using an axial coding scheme where coded segments were grouped into successively higher order categories resulting in emergent themes. In the following sections, we describe the sensors deployed in their homes, the data visualizations, and then followed by the results of the analysis.

## 5.3 Sensor Deployment and Data

The dwellSense sensor suite consisting of the smart pillbox, phone sensors, and augmented coffeemaker was deployed for over 18 months in the apartments of community-dwelling two older adults, who were living alone. We encouraged participants to carry on with their lives as normal and avoid being extra careful in the performance of their everyday activities just because their activity was being tracked or because they were participating in a study. The length of the study helped to ensure that participants behaved naturally.

### 5.3.1 Smart Pillbox

In the first two months, we continually revised and reintroduced more robust versions of the sensors. The first version of the pillbox was modeled after a commonly used pillbox that had multiple (four) slots for each day of the week. This style of pillbox could accommodate the majority of the population because it could track pills that are taken multiple

times per day. However, this first design of the pillbox proved to be difficult to maintain because it required 28 (4 slots/day x 7 days a week) independent closure switches for each pillbox door to be mounted on the pillbox, which presented us with 28 individual points of failure that could compromise the accuracy of sensing. Furthermore, the design required a more sophisticated microcontroller with at least 28 separate inputs that drew more battery power and required more frequent visits to maintain. After testing this 28-slot pillbox for a few weeks, we realized it was not a scalable solution. The actual pillboxes that the pilot study participants used had only one slot per day (7 slots total). Thus, we scaled down the prototype to a 7-day pillbox. This revised version required fewer parts, a simpler microprocessor, consumed less power, was packaged better, and most importantly, matched the type of pillbox that the pilot study participants used. To minimize the potential for disruptions to their pre-established routines, we replaced their pillbox with our instrumented pillbox that had the exact same size, lettering, and shape as their existing pillbox. The pilot study participants were much happier with the new box as it was more similar to what they were used to (compared to our first version) and had a more solid feel. The 28-slot pillbox design was functional, but they also felt the box was designed for people “more sick” than they were because they did not have to take medications more than twice a day. Pilot participant P1 normally used only one pillbox in which she placed both her morning and evening pills, so she was given one pillbox to replace her existing one. Pilot participant P2 wanted to track only her morning medications but wanted more than one box because her daughter (a trained nurse) refilled her pillboxes once every two weeks, so she was given two boxes.

### 5.3.2 Phone Sensor

The phone sensor was installed on the phone line. The phone line was split at the wall outlet (or cable box) into three identical lines. One line went to the user’s phone to maintain normal phone service, another line went to the laptop’s modem that logged incoming calls and Caller ID if the resident had that service. The third line went into the phone sensor. The phone sensor was powered off the mains with a power adapter. Initially, the mains AC/DC power adapter (colloquially called a “wall wart”) introduced some white noise on the phone line that annoyed the participants. After testing various power adapters that had different amounts of shielding and circuitry, we found that higher voltage (over 9 volts) adapters with more robust transformers and filters introduced the least amount of interference and so three months into the study, the phone circuits were upgraded with new power adapters. One pilot study participant P2 received her phone service through her cable service provider (Comcast) and thus had a digital system. The phone circuit could not decode the DTMF signals from this digital box. (In subsequent deployments in homes with Comcast-based phone services, the phone circuit was able to decode the DTMF signals without problems, as long as the phone line was split upstream of any devices connected to the cable box. It is unclear whether other devices connected to P2’s cable box interfered with the DTMF decoding or whether her particular cable box somehow made it difficult to decode the DTMF signals.)

### 5.3.3 Instrumented Coffeemaker

The instrumented coffeemaker was introduced approximately 3 months into the deployment after the pillbox and phone sensors were introduced. Our augmented coffeemaker replaced their existing coffeemaker that typically had a simpler design (with only one button to turn on the machine) than the coffeemaker we introduced. The experimenter explained and demonstrated how to operate the new coffeemaker and tested the participant’s understanding by having them demonstrate how to use the coffeemaker to make coffee. To make sure the new coffeemaker did not unduly interfere with their coffee-making routine, their routine was videotaped once with their old coffeemaker and once with the new coffeemaker. The videos did not show any significant changes to the routine due to the difference in operation between the old and new coffeemaker. Participants rarely used the “advanced” functions of the new coffeemaker not found on their old coffeemaker such as the timer, 3-cup mode, and digital clock (they did not bother or know how to reset the clock when the coffeemaker was unplugged). In this pilot deployment, some of the switches added onto the

coffeemaker such as those to track the carafe placement and reservoir door did not stay properly attached over time due to the heat, moisture, and general wear and tear of normal use. The robustness of subsequent versions of the coffeemaker were improved by using more durable (yet non-toxic and non-reactive) adhesives and improving the shape of the physical mounting points of the switches to relieve the stress on particularly sensitive joints. However, even after these modifications, the switch that detected whether the carafe was in place was still prone to falling off. Thus we were careful about ignoring spurious carafe switch events when the switch was found to have fallen off. Due to the repeated problems with the carafe switch, the coffee making data was not used in the analysis for this pilot study.

Participants reported that they enjoyed using the “new” coffeemaker because it made very good coffee. The deployment and acceptance of the coffeemaker is a good example of how people are willing to adopt a new technology if it provides a particular value to them, in this case, making good coffee. From a research or clinical perspective, the important functionality of the instrumented coffeemaker is the sensing, but for the consumer/patient/user, the important functionality is its ability to make coffee. Sensing alone does not provide direct immediate value to the user (particularly in this case because we wanted the sensing to be as unobtrusive as possible), but the fact that it made good coffee was enough to make them happy to use the machine, which has the side effect of logging data about their abilities.

#### 5.3.4 Deployment Data

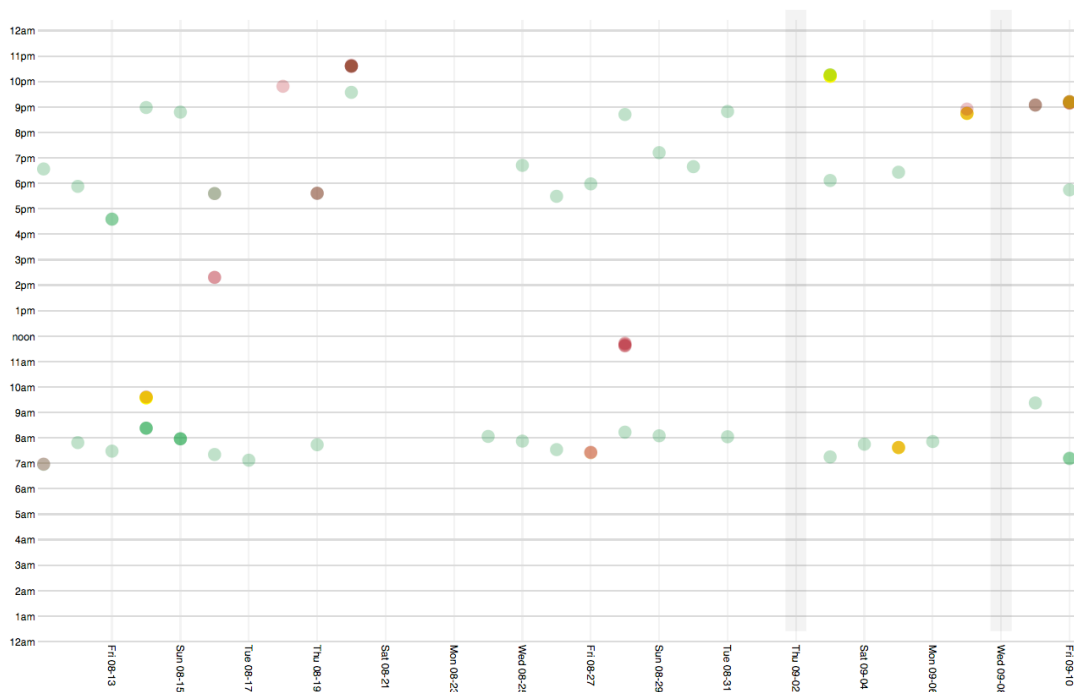
Throughout the deployment, a researcher visited the apartments every two weeks to replace batteries, debug sensors, and ensure that the sensors were not getting in the participant’s way. The sensors had fairly well defined points of failure (for example, each switch and its connections) that could be easily identified either remotely or during visits. The accuracy of the sensors was verified through a combination of lab testing, field testing, and observations of use during bi-weekly visits to the apartments. Upon discovering a faulty sensor during a visit, the researcher would flag the data up through the date of the last known time the sensor was known to be functioning correctly. In subsequent data analysis, the data flagged as inaccurate was ignored. On a few occasions we were unable to collect data for one or more consecutive days due to a power loss or error in the logging script. These days were flagged to be ignored in subsequent data analysis.

The data from first two months (March and April 2010) of the deployment were not included in the data analysis because during these months, we continually revised and reintroduced more robust versions of the pillbox and phone sensors. Six months into the study in September 2010, we performed the first informational intervention (a reflection session) with the pilot participants to investigate the effect reflecting on the task performance data would have on awareness and behavior (RQ3). At this 6-month point, we had approximately four months of validated pill taking and phone use data.

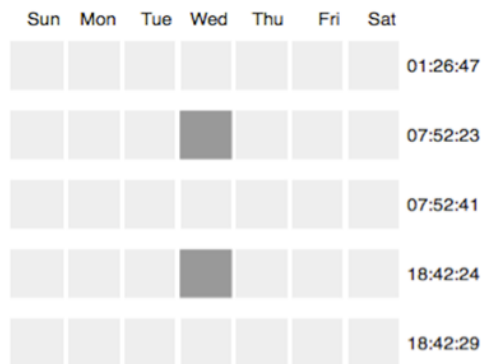
In these four months (125 days) from May 2010 to September 2010, P1 had an opportunity to take her pills (both morning and evening combined) in 250 instances. Of these instances, 86 instances were ignored because either the system was down due to power outages, sensor failure, or software bug (37 instances) or the participant was away from her home (49 instances). This left us with 164 instances of logged pill-taking data for P1 in the four months prior to the retrospective reflection session. P2 had an opportunity to take her morning pills 121 times during this period (her evening pills were not tracked because she wanted to use a very different looking case for them). Of these instances, 43 instances were ignored because the system was down. She was not away from her home during this period. This leaves 78 instances of logged pill-taking data for P2 in the four months prior to the retrospective reflection session. We visualized the data using charts, described in the following section.

### 5.4 Data Visualizations

For both pill taking and phone use, a high-level, long term view showing performance over weeks or months and a low-level, short term view showing the specific details about the task performance for one day were available.



**Figure 5-2. Long-term visualization of pill taking. A dot represents an opening of a pillbox door. The y-axis is the time of day, and the x-axis is the date. A green color indicates that door for that day of week was opened, otherwise the dot is colored red. A yellow color indicates the door was not closed after it was opened.**



**Figure 5-1. Short-term pill taking visualization showing pillbox door states and times for a particular day. The user opened the Wednesday door once in the morning and once in the evening.**

For pill taking, the long-term visualization (Figure 5-2) showed the date and time of every instance when a pillbox door was opened over a user-configurable time span that ranged from a week to multiple months. Each mark’s color represents whether the door was left open until the next pill taking episode (yellow) and whether the pillbox door’s label matched (green) or did not match (red) the current day of the week. The green color represents the most typical “correct” sequence of pill taking, that is, opening the correct pillbox door and closing it within a reasonable amount of time, before opening another one. Dots from multiple door openings can overlap and appear in darker shades. A grayed out column represents a day that we were not able to collect data due to a system problem. The short-term visualization (Figure 5-1) showed how the pillbox doors were opened throughout a particular day. For phone use, the long-term visualization showed the date and time of every outgoing phone call over a user-configurable time span of a week to

multiple months. Each mark was colored green if the call was not misdialled or colored red if misdialled. The metric used for marking whether a call was misdialled was if two numbers were dialed within a minute of each other and also had 70% of the digits in the first number overlap with the digits in the second number. In the short-term visualization (Figure 5-3), participants were shown the time, length, and number of every phone call made on a particular day. Another long-term visualization of phone use (Figure 5-4) included the total number of minutes spent on the phone for each day over the course of a time span ranging from a week to a few months.

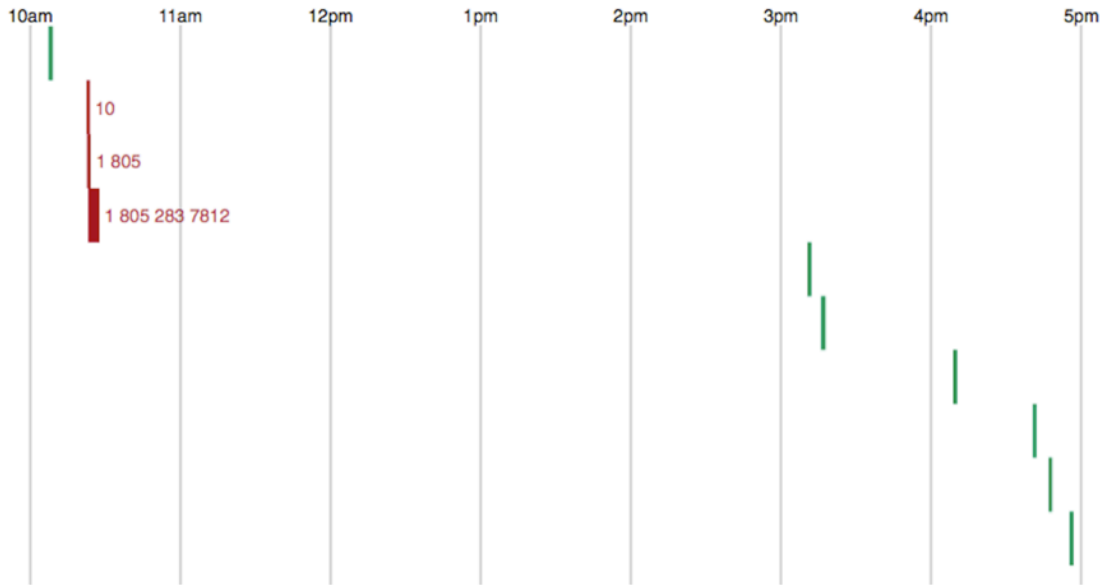


Figure 5-3. Detailed, short-term view of phone calls on a particular day, showing how the user at approximately 10:25am misdialed twice and successfully dialed the number on the third attempt. The horizontal width of the bars represent the length of the call. A bar is marked in red if it is part of an episode of misdialing. A green bar represents either a correctly dialed number or an incoming call if no number is displayed next to it.

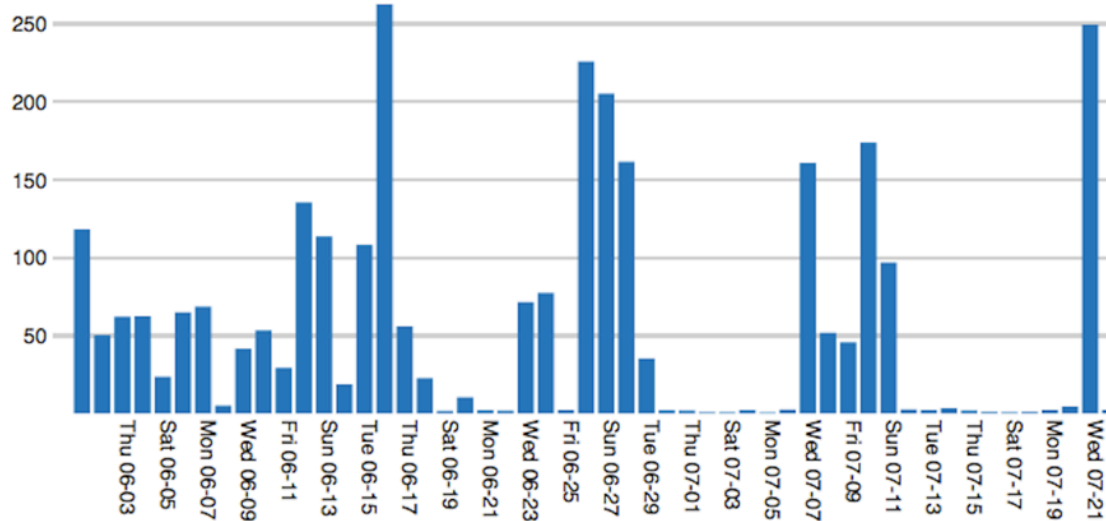


Figure 5-4. Long-term visualization showing the total number of minutes spent on the phone for each day.

## 5.5 Interacting with the Data

Based on participants' interactions with the visualizations of the task performance data we showed, we observed how they engaged with the data, what they paid the most attention to, and what other information they wanted to help interpret the data. Participants engaged in three different behaviors: looking for their mistakes in the data, investigating



and attempting to explain away these mistakes, and diving down into the details of their task performance to verify their explanations.

### 5.5.1 Looking for Anomalies/Mistakes

When presented with the visualizations of the data, participants attempted to find any mistakes or anomalies in their own behavior. The visualization (Figure 5-2) that showed long-term pill-taking performance for a week or more highlighted only positive examples of pill taking. The graph contains a dot for every instance a pillbox door was opened. It did not contain, for example, a marking to show when the pillbox was not opened that day. The only visual indication of a missed day is the rather inconspicuous between-dot whitespace, which can be difficult to see especially because the dots do not line up closely with each other. Nonetheless, we observed that P1 and P2 did not focus on the positive examples of pill-taking but rather went to the effort of going through the whitespaces on the graph and seeing if they lined up with a particular day to find instances of missed pills. Likewise for the phone use data, the two study participants looked at the outliers, the data points that showed a particularly high number of calls on a given day. It was typically these outliers that conflicted the most with the participant's self-perception of their behaviors. For example, P1 said she does not use her home telephone very much, typically making fewer than five calls per day. However, when she saw she made over 20 calls on one particular day, she was surprised to see that behavior. P1 also noticed that she spent well over two hours on the phone on one day, which was much more than she had originally reported before she saw the data. Thus, the concept of an anomaly can be either 1) a value that is largely different from the majority (and can be calculated statistically) or 2) a value that may be similar to the other values in the data set but is inconsistent or incongruent with the individual's expectations. Making these anomalies easy to spot is important when engaging people with data about their own behaviors.

### 5.5.2 Generating Explanations

After identifying the anomalies in their performance in the data (such as missed pills, opening the pillbox doors incorrectly, or unusually long phone calls), participants immediately tried to think of reasons why the anomalies might have occurred. Finding a reasonable explanation, other than they made a mistake, was important for the participants to know whether they were having a problem or not. Participants used a number of information sources in addition to the pill-taking and phone use visualizations including their memory, routines, and a wall calendar.

**Explaining with Personal Memory** The participants' first natural reaction to seeing an anomaly in the data was to think back to the events of that particular day or week to find an extenuating circumstance to explain the unusual behavior. P1 was particularly good at remembering recent significant events that helped to explain anomalies in her pill taking. For example, upon noticing that many of the pillbox doors were opened out of sequence a few days earlier, she recalled that she received a new supply of heart medication and was placing them into her pillbox to fill out rest the week. In contrast, P2 was less able to recall the details of recent events. When the sensor data showed that she did not interact with the pillbox three days ago, she tried to recall what she did that day that might have explained this error. Even though P1 was able to recall recent experiences adequately, both P1 and P2 eventually had difficulty relying solely on their memory to recall the personal experiences important for explaining anomalies in their behaviors and had to resort to other means such as their routines.

**Explaining with Routines** Without an explicit recollection of an event or circumstance that would explain why an anomaly such as a missed pill or a misdialled telephone call might have occurred, participants thought about their routines and whether the anomaly might fit within one of their many variations on their routines. For example, when noticing a few instances of taking her morning pills much later (at 9am) than she normally would have (7am), P1 reasoned that she must have slept in on those mornings. P2 was less able to draw on specific memories of events that might explain anomalies in her pill taking. When noticing in the data that she took her pills very late at night only two

days ago, P2 reflected on one of her routine behaviors that she often falls asleep on the couch during the evening which accounts for the lateness of the pill taking.

**Explaining with the Calendar** When unsuccessful in finding either a specific circumstance or routine to explain an anomaly in the task performance, P1 referred to her calendar for hints about what happened on the day(s) of the anomaly. The most common explanation P1 used to explain days with no pillbox activity was that she was away from her apartment, which she often recorded on her wall calendar. For example, P1 went to stay with her daughter for a few days in the second month of the study. While attempting to explain why there was no pillbox activity for that weekend, she noticed that her grandson's name was written in her calendar for that weekend and realized that he was returning from the Army and was home for a visit.

### 5.5.3 Confirming with Details

dwelSense could capture task performance at a fairly fine level of detail (*e.g.*, the specific time that a particular pillbox door was opened and every digit dialed for a particular phone call). We presented both a long-term view of the data usually spanning weeks or months and also allowed the participants to review the specific details of each phone or pill-taking episode in a given day. Both participants were able to understand the detailed information after it was explained by the researcher, but they expressed different interest in the detailed information. P1 was interested in knowing the details of when each pillbox door was opened and closed to make sure that she took her pill that day. She also used the details to confirm her explanations. For example, to explain why the log showed that she did not take her medications on Friday night, she remembered that she went to her nephew's party that evening and took her pills with her. She looked at the details of her pillbox interactions that day and saw that it took her 20 seconds in the morning, much longer than normal because she was moving her evening pills into her travel container.

## 5.6 Attitudinal Reactions to the Data

In addition to observing how the individuals reflected on the data and made sense of it to themselves, we found that the sensor data about their everyday performance provided the ground truth by which they could reaffirm or gain an accurate awareness of their functional abilities. After realizing the inconsistency in their routines through exploring the data, both individuals intended to “do something about it” and be more consistent to ensure safety. The participants also expressed opinions about sharing the information with members of their care network.

### 5.6.1 Supporting Accurate Awareness

Awareness of changes in functional abilities is key for successful aging, as it provides opportunities for the individual to make the appropriate adaptations to ensure she remains functional and avoid situations that threaten her safety. Prior to viewing any of the sensor data, both participants P1 and P2 were confident that they performed their pill taking regularly and almost never missed their medications. However, P1 and P2 differ in the accuracy of their confidence in their pill taking routine. P1's confidence in her routine actually matches her functional abilities. However, P2's pill taking routine is more erratic, showing instances of isolated days where she did not open the pillbox at all or opened up a pillbox door that did not match the day of the week. As a result, the sensor data had very different impacts on P1 and P2. For P1, the data provided a means to affirm her accurate confidence in her pill taking, whereas for P2 the data was useful for re-assessing her own (over-)confidence in her pill taking routine. Even though P1's awareness of her abilities was relatively accurate, she was initially surprised at the variability of when she took her pills during the day and how often she misdialled the telephone. Her feelings of surprise quickly transitioned to acknowledgement, as she was able to explain the variability and the number of misdials by accounting for them in natural variations in her routines, as described in the Generating Explanations section above. P2, on the other hand, had her confidence challenged when she saw the inconsistency and variability in her pill taking data.

### 5.6.2 Intention to be More Consistent

Based on a newly gained awareness of their abilities to take the right pills at the right time and correctly make telephone calls, the participants resolved to be more consistent in their routines to ensure their safety and adherence to their medications. P1, despite her relatively accurate awareness of her pill taking routine, decided she wanted to be more consistent in what time of day she takes her pills. A more consistent routine would make her feel more confident that she took them and would help her to ingrain in her brain a successful habit that will last into the future. Talking about how she will continue with her routine to move her pills from the box to the visible bowl on her counter, she said "I have to get more consistent in opening that box and putting them in [the bowl]. See, I'm so used to that routine. I keep that little black bowl on my counter for that reason." After seeing how often she was misdialing the phone, she said she wanted to buy a new phone with buttons that are easier to press so that she can be more consistent in her phone dialing and figure out whether the problem was caused by her old phone or her old arthritic fingers.

P2, after seeing the large variability and the unexplainable instances of missed pills, resolved to be more consistent and to pay more attention to her pill taking. She equated her poor pill taking performance with "messing with [her] life" because she currently is taking a "miracle" drug for Parkinson's disease and she certainly does not want to regress to a point where the Parkinson's symptoms re-emerge. She said, "I'm gonna set a time for me for my pills and try to adhere to that, say at the 11 o'clock news." She even began to question her evening medication taking routine (which is not monitored by our pillbox because her evening pillbox is a different type) and whether she was taking that properly. She considered whether or not to keep a written diary where she would check off everyday whether she performed important tasks like taking her pills.

### 5.6.3 Desire to Share Data and Potential for Misinterpretation

Both participants wanted to share their information with their family members so that others could know how well they are able to remain independent. P2 said her daughters, particularly the one who is a nurse, would want to see the data and help her mother fix any problems that might come up. Similar to previous findings (Hayes et al., 2006), participants wanted to keep their information private to just their own family, close friends/helpers, and their doctors. With sharing comes the additional potential for misinterpretation. P1 was concerned that others who would look at the data might not be able to determine whether the anomalies in the data (*e.g.*, missed or late pills or misdialing telephone) are benign or a cause of concern. She is able to look at the graphs and figure out whether the apparent missed pills are explained by being out of town or taken in some other acceptable way.

## 5.7 Behavioral Reactions to the Data

In addition to supporting an accurate awareness of their abilities, reflecting on the data also resulted in the two pilot study participants resolving to make a change to be more consistent in their medication taking routines. P2, in particular, was convinced by the objective data that she was not taking her medications as well as she should have been. She intended to not miss taking her morning pills and to set a particular time of day to take her medications. In the interview she thought about whether she would keep a diary as a record of her pill taking to remind herself if she has not taken her medications. P1 even though her self-perception was fairly consistent with the sensor data, she still wanted to improve (as it was fitting her personality as a former nurse to improve and remain independent). In the subsequent months, the sensors continued to record the task behaviors of the two pilot study participants. From these data, we identified whether the two pilot study participants were able to follow through with their intentions to improve their medication routines and telephone use.

### 5.7.1 Analysis of Medication Behaviors

This analysis of behavior after the reflection session considers the four months of actions following the reflection session. In these four months (122 days) from September 2010 to December 2010, P1 had an opportunity to take her pills (both morning and evening combined) in 244 instances. Of these instances, 82 instances were ignored because either the system was down due to power outages, sensor failure, or software bug (53 instances) or the participant was away from her home (29 instances). This left us with 162 instances of logged pill-taking data for P1 in the four months following the retrospective reflection session. P2 had an opportunity to take her morning pills 122 times during this period (her evening pills were not tracked). Of these instances, 51 instances were ignored because either the system was down in 41 instances or she was away for her home in 10 instances, leaving 73 instances of logged pill-taking data for P2 in the four months following the retrospective reflection session.

The four features of the medication taking routine considered in this analysis are: adherence, promptness, correctness, and the variance in the time of day the pills were taken.

Medication adherence characterizes how often pills are taken and is calculated as a percentage by summing all the pill-taking instances when the individual has taken their pills and dividing the sum by the total number of instances in a given time period. Contrary other studies, the percentage of medication adherence is not being used to classify individuals into categories such as “high adherence” or “low adherence” but instead is used as a linear measure of how adherent the individual is.

Promptness characterizes whether the pills are taken before the user-specified time-of-day threshold for late pills. We asked each participant to specify a time of day that they considered later than they normally would take them for their morning and evening pills. The measure of promptness was calculated as a percentage by summing the number of pill-taking instances before the late time and dividing it by the total number of instances over a given period of time.

Correctness characterizes whether the individual took the pills assigned for that particular day. A pill-taking instance was considered correct if the individual opened up the pillbox door that matched the current day of the week. Users could open up any other doors *in addition to* the door that matched the current day and the instance would still be considered correct. An instance would be marked as incorrect only if the user did not open the door that matched the current day of the week but opened other door(s) instead.

Variance in the time day measures how the time of day that medications were taken varied from one day to another within a given period of time. In the case for participants who took both morning and evening pills, we calculated the variance for each their morning and evening pills and summed the variances. The variance was calculated by converting the time of day to a decimal number between 0 and 24 and then taking the variance of the numeric times over a given time period.

For these case studies, statistical approaches appropriate for single-subject analyses were used instead of the inferential statistics typically used in studies with more than one subject. One common technique used in case studies is visual inspection where the data is presented visually in a graph in order to be evaluated subjectively for patterns of change. Graphs of P1's and P2's performance on taking their medication are shown in **Error! Reference source not found.** through Figure 5-12. To formalize the process of identifying change, another more objective, quantitative technique for identifying change within a single individual is the two-standard deviation band “2SD” technique (Krishef 1991). To use this technique, the mean and standard deviation are calculated for the data from the baseline phase, and a “band” is calculated bounded below by the mean minus two times the standard deviation and bounded above by the mean plus two times the standard deviation. This band is applied to the values in the intervention phase and if there are two consecutive values that occur outside the range of the two-standard deviation band, then the conclusion is that there has been a significant change in data. The intuition is that if there was no difference between the two phases, then the values in the intervention phase would also fall within the 2SD band 95% of the time. Thus, two consecutive points in

the intervention phase that exist outside the 2SD band would be fairly unlikely (at most 5%) given the variation found in the baseline phase. Although there sometimes can be disagreements between the conclusions reached via visual inspection and the 2SD technique, in the case studies presented in this study, the reader will likely reach the same conclusions using both techniques as the patterns are fairly apparent.

The following graphs show Participant P1's adherence, promptness, correctness, and variance in time taken before and after reflecting on the sensor data from dwellSense in August 2010. An interpretation of the graphs is included in the subsequent paragraph.

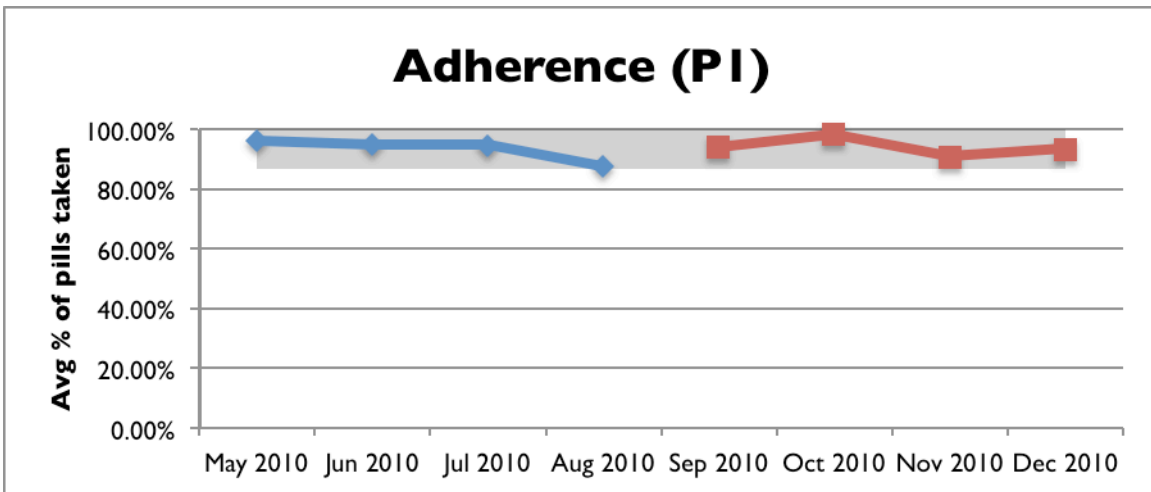


Figure 5-5. Participant P1's medication adherence rates before and after reflecting on her data in Aug 2010. The rates following the reflection session fall within the two standard deviation band (gray) of the pre-reflection rates. The data reflection does not appear to have affected her (already good) adherence behavior.

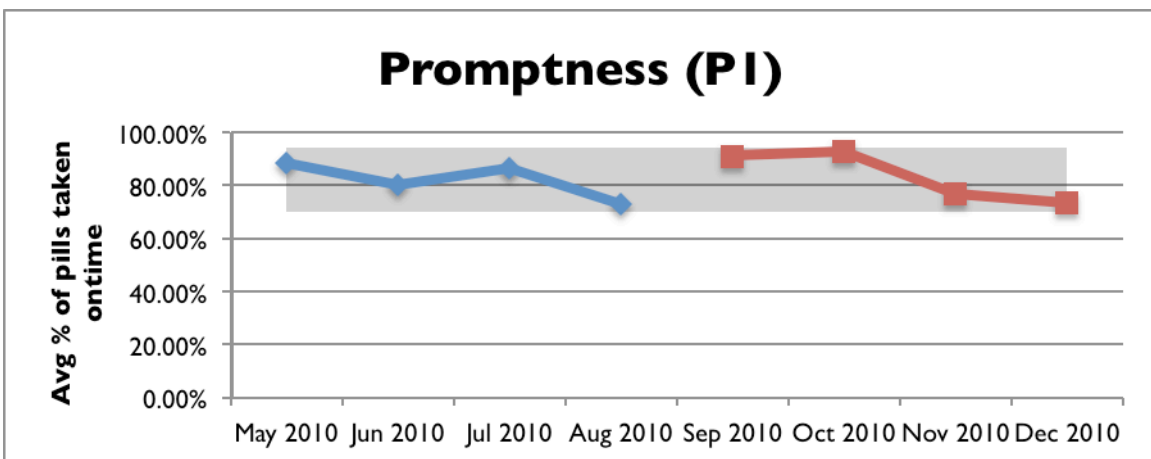


Figure 5-6. Participant P1's medication promptness before and after reflecting on her data in Aug 2010. The rates following the reflection session fall within the two standard deviation band (gray) of the pre-reflection rates. The data reflection does not appear to have affected her (already good) promptness behavior.

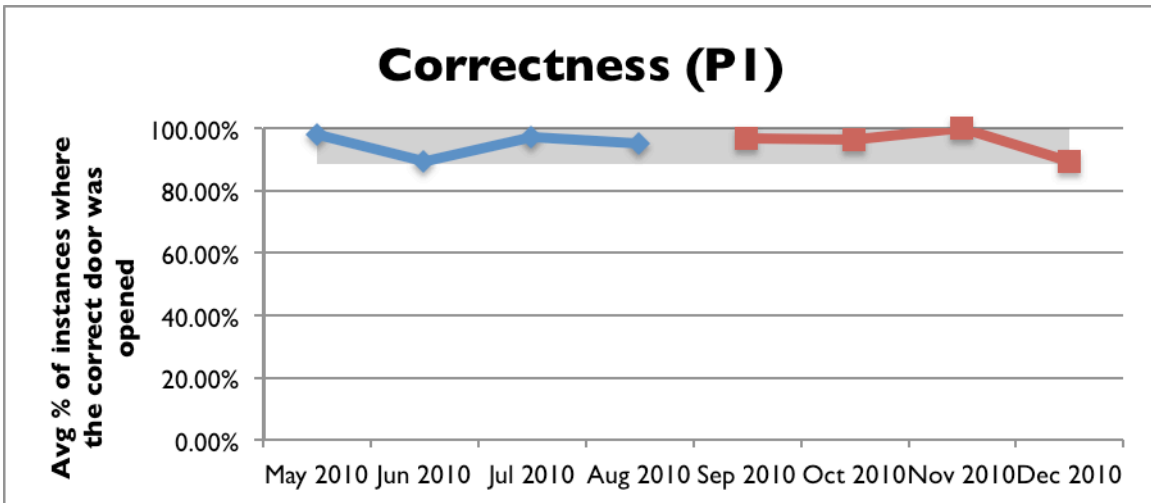


Figure 5-7. Participant P1’s medication correctness rates before and after reflecting on her data in Aug 2010. The rates following the reflection session fall within the two standard deviation band (gray) of the pre-reflection rates. The data reflection does not appear to have affected how often she opens the correct pillbox door.

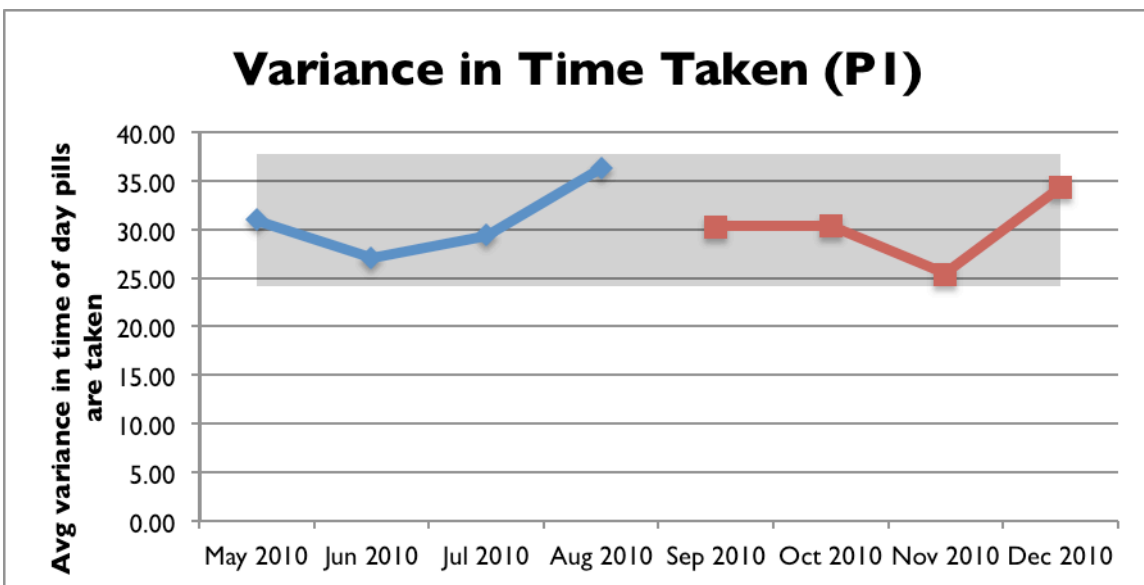


Figure 5-8. The average variance in the time of day Participant P1 took her medications per month before and after reflecting on her data in Aug 2010. The average variances per month following the reflection session fall within the two standard deviation band (gray) of the pre-reflection rates. Reflecting on the data does not seem to have changed the variance in the time of day she took her medications.

The results show P1 did not change her pill-taking behaviors significantly after the retrospective reflection session, according to both the visual inspection and the 2SD band statistical approach. This is not surprising given the fact that P1 already felt she was doing adequately well and her self-perception of her pill-taking behaviors matched well with the sensor data shown to her in the reflection session. Through visual inspection, an observer can notice that the level of data points during the baseline and post-intervention phases in the four graphs above are similar and the trajectories of the values are relatively flat over time. A careful observer may also notice that the values for promptness (Figure 5-6) is a little higher (more prompt, which is good) just after the intervention and the values for the variance in time taken is a little lower (less variation in the time of day the medication was taken, which is good) just after the intervention.

However, the 2SD approach demonstrates that most (if not all) the values in the post-reflection phase in all the graphs fit within the 2SD band calculated from the variation of the data during the baseline phase. Thus the conclusion is that P1 did not significantly improve or decrease her medication adherence (Figure 5-5), promptness (Figure 5-6), correctness (Figure 5-7), or variance in time taken (Figure 5-8) as a result of the reflecting on the sensor data of her medication taking behaviors. Again, this is not surprising, as P1 already had a fairly accurate self-perception of her abilities even before she reviewed the data. The data did not trigger any a sense of dissonance about her ability to carry out her medication taking, so she did not feel she had to change very much. During the reflection session, she did say she wanted to be more consistent, but the degree of improvement in her actions was either negligible or too small to be detected in the analysis for this study.

The following graphs show Participant P2's adherence, promptness, correctness, and variance in time taken before and after reflecting on the sensor data from dwellSense in August 2010. An interpretation of the graphs is included in the subsequent paragraph.

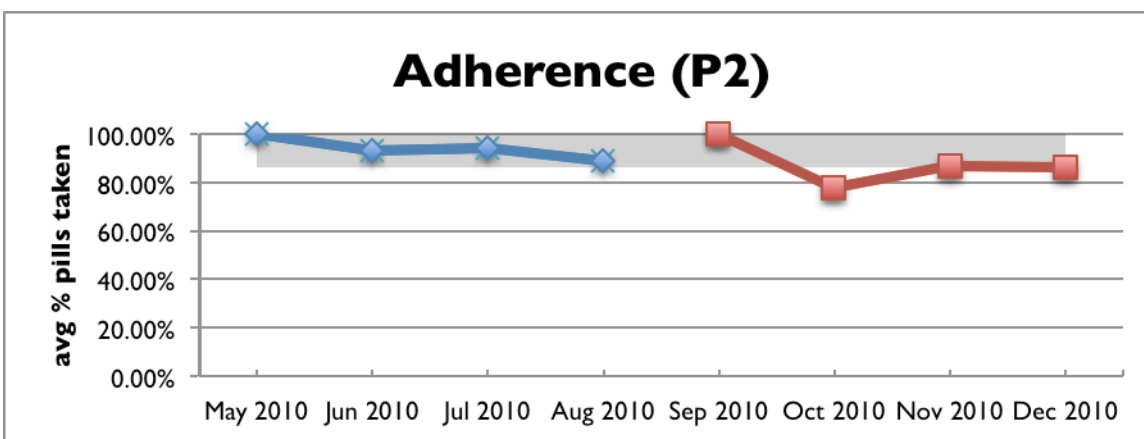


Figure 5-9. Participant P2's medication adherence rates before and after reflecting on her data in Aug 2010. Most of the rates following the reflection session fall within the two standard deviation band (gray) of the pre-reflection rates, except for one in October. One data point is not sufficient to constitute a significant change. Thus, the data reflection does not appear to have affected her adherence behavior.

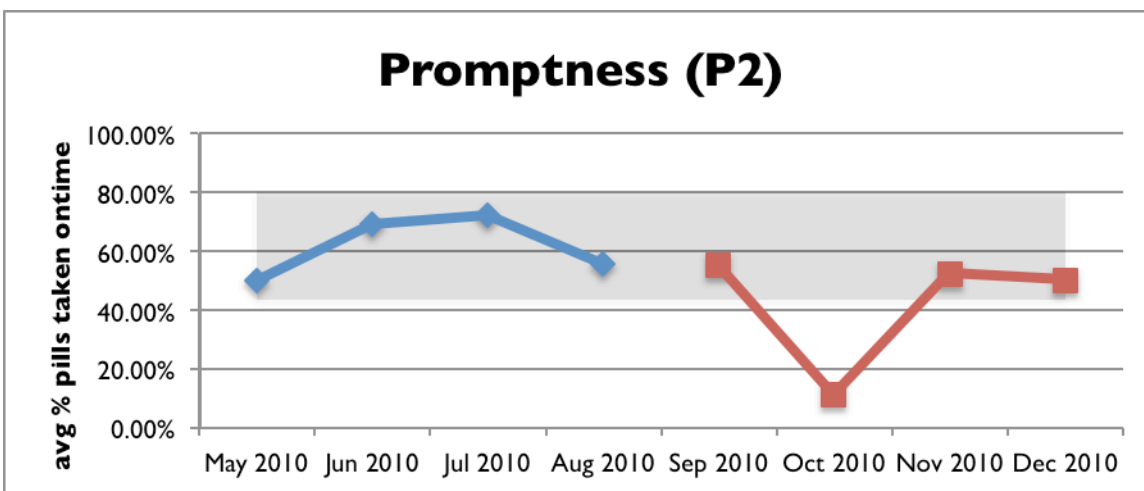


Figure 5-10. Participant P2's medication promptness before and after reflecting on her data in Aug 2010. Most of the rates following the reflection session fall within the two standard deviation band (gray) of the pre-

reflection rates, except for one in October. One data point is not sufficient to constitute a significant change. The data reflection does not appear to have affected her promptness behavior.

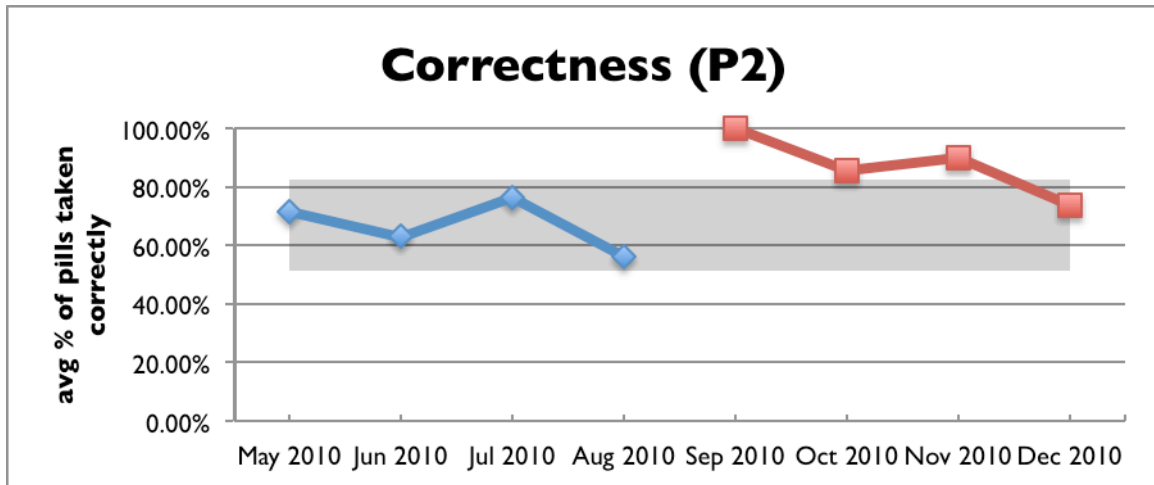


Figure 5-11. Participant P2's medication correctness rates before and after reflecting on her data in Aug 2010. The correctness rates in the three months following the reflection session fall outside the two standard deviation band (gray) of the pre-reflection rates. Thus, the data reflection appears to motivated her to select the correct pillbox door significantly more often.

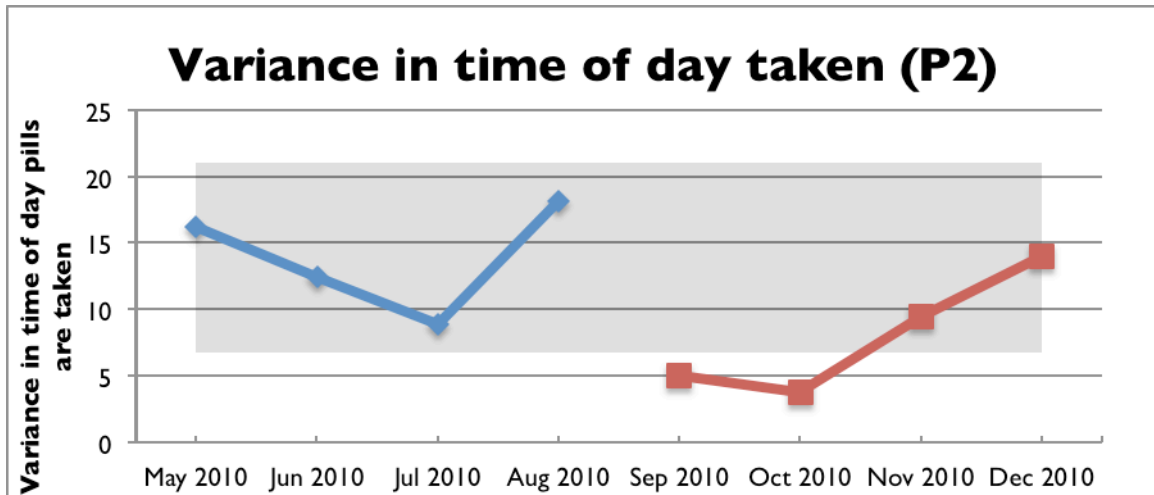


Figure 5-12. The average variance in the time of day Participant P1 took her medications per month before and after reflecting on her data in Aug 2010. The average variance in time taken in time taken in the two months following the reflection session fall outside the two standard deviation band (gray) of the pre-reflection rates. Thus, the data reflection appears to motivated her to be less variable in the time of day she is taking her pills.

The analysis of the data for P2 reveals that she was able to make a change in her pill-taking behaviors after the retrospective reflections session. In particular, both visual inspection and the 2SD band approach show that P2 improved significantly in the correctness (selecting and opening the right door on the pillbox) of her pill taking (Figure 5-11) and the variance in the time she took her pills (Figure 5-12). Through visual inspection, an observer can see that the data points in the post-intervention phase for correctness are higher than the data points in the baseline phase (Figure 5-11). The 2SD band technique concurs with the visual inspection showing that the first two months following the reflection session exist above the 2SD band. Thus, P2 was able to improve significantly how often she selected the



correct pillbox door to open in the two months that followed the reflection session. The 2SD band technique was particularly effective for these data because the variation in the baseline phase is fairly small. A similar result can be seen in the data for the variance in the time of day that the pills were taken. Through visual inspection, an observer would notice that the variance in the time of day varies quite a bit over time for P2 (Figure 5-12). In the month prior to the reflection session, the variance in time of day is rather high and seems to be on an upward trajectory. However, the variance in the time of day drops fairly dramatically just after the reflection session. The 2SD band approach corroborates the results from the visual inspection. The variance in the time taken in the two months just after the reflection session fall outside the 2SD band calculated from the baseline data. Thus, P2 was able to be less variable in the time of day she took her morning medications in the two months following the reflection session. Visual inspection also shows that this drop (*i.e.*, improvement) in the variance in the time of day is temporary, lasting only about two months before it returns back closer to levels found in the baseline phase. Visual inspection and 2SD band analyses of the data for other features of the medication taking routine (adherence in Figure 5-9 and promptness in Figure 5-10) showed no significant change in level. However, the data do demonstrate that P2 was able to follow through with her intentions to open the correct pillbox doors according to the day of the week and also to take her medications closer the same time of day. These changes in correctness and in the variance in the time of day she took her pills were somewhat temporary, lasting approximately two months before it reverted back to the level before reflecting on the sensor data.

### 5.7.2 Analysis of Phone Use Behavior

During the reflection sessions, participants also reflected on their phone use data including how often they made outgoing calls, received incoming calls, the average number of rings before answering the phone, and how often they were dialing the phone. In this case, how well they were dialing the phone corresponds to the rate at which they dialed the phone correctly (instead of misdialing the phone). A phone call is considered a misdial if the call lasts less than 60 seconds, has at least 70% of its digits overlapping with the following call and/or transposed digits of the following call. Measures for how often individuals made outgoing calls provides a fairly coarse measure of cognitive or functional ability, although individuals will likely misdial the phone more often before they deciding to use the phone less often. The number of incoming calls does not provide a direct measure of functional or cognitive ability but it does provide a measure of socialization and was included in the reflection session as an interesting feature to give the participants a broader view of their activities.

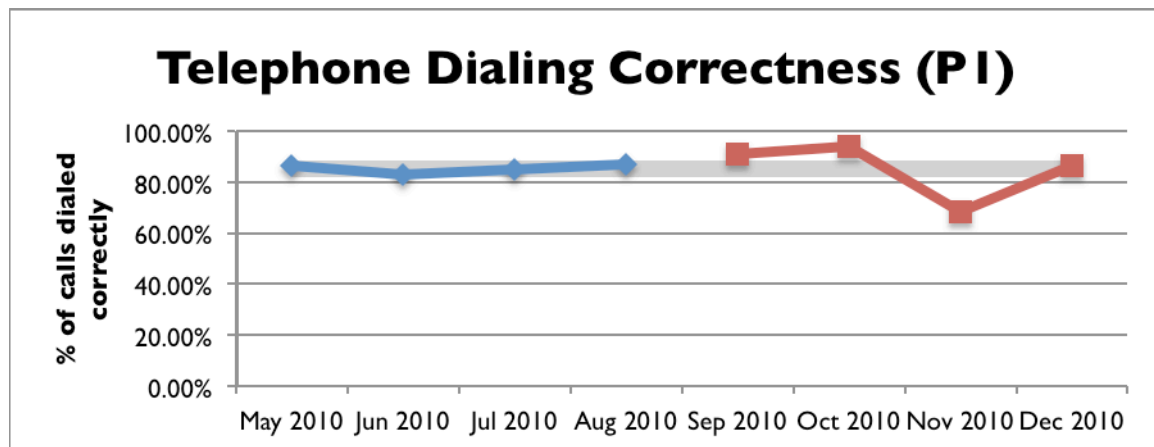


Figure 5-13. The percentage of outgoing calls per month that were dialed correctly (*i.e.*, not misdialed). The first two values after the reflection session (red) in August 2010 appears outside the two standard deviation band (gray) calculated from the pre-reflection baseline data (blue). These data show that P1 was able to dial her phone significantly more correctly following the reflection session.

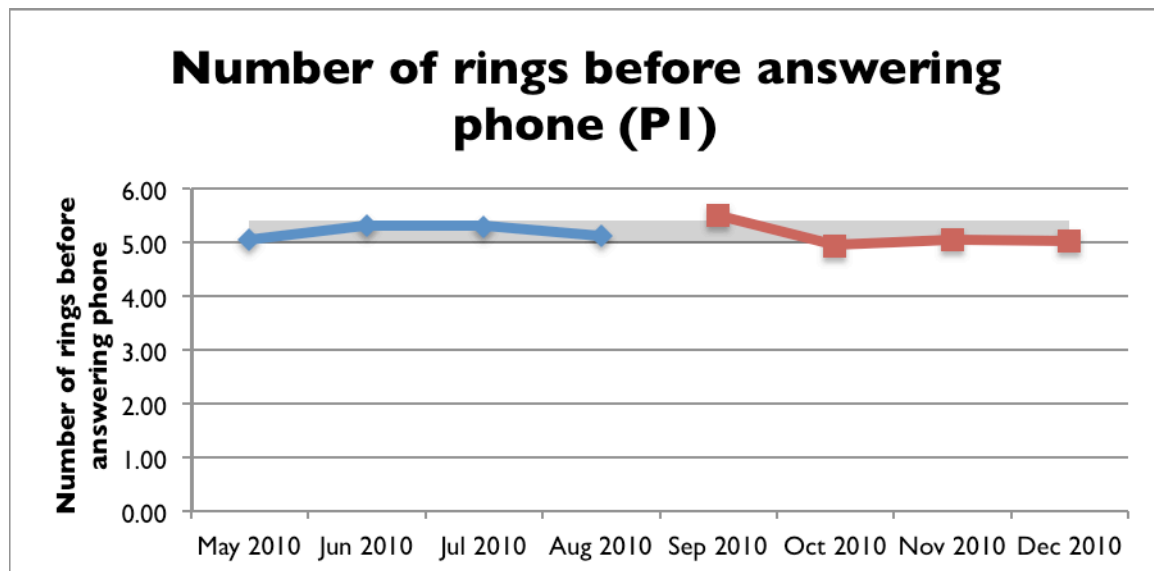
The analysis of P1's phone use behavior reveals that she increased frequency of dialing correctly in the two months following the reflection session. With visual inspection in **Error! Reference source not found.**, the reader can notice might notice that the values for September 2010 and October 2010 appear a little higher than the pre-reflection rates. Indeed, the two standard deviation band methods shows that the first two months exist outside the 2SD band calculated from the pre-reflection monthly correctness values. When considering changes in the volume or number of outgoing calls made by P1 before and after the reflection session with the visual inspection method in **Error! Reference source not found.**, the reader may notice P1 may have made fewer calls (averaging approximately 65 calls per month in the months following the reflection session, instead of an approximate average of 88 per month before the reflection session. However, the two standard deviation band method shows that values, despite their decrease, does not exist outside the 2SD band calculated from the pre-reflection data. Similarly, the data for the number of rings in **Error! Reference source not found.** before P1 answered the phone does not show a significant change according to the 2SD band method.

The analysis for P2's phone use behavior excludes the correctness metric because the phone sensor could not decode the DTMF signals from her digital-based phone line and could not provide information about what numbers she was dialing. The data show that P2 did not make any significant changes in the number of outgoing calls (**Error! Reference source not found.**) nor did she increase or decrease the time (or rings) to answer the phone (**Error! Reference source not found.**). The lack of change in the data are not surprising, as P2 did not feel her phone use behaviors were that interesting or important to change.

## 5.8 Design Recommendations

Based on our observations of how the older adults in our case studies explored and used the information to make sense of and reflect on their own functional abilities, we provide recommendations for designing home sensing systems that support self-reflection for older adults.

We observed that users were eager to not just look at the visualizations at a glance but actually spent the time to study



**Figure 5-14.** The number of rings before P1 answered an incoming phone call, averaged over each day in a month. The values after the reflection session (red) in August 2010 all exist within the two standard deviation band (gray) calculated from the pre-reflection baseline data (blue). These data show that P1 did not make a significant change in answering the phone more quickly or more slowly.

them to find instances where it looked like they made a mistake. Similar to the process of finding key events in data

used in intelligence analysis (Chin, Kuchar, & Wolf, 2009), users looked for anomalies. However, in contrast to intelligence analysis, which requires a high amount of interpretation by the analyst, anomalies in task performance can be more easily identified computationally. Thus, we recommend from a usability standpoint that *the instances of anomalous behaviors likely to be caused by the user should be highlighted or at least represented in a way that requires minimal analysis by the user*. As a negative example, the multi-month visualization (Figure 5-2) for pill taking only included marks where the pillbox was opened, relegating the representations of missed pills to narrow, difficult-to-notice columns of white space.

The contrasting outcomes between our two participants highlight the need for supporting better explanations of anomalous behaviors. Many older adults have difficulty remembering their recent experiences due to either neurological conditions (like P2) or simply benign declines in memory associated with aging (like P1). To be able to identify whether an anomalous behavior is acceptable or a mistake, the context of the behavior needs to be available. Thus, designers of embedded assessment systems should *provide tools to allow the user to retrieve the context of the data such as special events from calendars, people encountered that day, or to-do lists in addition to just presenting the data by itself*.

In addition to providing context for explanation, *providing the low-level details of the behavior can support a better understanding, explanation, and fixing of the behavior*. For example, seeing that a particularly long telephone number is being misdialled helps the user to understand they might have a problem with digit span memory and have to pay attention more when dialing that number. We also found that there was little demand to see the details of behaviors that were judged as “good” (such as correctly dialed phone calls or days where pills were taken). Thus designers, if faced with a shortage of resources, can focus on providing the details for the anomalous cases. For example, when a system is able to detect an anomaly in near real time, the system can increase the resolution of sampling at the cost of a temporary increase in battery or memory consumption.

The value of embedded assessment data extends beyond the individual monitored to other members in their care network such as relatives and clinicians. The users in our case studies suggested that the information be shared with their relatives so they can look after them more closely. However, users were concerned that others might misinterpret their seemingly errorful but explainable behaviors as mistakes. Therefore, designers should *allow for collaborative sharing and exploration of behavioral data or support annotation of the data before sharing with others to avoid misinterpretations*.

Based on these design recommendations, we incorporated new features into the next version of dwellSense described in the next chapter. Due to the scope of the project thus far, we did not support direct sharing of the content beyond the individual, so we did not incorporate the last design recommendation about support annotations or collaborative reflection. We leave this feature to be added in our future work.

## 5.9 Pilot Study Summary

These two case studies address research question RQ4 by demonstrating one of most important potential benefits of embedded assessment data—that it helps older adults with managing their awareness of their functional abilities. We found that the objective data collected on her task performance allowed an older adult to adjust her inaccurate awareness of her functional abilities as well as for another older adult to affirm her accurate awareness of her abilities. As a result, they were empowered to make the appropriate adaptations to be more consistent and aware of their pill taking and phone use to safeguard their independence. Furthermore, to avoid misinterpretation when sharing performance data, designers should support joint viewing or at least allow the older adult to annotate and explain their performance.

We found that our participants looked for and focused on anomalies in the data (*e.g.*, missed pills or misdialled phone calls) that may indicate a mistake that might be their fault. They tried their best to explain away the anomaly by thinking of an event, circumstance, or reason why that anomaly might actually be acceptable. They drew first on their

own memory of events to find an explanation. Often lacking a specific explanation from their declining memories, the older adults drew next on their routines in an attempt to make the anomaly acceptable by placing it within one of their routines. They then consulted other sources of date-specific information such as calendars and diaries if they were available. Designers can support this investigation process by clearly marking the anomalies and can support the explanation process by providing the date-specific context that gives hints as to what activities might have occurred on particular days.

The impact of reflecting on the objective sensor data also included changes in behavior in addition to self-awareness and intentions to improve their behaviors. For the individual who was already fairly consistent in her task performance, we were unable to find significant improvements in her task performance behaviors in the months following the reflection session. In contrast, for the individual who had a poor self-awareness of her abilities, she was able to follow through on improving on opening the correct pillbox door and being less variable in the time of day she took her medications. However, these improvements in performance appear to be rather short-lived, lasting only about one to two months after the reflections session. The results of this pilot study demonstrates that a one-time informational intervention consisting of reflecting on data visualizations of individuals can result in changes in awareness, intentions to improve performance, and actual performance on tasks important for individuals, particularly for individuals who may not an accurate insight into their own abilities.

Having demonstrated the usefulness of reflecting on embedded assessment data in a case study with two older adults, we will discuss in the next chapter how we applied what we learned when engaging individuals with their own data. Using the User Reflective Design Framework, we now move from engaging users with their own data (which is what we did in the pilot study) to finding new opportunities for additional forms of sensing and feedback to improve the ability of the system to support individuals in reaching their goal to remain functionally capable. The next chapter will also describe a larger deployment of the revised dwellSense system to identify whether a larger population could benefit in the same way as the two case study participants benefitted from reflecting on the sensor data about their task performance.

## 5.10 References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2), 179–211.

Chin Jr, G., Kuchar, O. A., & Wolf, K. E. (2009). Exploring the analytical processes of intelligence analysts. *Proceedings of the 27th international conference on Human factors in computing systems* (pp. 11–20). Retrieved from <http://dl.acm.org/citation.cfm?id=1518704>

Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., et al. (2008). Activity sensing in the wild: a field trial of ubifit garden. *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (pp. 1797–1806). Retrieved from <http://dl.acm.org/citation.cfm?id=1357335>

Erickson, T., Podlaseck, M., Sahu, S., Dai, J. D., Chao, T., & Naphade, M. (2012). The dubuque water portal: evaluation of the uptake, use and impact of residential water consumption feedback. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (pp. 675–684). Retrieved from <http://dl.acm.org/citation.cfm?id=2207772>

Fishbein, M. (1980). A theory of reasoned action: some applications and implications. *Nebraska Symposium on Motivation*. *Nebraska Symposium on Motivation* (Vol. 27, p. 65).

Hayes, T. L., Hunt, J. M., Adami, A., & Kaye, J. A. (2006). An electronic pillbox for continuous monitoring of

medication adherence. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE (pp. 6400–6403). Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=4463275](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4463275)

Janz, N. K., & Becker, M. H. (1984). The health belief model: A decade later. *Health Education & Behavior*, 11(1), 1–47.

Kuriansky, J. B., Gurland, B. J., Fleiss, J. L., & Cowan, D. (1976). The assessment of self-care capacity in geriatric psychiatric patients by objective and subjective methods. *Journal of Clinical Psychology; Journal of Clinical Psychology*. Retrieved from <http://psycnet.apa.org/psycinfo/1976-28120-001>

Krishef C.H. (1991) *Fundamental approaches to single subject design and analysis*. Malabar, FL: Krieger Publishing Company.

Lee, M. and Dey, A. Reflecting on Pills and Phone Use: Supporting Awareness of Functional Abilities for Older Adults. *Proceedings of CHI 2011*, (2011), in press.

Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 557–566). Retrieved from <http://dl.acm.org/citation.cfm?id=1753409>

Little, A. G., Hemsley, D. R., Volans, P. J., & Bergmann, K. (1986). The relationship between alternative assessments of self-care ability in the elderly. *British journal of clinical psychology*, 25(1), 51–59.

Maitland, J., Chalmers, M., & others. (2010). Self-monitoring, self-awareness, and self-determination in cardiac rehabilitation. *CHI'10 Proceedings of the 28th International Conference on Human Factors in Computing Systems* (pp. 1213–1222).

Mamykina, L., Mynatt, E., Davidson, P., and Greenblatt, D. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, (2008), 477–486.

Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of consulting and clinical psychology*, 51(3), 390.

Saxton, J., Morrow, L., Eschman, A., Archer, G., Luther, J., & Zuccolotto, A. (2009). Computer assessment of mild cognitive impairment. *Postgraduate medicine*, 121(2), 177.

# 6

## The Time Dimension of Reflection

The results of the pilot deployment highlight the effectiveness of objective sensor data to help individuals build or maintain an accurate self-perception of their functional abilities. Individuals engaged in a sensemaking process in which they identified anomalies in the data, generated explanations, and then confirmed their explanations with details in the sensor data. The most difficult step in the sensemaking process was generating explanations, particularly explanations for events that occurred in the distant past. Memories of routine events in the distant past are difficult to recall even for younger adults and older adults without cognitive decline. Time has been found to be a particularly poor memory cue in cued recall tasks for autobiographical memory (Conway & Bekerian, 1987). This human limitation provides an opportunity for computation to play a role in helping individuals in their sensemaking process. In particular, providing individuals with the means to reflect on their personal data in near real time can help individuals make sense of their data without requiring them to engage in a difficult (cued) recall task about the distant past. With making sense of real time feedback, the context of the data and the experience associated with it is likely to be easier to recall than experiences from a particular time in the distant past. In this chapter, we explore the dimension of time in reflecting on personal sensor data, comparing reflection on data presented and updated in real time and reflection on data presented as a long term trend. We conduct a study comparing different ways of reflecting on data about how well individuals carry out their daily tasks. The results of this investigation will provide insights into how to design reflective interactions for a design personal sensing system and how to support self-awareness and behavior change using data.

### 6.1 Supporting Explanations of Data Events

Explaining anomalies in the data is a critical step in the sensemaking of personal sensor data. To explain an anomaly, an individual needs to be able to draw on the factors that are related to or may have caused that particular event, such as travel out of town explaining an instance of missed medication, being sick explaining a day with oddly low physical activity, or relatives visiting explaining a week of unusually high home energy use. Recalling these factors about an anomaly from the past can be difficult because of limitations of human memory, particularly if reflecting on routine behaviors and actions.

To provide the context, we can follow two approaches: 1) reinstating a past context or 2) leveraging the current context. The first approach aims to reinstate a past context by capturing these factors from the past and making them available to the individual when she engages and reflects on the data. This first approach can be thought of as “assistive” in the sense that it provides the context that the individual is unable to reinstate in her mind, and thus compensates for a form of “disability” in the lack of information necessary for explaining the anomaly. This approach to reinstate a past context has the potential to help with explanations but can only be realized if the system is able to overcome a number of challenges: a) first identifying what these factors are, b) sensing and recording these factors for later use, and c) presenting these data streams in way that allows the individual to understand how these factors relate to the actual event. Prior research has investigated the value of specific types of contextual cues such as location (Kalnikaite et al., 2010), physical artifacts (Carter & Mankoff 2005), kinesthetic cues (Tan et al., 2002), digital actions (Hailpern et al., 2011), however, when designing a personal sensing systems, which contextual cues to choose might not be evident at the time when the system is designed. Sensing technologies, as advanced as they are, still may not be able to detect the unexpected factors that are relevant for explaining anomalies, for example, mood, special appointments, and acceptable deviations in routines. Thus, reinstating a past context is possible, but can be difficult in practice.

Another approach for making sure the individual has the correct context for generating explanations of anomalies in the data is to rely on the current context as the basis for explanations. Instead of reinstating a context from the past, we can design a reflective interaction that encourages the individual to use cues from their present situation (such as recent memory and environmental cues) to explain a current or recent event. One such reflective interaction can be enabled by providing data about the individual's own behaviors in real time. With real time feedback, the individual can receive automated feedback on her task performance as she completes or soon after she completes her tasks. She can view this feedback and use her recent memory of the tasks she recently performed (a few seconds, minutes or hours ago, rather than weeks or months ago) to understand what might have influenced her performance.

Real time feedback can typically be shown using a dashboard style display that shows the current state of the system (or in this case, the task performance of the individual) and is updated within seconds or minutes. Dashboard displays do not typically provide the previous states of the system and thus do not provide any long-term information about how the individual is performing. A dashboard display encourages individuals to reflect on it often and regularly "in the moment" and can potentially support a more continuous awareness of the individual's own performance and abilities. More frequent reflection can lead to more opportunities to take action and see the results of their actions quickly reflected in the display. For example, a dashboard display that shows what time an individual has taken her morning pills can be used by the individual to check whether she had taken her morning pills when she is in doubt.

In contrast, providing data as a single series of events from an extended period of time (for example, in a chart or graph) allows individuals to see the trends and patterns in their behaviors. They are able to visually compare their performance from among multiple days to identify when their behaviors may have been different than normal. The behaviors captured and displayed by the system are directly available for comparison across time. With a long-term representation, individuals can spot the trends in their behaviors over time and are able to see a larger overall pattern if it exists. With only a real-time view, spotting trends requires the individual to remember and recall previous data values and their associated timestamps in order to mentally compare their current performance with previous performance. Combining both real-time and long-term feedback is a possibility, but it introduces additional complexity in both the system design as well as in the interaction design. Thus, when representing data for reflection, it is important for system designers to know the tradeoffs between either explicitly showing data values from a long period (but typically lacking the explicit context necessary to explain data values from the distant past) or explicitly showing data values continuously as the actual behaviors are performed or sensed (but lacking the ability to highlight trends in the recorded behavior because only the current data point is available). This chapter investigates the differences between single-time long-term feedback and continuous real-time feedback on the ability of the individual to be more self-aware of their behaviors and to improve their behaviors.

## **6.2 Background**

In this section, we briefly describe related work about feedback, goal setting, and medication adherence that underpin the hypotheses proposed in this chapter as well as to help interpret the results.

### **6.2.1 Feedback**

Providing feedback has long been a part of programs for health behavior change. Feedback can be categorized as generic, targeted, or personalized (Kreuter, Strecher, & Glassman 1999). Generic feedback provides an individual with information that applies to a larger population, for example, smoking cessation tips for someone trying to quit smoking. Targeted feedback is more specific than generic feedback but still does not take individual differences into account. An example of targeted feedback can be advice about the importance of mammograms for women over age 50. Personalized feedback, the type of feedback of focus in this dissertation, is the most specific level of feedback and takes into the account the parameters that make the situation of the individual more specialized. Typically, personalized

feedback is coupled with some sort of assessment of the individual (for example, their habits, abilities, preferences, genetics, etc.) so that the feedback can be customized for the individual's specific parameters. Personalized feedback can be either normative (in which the individual is compared with others in the population) or ipsative (in which feedback is presented to be compared within an individual over time). An example of normative personalized feedback might compare the walking speed of an older adult with other older adults of the same age group whereas an example of ipsative personalized feedback would compare the walking speed of an older adult with her walking speed from one year ago.

Personalized feedback has been shown to be useful across a number of domains to support greater awareness and to support health behavior change. A study of exploring different interventions for encouraging healthy habits in consuming alcohol for students on a university campus found that personalized feedback was helpful reducing the rate of drinking and harmful consequences when compared with non-personalized feedback (Baer et al., 1992). Ipsative personalized feedback has also been found to help reduce fat intake and increase the intake of fruits and vegetables (Brug, Steenhuis & DeVries 1999), an example of not just discouraging a negative behavior but also encouraging a positive behavior. Smokers experienced higher smoking cessation rates with personalized feedback (Prochaska et al., 1993), which helped motivate individuals who were not yet contemplating quitting smoking to consider quitting, moving them along the stages in the Transtheoretical Model of Behavior Change (Prochaska & DiClemente, 1983).

Personalized feedback is also commonly used in a counseling technique called Motivational Interviewing (Millner & Rollnick, 1991) designed to help individuals who have an addiction to move along the stages of the Transtheoretical Model of Behavior Change. The role of personalized feedback, such as logs of the individual's own habits or actions, helps highlight a discrepancy between where the individual current is and where the individual wants to be in the process of change. Personalized feedback provides the current status of the individual, which the individual can use to motivate herself to move forward closer towards the desired state. In Motivational Interviewing, the counselor is present to talk through the non-threatening feedback and raise the self-awareness of the individual. In the same way, the non-threatening, personalized feedback based on the objective sensor data about task performance, can be used by older adults to identify a discrepancy between their current self-awareness of their task performance and their actual task performance. Personalized feedback is also useful in setting and tracking progress towards goals, a topic discussed in the next section.

## 6.2.2 Goal Setting

Setting goals is one strategy used to increase performance. In particular, setting specific goals to achieve a task, when combined with personalized performance feedback about how well the individual is meeting her goals, leads to higher performance than when setting vague goals (Locke & Latham, 1990). Setting goals have been found to increase the amount of effort the individual devotes to a behavior, the longer they stick with the behavior (persistence), and the less likely they are to be distracted (higher concentration). Factors that influence the effectiveness of goal setting include a strong commitment to the goal, the complexity of the goal (not too low, but not too high), personal abilities, external obstacles, and frequent feedback on goal progress. This last factor, frequent feedback, is the subject of the study in this chapter which will investigate whether more or less frequent feedback is helpful for helping individuals improve their task performance. Similar to the idea in Motivational Interviewing frequent feedback can be an effective way of stimulating effort to perform highly because it can highlight the difference between the individual's current state the desired goal state (Bandura 1997).

In addition to setting goals, creating and setting sub-goals or "proximal goals" also provides the cognitive scaffolding that support higher performance (Locke & Latham, 1990). Using strategic analysis to break larger goals into the small, more attainable goals reduces the cognitive burden of developing an overall strategy to achieve a larger goal all at once. Instead, each sub-goal determines the amount of effort as well as the specific strategies necessary to achieve it.



Furthermore, progress on achieving sub-goals provides individuals with the frequent feedback that not only make the sub-goals feel more tangible and rewarding but also shows progress towards the larger overall goal. Setting sub-goals can also enhance self-efficacy (the perceived ability to carry out the task in spite of obstacles) and satisfaction (Bandura & Schunk, 1981).

However, setting goals without feedback often does not lead to greater performance. Likewise, providing feedback on performance without a set goal does not necessarily result in better performance (Latham, Mitchell, & Dossett, 1978). In addition to the feedback on task performance, other intrinsic rewards associated with goal setting and attainment include self-satisfaction, when the individual feels that she has mastered the task or has enhanced an aspect of the self. In the context of the benefits of high achievement in healthy aging, individuals who perform highly in tasks like medication adherence can feel that they can master the new task of taking pills or can maintain their ability to take their pills. In other words, by being able to perform IADLs well without any difficulty can allow individuals to feel not as “old”. The framing of goals is also important. Frequent feedback might actually result in lower achievement for goals that are framed negatively (to avoid a negative behavior such as avoiding eating fatty foods or avoiding smoking cigarettes) because through the frequent feedback, individuals are reminded frequently about their failures, which can lead them to abandon the goal (Cochran & Tesser 1996). In the context of this study, all study participants, like most other older adults, have made a personal goal to take their medications consistently every day because they believe that medications are important for their maintaining their health. Individuals had this goal, whether explicitly or implicitly, even before the beginning of the study, and thus providing different forms of feedback (real-time vs. long-term) has the potential to influence how they are able to perform in their medication taking.

### **6.2.3 Medication Taking and Adherence**

The main behavior of focus in the study described in this chapter is medication taking. Medication taking is recognized by both older adults and clinicians as an important task for maintaining independence. From a sensing perspective, it also is performed frequently and thus provides frequently opportunities for assessment. Medication adherence is typically defined as the proportion of prescribed doses taken on the correct day, and this is the definition used in this study. For example, if individuals are supposed to take their medications twice a day, and if they are 100% adherent to their medications, then it is expected that they would have performed the medication-taking task 2 times a day, 14 times in a week and etc.

Medication adherence is important not only from a personal health perspective but also from a societal perspective. Proper medication adherence was found to reduce medical costs and lower hospitalization rates. In fact, higher medication costs were found to be offset by the savings in medical and hospital costs (Sokol et al., 1995). Problems with medication non-adherence are most common for individuals who are prescribed a new drug for hypertension, hyperlipidemia (high cholesterol), and diabetes, which are conditions common among older adults (Fischer et al., 2010). It is also common for individuals to have more difficulties with taking medications on time rather than missing them altogether (Choo et al., 1999).

Attempts to improve long-term medication adherence rates among patients tend to be most effective when combined together and as a result, are often complex, labor-intensive, and expensive. A review of interventions to improve medication adherence and clinical outcomes found that the most effective intervention to improve long-term medication adherence involved combinations of counseling, reminders, close follow up, supervised self-monitoring, and feedback (Raynes et al., 2008). Thus, there is a need for lower-costs solutions to improve medication adherence. Furthermore, measuring medication adherence is difficult, with majority of the studies reviewed by Raynes and colleagues relying on subjective self-reported measures of medication adherence. The Medication Event Monitoring System pill bottle cap is occasionally used to collect data about how individual take their medications, but it is limited

to interactions with a specialized pill bottle (Olivieri et al., 1991). Thus there is a need for objectively collected data about long-term medication adherence.

To explore how real-time and long-term feedback influences how an individual achieves their goal to taking their medications, we conducted a field study described in the next section in which we deploy dwellSense version 2.0 to capture, assess, and feedback how individuals carry out IADLs including how they take their medications.

## 6.3 Study Design

In order to investigate the differences between single-time long-term reflection and continuous real-time reflection, we conducted a study in which we compared two groups of individuals, with each group using one of these types of reflection. In order to support reflection on personal data, we deployed a revised version of dwellSense (called dwellSense 2.0) that could sense and assess how individuals perform activities of daily living such as medication taking, phone use, and coffee making as well as provide different forms of feedback for individuals to reflect on. For a full description of dwellSense 2.0, refer to the next section. dwellSense 2.0 provided two ways to reflect on the data, either through a continuous real time display in the form of a tablet (Figure 6-1) or a single-time session reflecting on trended long term data. We measured the impact of different reflection interactions on individual's self-awareness and their ability to maintain or improve their behaviors.

### 6.3.1 dwellSense version 2.0

We revised the design of dwellSense by adding two additional features based on opportunities identified from the analysis of the sensemaking process from the pilot study. These two opportunities stem from challenges that individuals encountered in the sensemaking process, in particular, generating explanations of events. A critical piece of context, whether the individual was home or away, was missing from the first version dwellSense, so we added motion sensing to the suite of sensors in dwellSense 2.0. Due to the difficulty in explaining events from the distant past, we added real-time feedback to the dwellSense 2.0.

#### 6.3.1.1 Motion sensing for additional context

The first challenge found in the sensemaking process is knowing when to trust data about medication adherence. One of the main reasons for distrusting the medication adherence data is because dwellSense cannot sense medication taking events outside the range of the apartment. Individuals would occasionally travel and stay overnight at their relative's home or at the hospital, and so dwellSense, not knowing any better, would interpret these days as instances of missed medications. Individuals struggled when thinking about whether they were home or not to explain away instances of missed medications. Thus, either dwellSense would have to accurately capture these medication taking instances away from the home or it must minimally be able to know when the individual is not home and not mark those days as missed medications. The fact that the individual is home is something that can be easily detected with sensing. To that end, we added passive infrared motion sensing at locations in the apartment (kitchen, living room, kitchen) that would detect whether an individual is present in the home or not. The motion sensor data can not only be used as an additional data stream or piece of context to present to the individual in retrospective sessions, but the task recognition algorithms used in dwellSense can take advantage of this additional piece of context to know when to excuse the individual when she is away from her home.

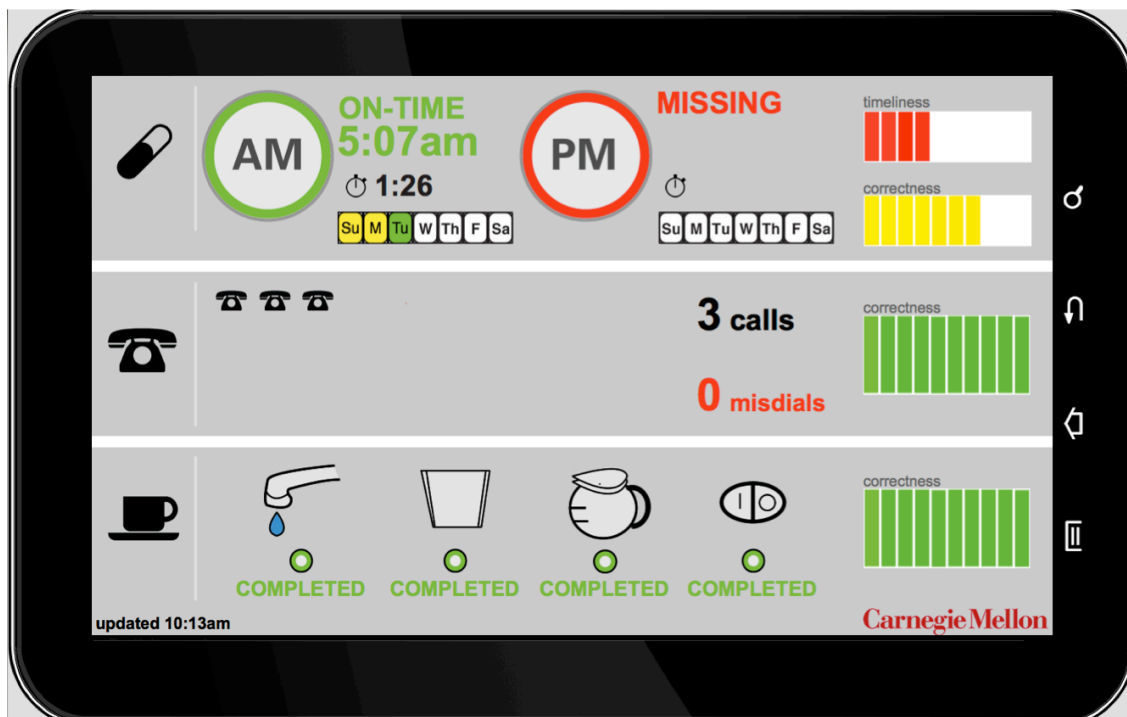
#### 6.3.1.2 Real-time feedback to encourage reflecting in the moment

The second challenge in the sensemaking process is generating explanations about events that occurred in the distant past. This challenge was already discussed in the introduction of this chapter as well as in Section 6.1. To address this challenge, dwellSense 2.0 provides real-time feedback in the form of a tablet-based display (Figure 6-1) situated in the home of the older adult. The content of the display updates in near real time and shows how the individual has performed their medication taking, phone use, and coffee making (for details, see the Appendix A for the user guide to

the display). The display is programmed to reset its content while the individual is sleeping at night so the information about the next day's task can be presented. The tablet-based display was designed to be used only to display information and thus was programmed to operate in a kiosk mode that made other features of the tablet (such as the browser and games) inaccessible to the individual. The individuals in our study had no interest in using these other function of the tablet but reported that visitors such as their grandchildren tried to use it (unsuccessfully) to play games or go onto Facebook.

We asked the individuals to put the tablet in a place that would be visible and encouraged them to look at the display at least twice a day. Individuals typically placed the display in their living room rather than the bedroom (because it was too bright at night) or kitchen (because there was limited counter space). The tablet was propped up like a photo frame so that individuals could see the screen without picking it up. The tablet was plugged into a power outlet so that it could be an always-on ambient style display. The tablet display did not require any interaction from the user to display information, in fact, the individual just has glance over (and perhaps make sure their glasses are on) to see the information. Higher-level details are available in large text and contrasting colors where as the lower-level details (like the particular pillbox doors that were opened) can be found in the smaller text. Individuals could choose to simply look at the high level details or they could step up closer to the tablet or pick it up to see some of the details.

The real time display has both front end and back end components. The front end of the display is implemented as an application for an Android tablet. We choose the Samsung Galaxy Tab 10.1 because it had a large, bright display with a good resolution well-suited for older adults. Because the app itself was designed to be more of a display than an interactive application, the app polls a remote server for updates every minute and renders a HTML page generated from the remote server. The remote server performs the back end processing and generates the content for the real time display. The basic dwellSense data flow architecture from the first version was used in dwellSense 2.0, where the sensor data collected by a laptop in the individual's home is uploaded (in this case, at least once every 30 minutes) to the remote server where it is stored, processed, and served up accordingly. In the case of the real-time display, the real-time content is generated by a server application (written in Python) that uses the Mako HTML/CSS templating engine to generate markup suitable for rendering on the tablet. All processing of the sensor logs were performed on a server located at the CMU campus, rather than on the laptop in each apartment because it was easier for researchers to monitor the processing. The sensor logs from the apartment were uploaded and processed at least every half hour.



**Figure 6-1** dwellSense real-time feedback display shows how individuals perform everyday tasks such as medication taking, phone use, and coffee making. Implemented and deployed as an Android app, the display updates at least once every half hour with updated task performance data. Each task is rated with a colored bar on the right edge. The more shaded cells in the bar, the better the performance rating.

### 6.3.2 Participants

In order to examine the differences between reflection interactions, we deployed dwellSense 2.0 into the homes of 14 community-dwelling older adults. Two of these individuals were removed from the study because they were not able to integrate our pillbox sensors into their routines because the relative or nurse who filled the individual's pillboxes would not work with only two boxes. This left us with 12 participants for the duration of the study (see Table 6-1).

These older adults lived in a low-income senior highrise building and due to income and a lack of strong social connections were most at risk for losing their independence and thus may benefit the most from the monitoring that dwellSense can provide. These individuals also had at least one chronic condition (such as high cholesterol, high blood pressure, chronic obstructive pulmonary disorder, and diabetes) that had a significant impact on their wellbeing and also required them to take medications regularly. Every participant used the instrumented pillboxes we gave them. We installed the phone sensor into every home except two individuals (L01 and L13) who did not had a telephone landline in their home. About half the individuals used a coffee maker on a regular basis and were willing to use our instrumented coffee maker. Others made coffee either with instant coffee or drank tea.

The participant recruitment process began with sending a brochure to each resident in a nearby low-income senior highrise with whom we had a previous connection. Residents were invited to an information session where they could learn about the study. Those who were interested in being considered for the study completed a screening questionnaire (Appendix B) that asked them about what tasks they perform on a regular basis, their chronic conditions, and their willingness to use our instrumented devices in their everyday lives. We received 26 screening questionnaires back. From those who completed the screening question, the research team used the inclusion/exclusion criteria to screen

individuals. Qualified individuals were contacted to schedule a visit to their home to enroll them in the study and to observe how they carry out their daily tasks. During the first visit, individuals gave their informed consent to participate in the study. Individuals were told that the study would last approximately 8-9 months and that their participation was completely voluntary, meaning that they could withdraw from the study at any time.

Individuals were selected to be in one of two experimental conditions: real-time or long-term, corresponding to the type of reflection interaction we had them engage in. Due to the fact that the real time group required a broadband connection, we selected individuals within range of two wifi hotspots we set up in the building. One hotspot was located on the 5<sup>th</sup> floor and the other was located on the 8th floor. The distribution of tasks monitored were balanced as much as possible between the two groups; for example, there were five individuals who used all three sensors, so we made sure the distribution was 2 in one group (long-term) and 3 in the other (real-time), rather than 1/4 or 0/5. The average ages of the real-time and long-term groups were 66.8 years and 67.7 years, respectively. There were only three males in the study, one was in the real-time group and the other two were in long-term group. Thus, the groups were balanced with respect to tasks, age, and gender.

Participant Number	Gender	Age	Experimental Condition	Medication Use	Phone Use	Coffee Making	Chronic Conditions
L01	F	66	Real-time	•		•	Cancer, diabetes, bipolar
L02	M	58	Real-time	•	•		High blood pressure, COPD
L03	M	76	Long-term	•	•	•	High Cholesterol
L04	M	56	Withdrawn				High blood pressure
L05	F	72	Real-time	•	•	•	High blood pressure, Arthritis
L06	F	67	Real-time	•	•	•	High blood pressure
L07	F	75	Long-term	•	•	•	High blood pressure, COPD
L08	F	88	Withdrawn				Cancer
L09	F	55	Real-time	•	•	•	COPD, Depression
L10	F	83	Real-time	•	•		COPD, High blood pressure
L11	F	65	Long-term	•	•		Bipolar disease
L12	F	66	Long-term	•	•		COPD, Arthritis
L13	F	52	Long-term	•			Diabetes, Arthritis
L14	M	72	Long-term	•	•		Cancer


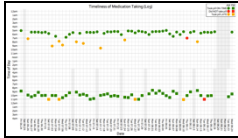
**Table 6-1. Twelve study participants, the tasks tracked for each individual, and their chronic conditions. 2 of the initial 14 participants were withdrawn from the study due to an inability to use our sensors to track their tasks.**

### 6.3.3 Deployment Timeline

We deployed dwellSense 2.0 into the homes of 12 community-dwelling older adults to sense, record, and assess their instrumental activities of daily living for 12 months (Table 6-2). Participants were given a \$30 grocery gift card once a

month during the study duration. The researcher’s interactions with the participants during each visit was recorded on a visit sheet (Appendix C) that documented what systems errors were addressed and what research activities (such as completing questionnaires) the participants performed.

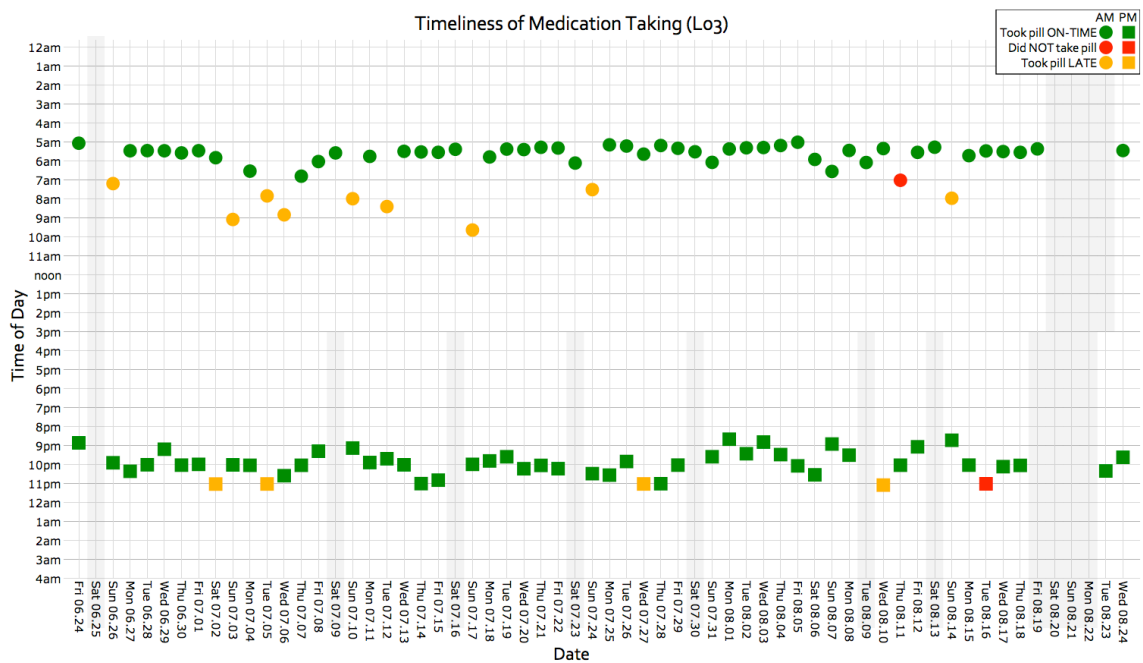
In the first two months after the sensors were introduced into their homes, no informational intervention was provided, but rather the individuals were told to carry on with their normal lives in their homes, with the dwellSense sensors installed. These first two months acted as a “baseline” period by which we can compare the relative effects of long-term versus real-time reflection. At the beginning of the third month of the deployment, we introduced the real-time display (as discussed in 6.3.1.2) into the homes of individuals who were selected to be in the real-time group. When the display was initially introduced, a researcher explained how to read the display and what the information meant. The researcher would informally quiz the individual to assess whether the individual was interpreting the information on the screen correctly. The researcher would make sure the individual understood the meaning of the displayed information before leaving the apartment. In subsequent visits, the researcher would also have the individuals explain the details of the display and correct the individual if they misinterpreted the real-time data. Individuals in the real-time group had the real-time display in their home for months 3 and 4.

	Data presented	
Month	Real-time Feedback Group	Long-term Reflection Group
		
1	baseline (nothing)	
2	baseline (nothing)	
3	Real-time display	No feedback
4	Real-time display	No feedback
end of 4	Real-time display	Long-term reflection on 2 months of data
5	Real-time display	No feedback
6	Real-time display	No feedback
7	Real-time display	No feedback
8	Display removed	No feedback

**Table 6-2 Deployment timeline for larger deployment of dwellSense into the homes of 12 older adults.**

In contrast, individuals in the long-term group did not receive any informational feedback during months 3 and 4 and could act as a “control” group for the real-time group. Significant differences between these two groups during months 3 and 4 can be attributed to the effect of real-time display introduced to the individuals in the real-time group at

beginning of month 3. At the end of month 4, individuals in the long-term group received their informational intervention by participating in a retrospective reflection session with the researcher. During these reflection sessions in the homes of the individuals, researchers showed each individual charts (Figure 6-2) of their behaviors from months 3 and 4 on a computer screen. The researcher followed a similar procedure as in the retrospective reflection sessions in the pilot study, where the researcher would begin with a recent, short term view of the data and slowly increase the time range to show up to 2 months of data (from months 3 and 4 of the deployment). The content of the long-term charts mirrored the data that would have been shown in the real-time display. One difference was the long-term view did not have an explicit rating for how well each task was performed, but the real-time display did show a rating of how well each tasks was performed in the form of the colored bars on the right-hand edge of the display. Nonetheless, the important distinction between the real-time and long-term reflection interactions is focused on the subtle differences in data content, but rather in how the data content is review either temporally distributed as tasks are performed or temporally isolated in a *post-facto* retrospective session.



**Figure 6-2** dwellSense long-term visualization of medication taking behavior showing approximately 8 weeks of historical data. Each dot represents when the individual has taken or not taken his morning pills (circle dots) or evening pills (square dots). Grayed out sections represent times when the system was unavailable or the individual was away from his home.

To ensure individuals in the long-term group could understand the content of the long-term charts correctly, the researcher would explain to the individual how to interpret the long-term graphs and would test the individual’s comprehension with directed questions to verbally describe their interpretation of the data. The researcher would correct them and retest them until they could correctly understand what the data was showing. After the individual had a good understanding of how to read the graphs, they were asked to think aloud and point out what they found interesting, unusual, or wanted to ask questions about. These retrospective reflection sessions lasted approximately one hour. During that hour, individuals were confronted with the objective sensor data about their everyday behaviors.

The effects of the real-time feedback and retrospective long-term feedback on the individual’s self-awareness of their behaviors, their subjective ratings of their abilities, and their actual task behaviors were measured beginning from month 3 to the end of the deployment.

## 6.4 Hypotheses

In this study, we address Research Question #5 (RQ5) to investigate the differences between real-time and long-term reflection interactions. Presenting information in real-time has two potential benefits: 1) it is easier for the individual to recall and think about the (current or very recent) context that may have caused an anomaly in the data and 2) providing that immediate feedback can help individuals carry out their tasks understand how they are making progress on their sub-goals of taking their medications today. Goal setting theory expects that more frequent feedback will produce more opportunities to see progress towards the goal and help increase the amount effort devote to the task.

One drawback of real-time feedback is that it only presents the current state and does not typically incorporate historical information about past performance. To make comparisons with previous performance, the individual must use her memory to recall previous instances. Real-time feedback not only lacks information about the past but also makes it difficult to identify trends, particularly subtle trends, in behavioral data. The difficulty in spotting subtle trends in real-time data can be likened to the difficulty that the proverbial frog has in noticing the pot of water it is sitting in is slowly getting hotter before it is cooked (perish the thought!). Without contextualizing data points within a larger trend, the real-time data can appear noisy and either mask trends or make it seem as if changes in abilities are occurring when they are not. Long-term feedback can show the trends over time, properly contextualizing data points among related data points. These trends can enable two important mechanisms of change in the Transtheoretical Model of Behavior Change: dramatic relief and consciousness raising (Prochaska & DiClemente, 1983). The trended data provides a significant amount of concrete information about how an individual has been performing which can be used as a form of dramatic relief leading to a raising of consciousness of a problem, especially if the discrepancy between the individual's self-perception and sensor data is large.

However, as discussed earlier, even though spotting trends and outliers may be easier with long-term representation, explaining them without a clear memory of the larger context of the external factors can be difficult. Thus, this study examines the differences between real-time and long-term feedback in how they affect actual behaviors, the accuracy of self-awareness of these behaviors, and the subjective ratings that individuals have about their abilities to perform these behaviors well. In this study, we consider and evaluate the following hypotheses about behavior change, self-awareness, and subjective self-ratings.

### Behavior Change

H6.1a – Individuals in the real-time feedback group will improve in their behaviors and sustain that behavior because the real-time feedback provides the immediate feedback necessary to correct errors.

H6.1b – Individuals in the long-term feedback group will not show any improvements in their behaviors until the long-term feedback is provided, but after that, the improvement in behavior (as a result of greater awareness) will be short-lived because of the lack of continued feedback.

### Accuracy of Self-awareness of Behaviors

H6.2a – The real-time feedback from the display will increase the accuracy of self-awareness of behaviors for the real-time group and the increased level of accuracy will persist because the real-time feedback continues to re-orient them as their abilities change.

H6.2b – Reflecting on the long-term feedback will increase the accuracy of self-awareness of behaviors for the long-term group but the increased level of accuracy will not remain over time because they lack the continuous feedback.



### Subjective self-ratings of abilities

H6.3 – Individuals in the long-term feedback group will change their subjective ratings of their abilities more than the individuals in the real-time feedback group because the long-term group will be able to notice trends in their behaviors that they did not notice before about themselves.

## 6.5 Measures

In this field study, we consider three types of measure: 1) measures for task behavior 2) measures of self-awareness, that is, how accurately participants reported their behaviors, and 3) measures for the participant's subjective ratings of their ability to perform the task.

### 6.5.1 Measures of behavior

The behavior of focus is how people carry out the task of taking their daily medications/pills. How well people perform their medication taking can be determined by looking at adherence, promptness, correctness, and variance in time taken.

Medication adherence characterizes how often pills are taken and is calculated as a percentage by summing all the pill-taking instances when the individual has taken their pills and dividing the sum by the total number of instances in a given time period.

Promptness characterizes whether the pills are taken before the user-specified time-of-day threshold for late pills. We asked each participant to specify a time of day that they considered later than they normally would take them for their morning and evening pills. The measure of promptness was calculated as a percentage by summing the number of pill-taking instances before the late time and dividing it by the total number of instances over a given period of time.

Correctness characterizes whether the individual took the pills assigned for that particular day. A pill-taking instance was considered correct if the individual opened up the pillbox door that matched the current day of the week. Users could open up any other doors *in addition to* the door that matched the current day and the instance would still be considered correct. An instance would be marked as incorrect only if the user did not open the door that matched the current day of the week but opened other door(s) instead. Opening the correct pillbox door is only a proxy for selecting the correct pills to take because the augmented pillbox does not know what pills are in each slot. We assume that individuals are sorting their pills correctly when refilling their pillbox. More sophisticated sensing such as an ultrasensitive scale was considered for tracking which pills added or removed from the pillbox during pill taking but it was not feasible given the desired form factor.

Variance in the time of day measures how the time of day that medications were taken varied from one day to another within a given period of time. In the case for participants who took both morning and evening pills, we calculated the variance for each of their morning and evening pills and summed the variances. The variance was calculated by converting the time of day to a decimal number between 0 and 27 (with 24 corresponding to 12 midnight of the next day, 25 corresponding to 1am of the next day, *etc.* to account for taking evening pills after midnight) and then taking the variance of the numeric times over a given time period such as a week or month.

### 6.5.2 Measures of accuracy of self-awareness

In addition to measuring how people's behaviors changed over time, we also asked individuals to self-report the frequency and quality of their tasks in a monthly questionnaire (Appendix D). By comparing their self-reported task frequency and quality with task frequency and quality as captured by the sensors, we can calculate the accuracy of their self-awareness of their behaviors. The closer their self-reported frequencies are to the frequencies captured by the

sensors, the more accurate is the individual's self-awareness of their behaviors and abilities. Accuracy of self-awareness is closely related to the concept of insight, the ability to know one's own strengths and weaknesses accurately.

For a measure of the accuracy of self-reported medication adherence, we asked individuals once a month "How many times in the past week did you miss taking a pill?" and took the difference between their self-reported number and number of pills recorded as not taken in the sensor data. Likewise, to get a measure of the accuracy of the self-reported correctness, individuals were asked, "How many times in the past week did you open the wrong pillbox door?" and compared that response with the number of wrong doors opened according to the sensor data. A smaller difference corresponds to a higher accuracy in their self-awareness of medication adherence and correctness.

### 6.5.3 Measures of self-reported subjective ability

In addition to asking participants to self-report the specifics of their task behavior such as missed pills, they were also asked to rate their overall medication-taking ability once a month by answering the question, "How would you rate your overall ability to take your pills correctly in the past week?" Participants responded using a 6-point Likert scale: Very Poor, Poor, Fair, Good, Very Good, and Excellent. Participants also rated their promptness of their medication taking by responding to the question, "How would you rate how ontime you were in the past week?" Participants responded using a 7-point Likert scale: Never, Rarely, Occasionally, Sometimes, Frequently, Usually, Always.

## 6.6 Results

The dwellSense system was deployed for over eight months in the homes of 12 older adults. 2 individuals from the original 14 recruited and consented into the study had to withdraw from the study because they were unable to integrate our instrumented pillboxes into their routines. The remaining 12 individuals participated to the end of the study.

The results provide support for the statistically significant behavior change and sustained behavior in the real-time group as predicted in hypothesis H6.1a. However, the magnitude of the behavior change for the long-term group after their long-term reflection session did not reach statistical significance, but the trends in behavior are consistent with hypothesis H6.1b showing that there was a small improvement just after reflection but that improvement did not last longer than three weeks.

The results also provide mixed results about improvements in the accuracy of the individual's self-awareness. The results show that the individuals in the real-time group did not make any statistically significant changes in the accuracy of their self-reported mistakes (missing pills and opening up the wrong pillbox door), leaving hypothesis H6.2a without support. However, the long-term group did significantly reduce the number of self-reported missed pills immediately after the retrospective reflection session providing some support for hypothesis H6.2b, but this reduction in the number of errors did not persist into the follow up period, six weeks after the reflection session.

In terms of how the different forms of reflection affected the individual's self-rating of their abilities to carry out IADLs, the results support hypothesis H6.3, showing that while both groups changed their ratings after the introduction of their respective reflection interventions, the long-term group changed their ratings to a greater degree than the real-time group.

### 6.6.1 Reflection and Behavior Change

To test whether feedback (real-time and long-term) had an effect on pill-taking behavior (hypotheses H6.1a and H6.1b), we will focus on the 18-week time period (Table 6-3) immediately before, during, and immediately after the introduction of the feedback. We segmented the data from this 18-week period into three 6-week windows and then further divided each 6-week window into two 3-week halves to increase the number of data points for statistical testing.

The rate of adherence, rate of promptness, rate of correctness, and variance in the time of day was calculated for each 3-week window. 3-week chunks were chosen because it offered the best compromise between maximizing the number of data points (windows) and noisiness in the data.

The first 6-week window (comprising two 3-week chunks) represents a BASELINE phase where both groups received no intervention. Just after this first 6-week window, the real-time group received the real-time tablet display in their homes and continued using it until the end of the study. The second 6-week window represents the DISPLAY phase where the real-time group has the real-time display but the long-term group does not receive any informational intervention but instead carries on undisturbed as they were in the baseline period. In this way, the long-term group acts as the control for the real-time group during the DISPLAY phase. Just after the DISPLAY period, the long-term group reflected on 6-8 weeks of their pill-taking data. The subsequent 6-week period following this reflection session is called the POST-REFLECTION phase. Thus, we should expect to see each individual's baseline level of performance in the BASELINE phase, the effect of the real-time tablet display for the real-time group in the DISPLAY phase, and the effect of the retrospective reflection session for the long-term group in the POST-REFLECTION phase.

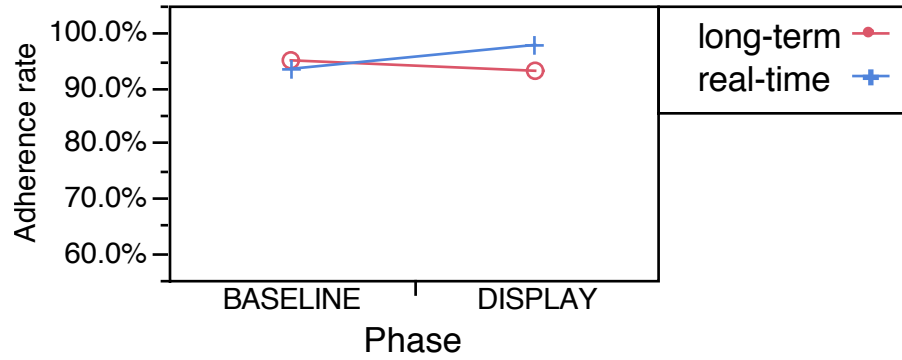
Phase	BASELINE						DISPLAY						POST- REFLECTION							
Sub-phase													IMMEDIATE			FOLLOWUP				
Week #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
Real-time group	nothing			nothing			Introduce display	display			display			display			display			
Long-term group	nothing			nothing				nothing			nothing			Long-term reflection	nothing			nothing		

**Table 6-3. Timeline for deployment. The first six weeks make up the BASELINE phase for both groups. Then the display is introduced to the real-time group during the DISPLAY phase. Six weeks later, the long-term group reviews a long-term view of the data, after which both groups are in the POST-REFLECTION phase, which is divided into IMMEDIATE and FOLLOWUP sub-phases.**

All analyses were performed using a one-way ANOVA repeated-measures design and the Huynh-Feldt correction when data variables did not satisfy conditions of sphericity (as determined by a significant effect with Mauchly's test). The results focus on the interaction effect between condition and phase, that is, whether the real-time and long-term conditions differ in how they change over time, as each acted as a control for the other condition at various time points in the study. Support for hypotheses H6.1a and H6.1b can be found if there is a significant interaction effect between condition and phase, with contrasts showing a change in the condition receiving the intervention. Overall, we identified that the real-time display had a significant impact on the real-time group's adherence, promptness, correctness, lateness, and variance in time taken (supporting H6.1a) whereas retrospective reflection had an effect only on promptness for the long-term group (providing only limited support for H6.1b).

#### 6.6.1.1 Real-time feedback and Adherence

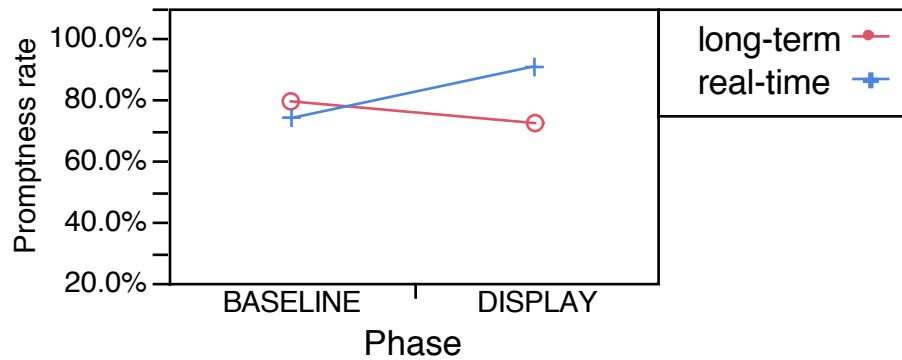
We predicted in H6.1a that the real-time feedback will help individuals increase their medication adherence rate and indeed the analysis reveals a (marginally) significant interaction effect ( $F[1,31]=2.94, p=0.0966$ ) between condition and phase when comparing the adherence rates in the BASELINE and DISPLAY phases for both the real-time and long-term conditions (Figure 6-3). The real-time group increased from a mean of 93.0% to 97.8% (a marginally significant contrast with  $F[1,31]=3.13, p=0.0866$ ) whereas the long-term group basically went unchanged from a mean of 95.0% to 93.5% (contrast was insignificant,  $p>0.10$ ). These results support H6.1a in that the real-time display helped improve the real-time group's adherence rate. Having the display either reminded them to take their medications or seeing the feedback helped individuals make a more conscious effort to not miss their medications.



**Figure 6-3. The significant interaction effect shows that real-time group increased in their medication adherence rate with the introduction of the real-time feedback display, while the long-term group stayed relatively unchanged because they received no intervention across these phases.**

#### 6.6.1.2 Real-time feedback and Promptness

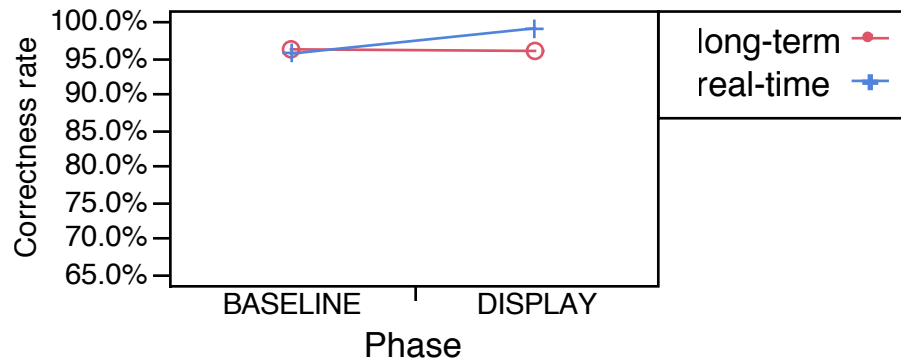
We predicted in H6.1a that the real-time feedback will help individuals increase their medication promptness rate and indeed the analysis reveals a significant interaction effect ( $F[1,31]=12.34, p=0.0014$ ) between condition and phase when comparing the adherence rates in the BASELINE and DISPLAY phases for both the real-time and long-term conditions (Figure 6-4). The real-time group increased from a mean of 74.0% to 91.1% (a significant contrast with  $F[1,31]=13.448, p=0.0001$ ) whereas the long-term group went from a mean of 79.5% to 72.4% (contrast was insignificant,  $p>0.10$ ). These results support H6.1a in that the real-time display helped the real-time group to take more of their medications on-time rather than late. Seeing the immediate feedback helped individuals make a more conscious effort to take their medications before a “late time” predefined by each individual.



**Figure 6-4. The significant interaction effect shows that real-time group increased in their medication promptness rate (took more of the pills ontime) with the introduction of the real-time feedback display, while the long-term group stayed relatively unchanged because they received no intervention across these phases.**

#### 6.6.1.3 Real-time feedback and Correctness

We predicted in H6.1a that the real-time feedback will help individuals increase their medication promptness rate but the analysis reveals no significant interaction effect ( $F[1,31]=1.18, p=0.2857$ ) between condition and phase when comparing the correctness rates in the BASELINE and DISPLAY phases for both the real-time and long-term conditions (Figure 6-5). There was a slight increase in correctness from the baseline to display phases for the real-time group but this difference was not statistically significant. The real-time group increased from a mean of 95.5% to 99.0% whereas the long-term group basically was unchanged from a mean of 96.1% to 95.9%, however, neither of these contrasts was statistically significant. The increase experienced by the real-time group was not large enough to support hypothesis H6.1a in increasing correctness (opening the correct pill box door) for the real-time group with the introduction of the real-time display. Individuals may not have been concerned with opening the correct pillbox door

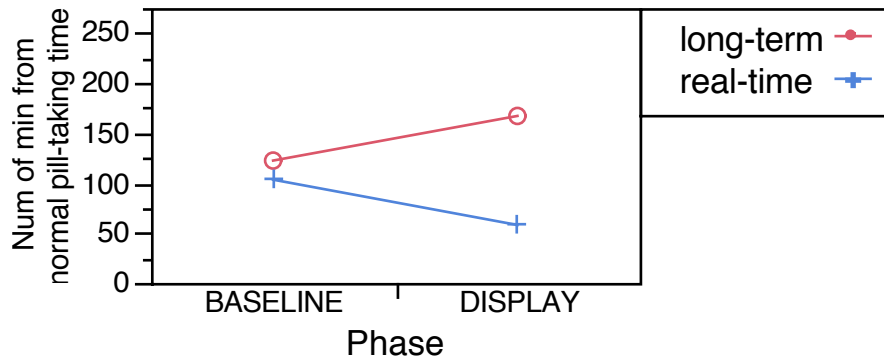


**Figure 6-5. No significant interaction effect between condition and phase in the percentage of correctly opened pillbox doors. The real-time display had little effect on whether individuals opened the correct door or not.**

because they had the same pills for every day of the week.

#### 6.6.1.4 Real-time feedback and Lateness

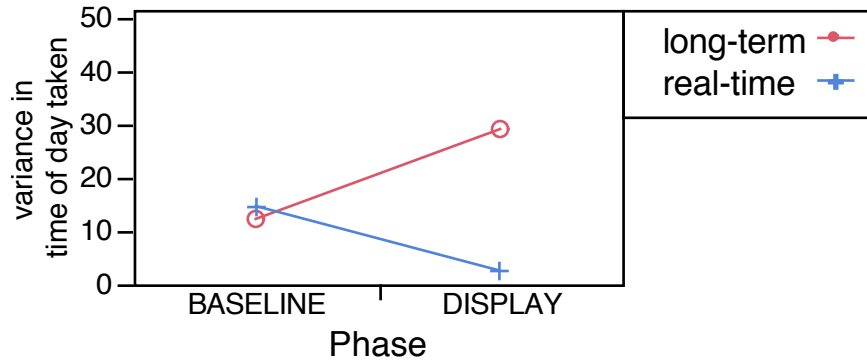
We predicted in H6.1a that the real-time feedback will help individuals decrease the time difference between the time people take their medications and the time they are (self-reportedly) supposed to take their medications. The analysis reveals a significant interaction effect ( $F[1,31]=32.63, p<0.001$ ) between condition and phase when comparing the adherence rates in the BASELINE and DISPLAY phases for both the real-time and long-term conditions. The real-time group increased from a mean of 103.3 minutes to 57.98 minutes (a significant contrast with  $F[1,31]=12.73, p=0.0006$ ) whereas the long-term group went from a mean of 122.43 minutes to 167.10 minutes (a significant contrast,  $F[1,31]=18.22, p=0.0002$ ). These results support H6.1a in that the real-time display helped reduce the real-time group's lateness. The presence of the display helped individuals take their medications closer to the time they were supposed to take them. It is interesting that this effect happened even though the display did not visually compare people's medication taking time with the time they are supposed to take it. The display does show the time of day the medications were taken, and this is probably adequate information for the individual to compare that time with the time they predefined as the time they are supposed to take the medication. We also observed a significant increase in the lateness for the long-term group, which is a little surprising. However, we attribute this increase in lateness to either a decline in the functional abilities associated with aging or a pattern showing that the individuals in the long-term group are "relaxing" and getting sloppier in their medication routines, which often happens during the course of an observational study.



**Figure 6-6** The significant interaction effect shows that average difference between time the real-time group took their pills and the time they are supposed to take their pills decreased with the introduction of the real-time feedback display, while the long-term group stayed relatively unchanged because they received no intervention across these phases.

#### 6.6.1.5 Real-time feedback and Variance (in the Time of Day)

We predicted in H6.1a that the real-time feedback will help individuals decrease the variance in the time of day each individual took his/her medications. A decrease in the variance in the time of day the medications are taken indicates that the individual is taking them at more consistently and the same time of day everyday. The analysis reveals a significant interaction effect ( $F[1,31]=11.51, p=0.009$ ) between condition and phase when comparing the adherence rates in the BASELINE and DISPLAY phases for both the real-time and long-term conditions. The real-time group decreased from a mean of 14.68 to 2.56 (a significant contrast with  $F[1,31]=7.14, p=0.012$ ) whereas the long-term group basically actually increased from a mean of 12.26 to 29.11 (with a significant contrast  $F[1,31]=4.43, p=0.0435$ ). These results support H6.1a in that the real-time display helped improve the real-time group’s variance in the time of day they are taking their medications. Having the display and seeing the feedback appears to have helped individuals make a more conscious effort to take their medications more consistently at the same time of day from one day to the next. The significant increase in the variance for the long-term group from the BASELINE to DISPLAY phases was somewhat surprising because the long-term received no interventions during these phases. Similarly to the findings regarding lateness, we attribute this increase in variance to the individuals in the long-term group “relaxing” in the study and getting sloppier in their medication routines, which often happens during the course of an observational study. The linear measure of lateness and the variance in the time of day may both be particularly sensitive measures for how much effort and attention people are paying to their medication routines.



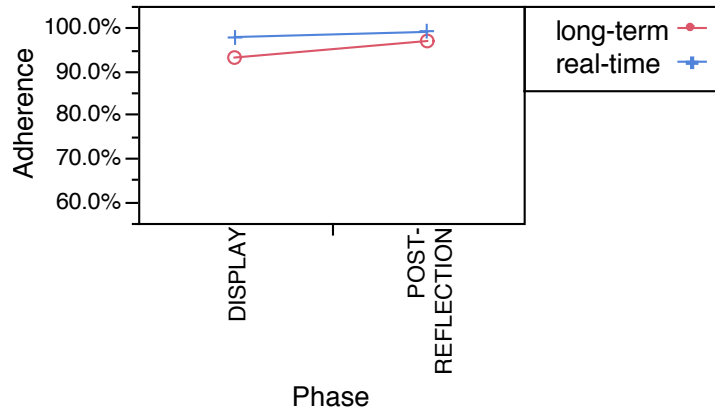
**Figure 6-7. The significant interaction effect shows that real-time group were more consistent in the time of day they took their pills (the variance in the time of day was smaller) after the introduction of the real-time feedback display, while the long-term group appears to grown less consistent over time.**

#### 6.6.1.6 Long-term feedback and Behavior Change

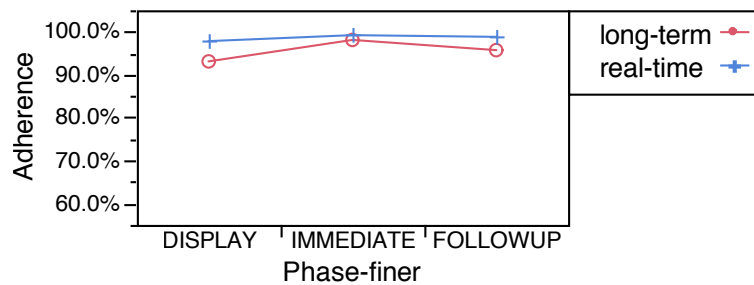
This section provides a summary of the upcoming five sections about long-term feedback and behavioral measures. The results show that reflecting on long-term data affects only promptness significantly. In an effort to identify whether there are short-term improvements in behavior masked by treating the POST-REFLECTION phase as one monolithic phase, we can break the 6-week POST-REFLECTION phase into two 3-week phases, IMMEDIATE and FOLLOWUP. Short-lived, immediate effects after the reflection session might be found in the IMMEDIATE phase but not persist into the FOLLOWUP phase. This finer-grained analysis measures of medication behavior (adherence, closeness, correctness, and lateness, but not including variance in the time of day) reveals a common “v-shape” pattern where the medication behavior is improved in the IMMEDIATE phase but the measure drops back to a level closer to the DISPLAY phase, likely due to the fact that they did not receive subsequent, frequent feedback to reinforce their improved behaviors.

#### 6.6.1.7 Long-term feedback and Adherence

We predicted in H6.1b that the long-term reflection session will help individuals in the long-term group increase their medication adherence rate but the analysis (Figure 6-9) reveals no significant interaction effect ( $F[1,31]=0.418$ ,  $p=0.5225$ ) between condition and phase when comparing the adherence rates in the DISPLAY and POST-REFLECTION phases for both the real-time and long-term conditions. Both groups increased in adherence between these two phases, with the long-term group going from a mean of 93.1% to 96.9% whereas the real-time group went from a mean of 97.8% to 99.0%, but these increases were not statistically significant. A finer-grained analysis (Figure 6-8) reveals an interesting temporal pattern of higher adherence in the IMMEDIATE phase (mean 99.2%,  $SE=2.8\%$ ) than in the DISPLAY (mean 97.8%,  $SE=2.0\%$ ) and FOLLOWUP (mean 98.8%,  $SE=2.8\%$ ) phases, but the differences among these phases were not statistically significant. This similar pattern occurs not only for adherence, but also for promptness, correctness, and lateness. Thus, reflecting on 6-8 weeks of historical data about their pill-taking behaviors does not appear to have motivated individuals in the long-term group enough to make a difference in being diligent in taking their medications everyday.



**Figure 6-9.** There was no significant interaction effect between phase and condition for the adherence rate when the long-term group reflected on 6-8 weeks of their own data. Reflecting on long-term data did not seem to change how often the long-term group took their pills.

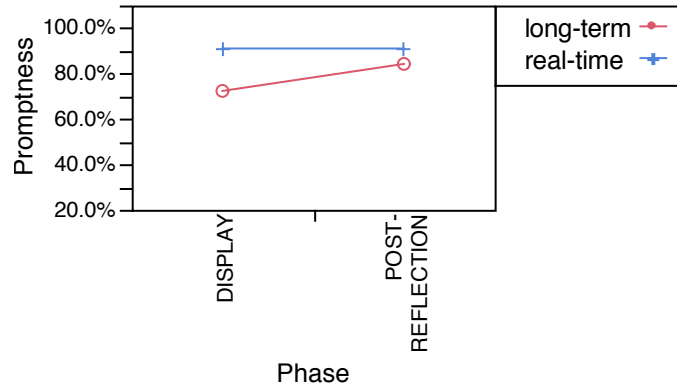


**Figure 6-8.** A finer-grained analysis of the phase following the long-term reflection session shows a slight increase in adherence rate during the IMMEDIATE phase (the 3 weeks directly following the session) but it drops back to baseline levels during the FOLLOWUP phases (3-6 weeks after the session).

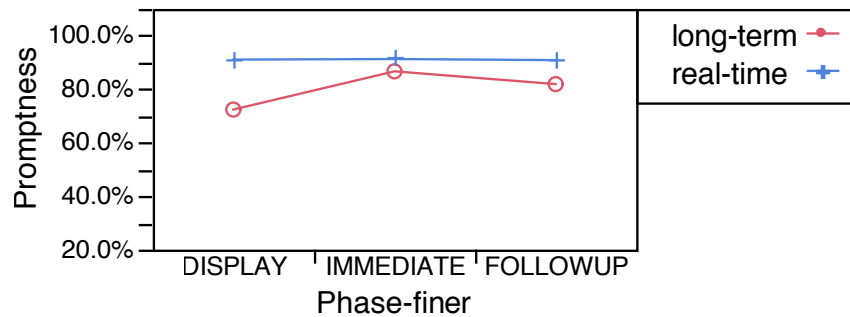
#### 6.6.1.8 Long-term feedback and Promptness

We predicted in H6.1b that the long-term reflection session will help individuals in the long-term group increase their medication promptness rate but the analysis reveals a marginally significant interaction effect ( $F[1,31]=2.92, p=0.0973$ ) between condition and phase when comparing the promptness rates in the DISPLAY and POST-REFLECTION phases for both the real-time and long-term conditions (Figure 6-11). There was an increase in adherence for the long-term group from a mean of 72.4% to 86.7% (a significant contrast,  $F[1,20]=5.16, p=0.0343$ ) whereas the real-time group was unchanged from a mean of 91.1% to 91.3%, as predicted due to the fact that the real-time group continued on with the display in both phases. Reflecting on 6-8 weeks of historical data about their pilltaking behaviors appears to have motivated individuals in the long-term group to be more on time when taking their medications in the 6 weeks following the reflection session.





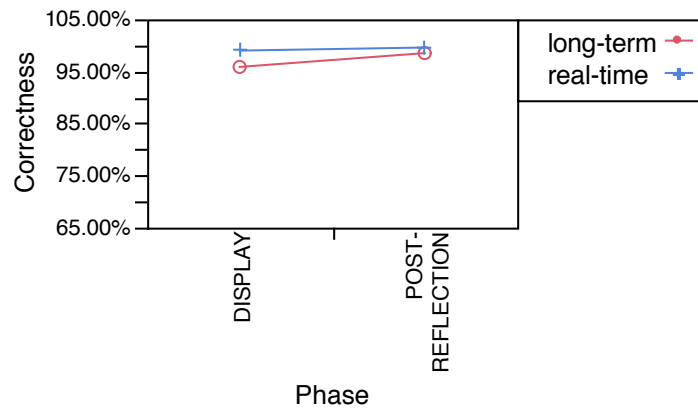
**Figure 6-11.** There was marginally significant interaction effect between phase and condition for promptness when the long-term group reflected on 6-8 weeks of their own data. The long-term group increased in promptness after reflecting on their data (a significant contrast).



**Figure 6-10.** A finer-grained analysis of the phase following the long-term reflection session shows an significant increase in promptness during the IMMEDIATE phase (the 3 weeks directly following the session) but the increase is short-lived and promptness drops slightly during the FOLLOWUP phases (3-6 weeks after the session).

#### 6.6.1.9 Long-term feedback and Correctness

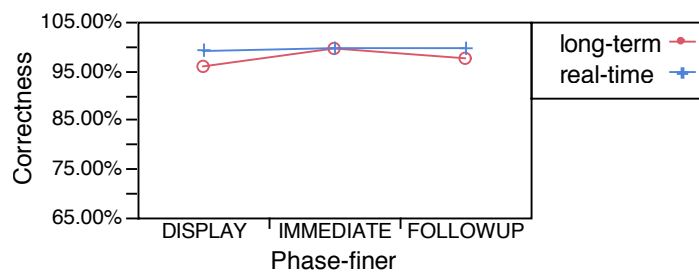
We predicted in H6.1b that the long-term reflection session will help individuals in the long-term group increase their medication correctness rate but the analysis (Figure 6-12) reveals no significant interaction effect ( $F[1,31]=0.4886$ ,  $p=0.4898$ ) between condition and phase when comparing the correctness rates in the DISPLAY and POST-REFLECTION phases for both the real-time and long-term conditions. Both groups increased in correctness between these two phases, with the long-term group going from a mean of 95.9% (SE=1.6%) to 98.5% (SE=1.6%) whereas the real-time group went from a mean of 99.0% (SE=1.4%) to 99.5% (SE=1.4%), but these increases were not statistically significant. A finer-grained analysis (Figure 6-13) reveals an interesting temporal “v-shape” pattern of higher correctness in the IMMEDIATE phase (mean 99.5%, SE=2.2%) than in the DISPLAY (mean 95.9%, SE=1.6%) and FOLLOWUP (mean 97.5%, SE=2.2%) phases, but the differences among these phases were not statistically significant. Thus, reflecting on 6-8 weeks of historical data about their pill-taking behaviors does not appear to have motivated individuals in the long-term group enough to make a difference in being more diligent in opening the correct pillbox doors.



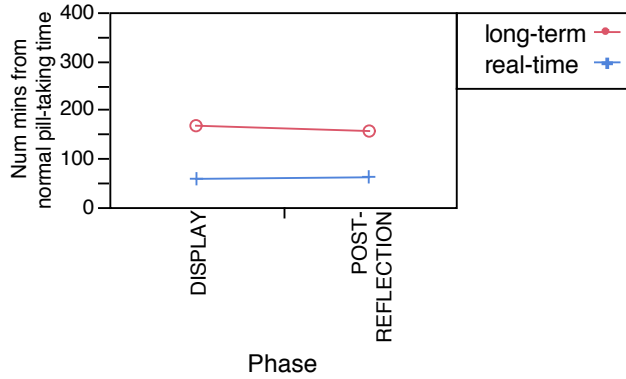
**Figure 6-12.** There was no significant interaction effect between phase and condition for the correctness rate when the long-term group reflected on 6-8 weeks of their own data. Reflecting on long-term data did not seem to change whether the long-term group opened the correct pillbox door.

#### 6.6.1.10 Long-term feedback and Lateness

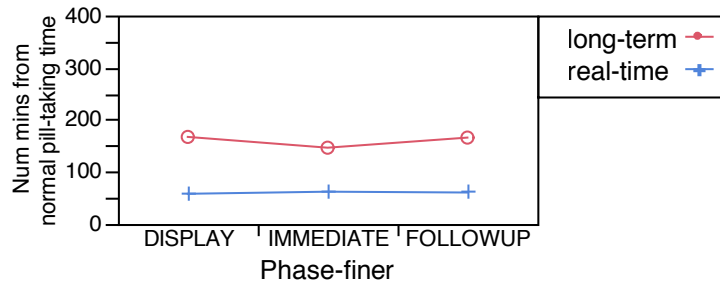
We predicted in H6.1b that the long-term reflection session will help individuals in the long-term group decrease their medication lateness but the analysis (Figure 6-15) reveals no significant interaction effect ( $F[1,31]=0.4151$ ,  $p=0.5241$ ) between condition and phase when comparing the correctness rates in the DISPLAY and POST-REFLECTION phases for both the real-time and long-term conditions. The long-term group went from a mean of 167.1 minutes ( $SE=26$ ) to 156.1 minutes ( $SE=26$ ) whereas the real-time group went from a mean of 57.97 minutes ( $SE=23.7$ ) to 61.22 minutes ( $SE=23.7$ ) but these changes were not statistically significant. A finer-grained analysis (Figure 6-14) reveals an interesting temporal “v-shape” pattern of less lateness in the IMMEDIATE phase (mean 146.4 minutes,  $SE=28.6$ ) than in the DISPLAY (mean 167.1 minutes,  $SE=26$ ) and FOLLOWUP (mean 165.8 minutes,  $SE=28.6$ ) phases, but the differences among these phases were not statistically significant. Thus, reflecting on 6-8 weeks of historical data about their pill-taking behaviors does not appear to have motivated individuals in the long-term group enough to make a difference in being more diligent in taking their pills closer to time they are supposed to take it.



**Figure 6-13.** A finer-grained analysis of the phase following the long-term reflection session shows a slight increase in correctness rate during the IMMEDIATE phase (the 3 weeks directly following the session) but it drops back to baseline levels during the FOLLOWUP phases (3-6 weeks after the session).



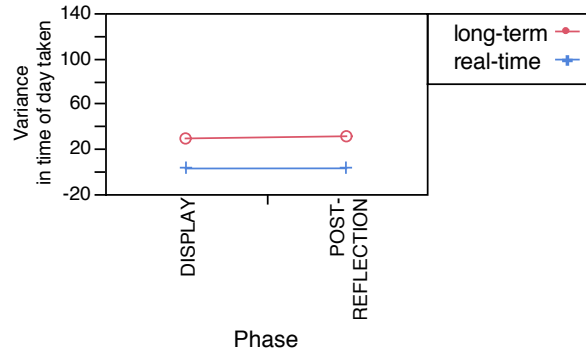
**Figure 6-15.** There was no significant interaction effect between phase and condition for the lateness when the long-term group reflected on 6-8 weeks of their own data. Reflecting on long-term data did not seem to change whether the long-term group took their pills closer to the time they should normally take it.



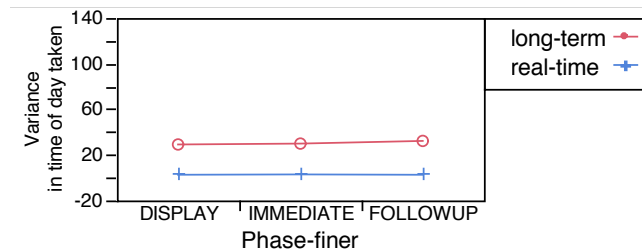
**Figure 6-14** A finer-grained analysis of the phase following the long-term reflection session shows a slight decrease in the lateness during the IMMEDIATE phase (the 3 weeks directly following the session) but it increases back to baseline levels during the FOLLOWUP phases (3-6 weeks after the session).

#### 6.6.1.11 Long-term feedback and Variance

We predicted in H6.1b that the long-term reflection session will help individuals in the long-term group decrease their variance in the time of day they are taking their medications but the analysis (Figure 6-17) reveals no significant change in variance. A finer grained analysis (Figure 6-18) with the POST-REFLECTION phase separated into IMMEDIATE and FOLLOWUP phases similarly shows that reflecting on the retrospective long-term data did not affect the individual's variance in the time of day she took her pills from day to day.



**Figure 6-17. There was no significant interaction effect between phase and condition for the variance in the time of day the pills were taken after the long-term group reflected on 6-8 weeks of their own data. Reflecting on long-term data did not seem to change whether the long-term group took their pills more consistently at the same time of day.**



**Figure 6-16. A finer grained analysis reveals no further trends than the coarser analysis. The variance (or spread) of the times of day pills were taken did not change as a result of reflecting on long term data.**

#### 6.6.1.12 Summary of Support for Hypotheses H6.1a & H6.1b

Overall, we identified that the real-time display had a significant impact on the real-time group's adherence, promptness, correctness, lateness, and variance in time taken (supporting H6.1a) whereas retrospective reflection had an effect only on promptness for the long-term group (providing only limited support for H6.1b).

<b>H6.1a</b>	Interaction effect between condition and phase. (In other words, did real-time display have any effect?)	Trend, as expected, but not statistically significant
ADHERENCE	Yes (p = 0.0014)	n/a
ONTIME	Yes, (p = 0.0966)	n/a
CORRECTNESS	No	Yes
LATENESS	Yes (p<0.001)	n/a
VARIANCE	Yes (p=0.0019)	n/a

**Table 6-4. Summary of findings in support for Hypothesis H6.1a. Overall, the results show that the real-time display did have significant impacts on how often people took their meds, whether they took them on time, how close to their normal time they took their pills, and how consistently they took them with respect to the time of day. The only aspect of medication taking that did not reach a level of significance was correctness, though the trend was consistent with the other aspects.**

<b>H6.1b</b>	Interaction effect between condition and phase. (In other words, did reflecting on long-term data have any effect?)	Temporary change only in IMMEDIATE phase (In other words, did reflecting on long-term data have at least a <i>temporary</i> effect?)
ADHERENCE	No	Yes
ONTIME	Yes (p=0.0973)	
CORRECTNESS	No	Yes
LATENESS	No	Yes
VARIANCE	No	No

**Table 6-5. Summary of findings for Hypothesis 6.1b. Overall, the effect of reflecting on long-term (6-8 weeks of) data was somewhat weak. It only significantly affect whether individuals took their pills on time or late. However, in the other measures, the results show a temporary increase in performance in the three weeks following the reflection session but this increase often does not persist beyond a three weeks.**

## 6.6.2 Accuracy of Self-Awareness

We measure the accuracy of self-awareness by taking the difference between the self-reported number of task errors and the number of errors calculated from the sensor data. The self-reported number of task errors (such as the number of instances they did not take their medications or opened the wrong pillbox door) was found by asking people in a monthly questionnaire to self-report their estimate from the previous week. The smaller the difference between the self-reported value and the sensed value, the greater the accuracy in the individual's self-awareness of her behaviors. We consider two features of a medication-taking routine: self-awareness of adherence and self-awareness of correctness.

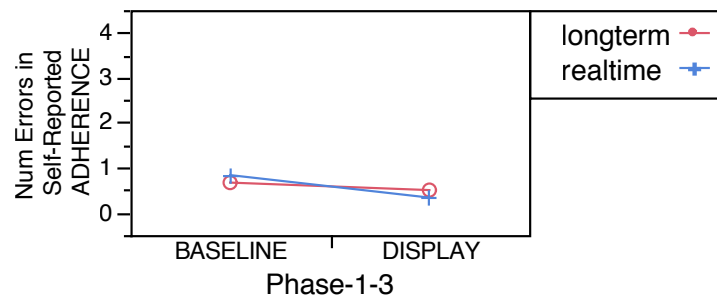
The results show that the real-time feedback display did not directly change the individual's self-awareness of their medication adherence mistakes (providing little support for H6.2a), but the long-term reflection did result in a significant increase in the accuracy of their self-awareness of their medication adherence mistakes (providing support for H6.2b). Moreover, neither the real-time nor long-term reflection interactions had any impact on how self-aware individuals were with about opening the wrong pillbox doors.

### 6.6.2.1 Self-awareness of medication adherence

Individuals were each asked to report how many times in the previous week she missed taking her medications. This number was compared with the number of missed days in the sensor data. The absolute value of the difference between these two numbers provides a measure of the self-awareness of medication adherence. Under ideal circumstances, the

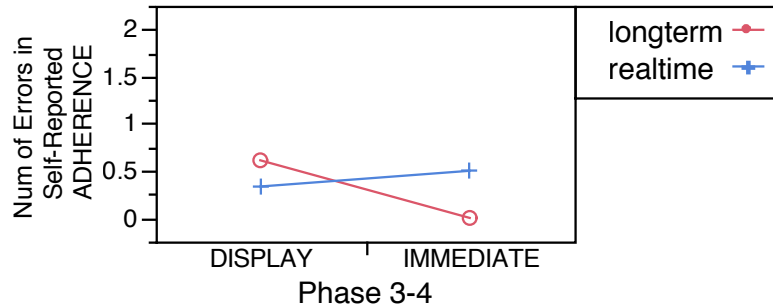
individual's self-report of the number of missed pills should match the number calculated from the sensor data, resulting in a difference of zero.

The real-time display appeared to have little effect in the self-reported errors in medication adherence, providing little support for hypothesis H6.2a. Analysis of the BASELINE to DISPLAY phases shows no significant interaction effect between condition and phase ( $F[1,22]=0.361$ ,  $p=0.5543$ ) which means that the group in the long-term condition (as the control) and the group in the real-time condition varied very similarly across these two phases (Figure 6-18). This result can be interpreted that the real-time group were just as accurate or inaccurate in their self-reported medication adherence errors with and without the display. However, as discussed in 6.6.1.1, the real-time group actually improved their medication adherence behavior as a result of the real-time display, moving from a 93% to 97.8% adherence rate. Despite this increase, the self-reports of individuals in the real-time group still had a similar discrepancy before and after the introduction of the real-time display. This interesting result shows that even though the real-time group did not improve in the accuracy of self-reports for medication adherence, they were not any worse, at least statistically. This means that with the display, individuals reported fewer missed medications, but their estimate of how many medications they missed was still off by the same amount before and after the introduction of the display. Therefore, the display plays a role in motivating behavior change (increase adherence), but it may also provide a commensurate level of awareness that allows individuals to notice their improvement. However, the level of awareness does not make them more accurate, but simply allows them to maintain their level of accuracy as they improve their pill-taking performance.



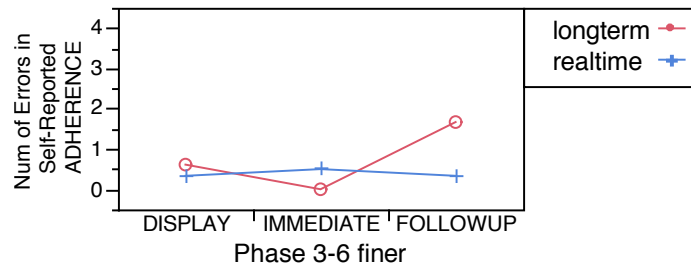
**Figure 6-18. The lack of an interaction effect between condition and phase shows that the introduction of the real-time feedback display did not have much of an effect on how accurate individuals were when reporting the number of medications they missed taking.**

Considering whether the results provide support for hypothesis H6.2b, we found that reflecting on long-term feedback was helpful in reducing the difference between the number of self-reported medications not taken and the number of medications not taken detected by dwellSense. In the analysis of the DISPLAY and IMMEDIATE phases (Figure 6-19), there was a significant interaction effect between condition and phase ( $F[1,22]=4.42$ ,  $p=0.0468$ ), with the long-term group decreasing the difference between the number of self-reported missed pills and the number detected by the system and the real-time group not changing much (Figure 6-19). The long-term group went from a mean difference of 0.61 (SE=0.14) to 0 (SE=0.22) (a significant contrast  $F[1,22.9]=6.81$ ,  $p=0.0157$ ) and the real-time remained relatively unchanged from mean difference of 0.33 (SE=0.22) to 0.50 (SE=0.22). The real-time group, acting as a control, did not change from the DISPLAY to the IMMEDIATE phases because no new intervention was introduced during this time. However, the long-term group was more accurate after reflecting on their long-term data in the IMMEDIATE phase, which supports hypothesis H6.2b in that reflection on long-term data supports a more accurate self-awareness.



**Figure 6-19. Reflecting on long-term data results more accurate self-awareness of the adherence rate for the long-term group when rated right after the reflection session. As expected, the real-time group does not change much during these phases.**

Further analysis of the FOLLOWUP phase, shows that the long-term group dramatically decreased the accuracy of their self-reported medication adherence, that is, there was an increase in the difference between their self-reported medication adherence and the sensor data. There was an interaction effect (Figure 6-20) between condition and phase ( $F[2,45.84]=6.39$ ,  $p=0.0036$ ) where the real-time group remained basically unchanged while the long-term group decreased in the IMMEDIATE phase and increased in FOLLOWUP phase. The long-term went from a BASELINE mean of 0.61 (SE=0.19), to an IMMEDIATE mean of basically 0.0 (SE=0.32) (a marginally significant contrast from BASELINE of  $F[1,45.8]=2.93$ ,  $p=0.0933$ ) to a FOLLOWUP mean of 1.67 (SE=0.23) (a significant contrast from the IMMEDIATE phase of  $F[1,45.8]=19.42$ ,  $p<0.0001$ ). The real-time group went from a BASELINE mean of 0.33 (SE=0.32), to an IMMEDIATE mean of 0.50 (SE=0.32) to a FOLLOWUP mean of 0.33 (SE=0.23). There were no significant contrasts in the real-time group. The increase in error in self-reported medication adherence indicates that the impact of improved self-awareness following reflecting on long-term data is temporary and that extra



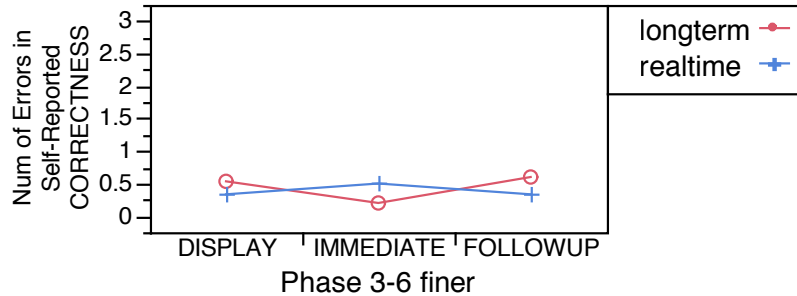
**Figure 6-20. The increase in error in self-reported medication adherence indicates that the impact of improved self-awareness following reflecting on long-term data is temporary and that extra insight does not persist beyond immediately after the reflection session.**

insight does not persist beyond immediately after the reflection session. In contrast, the real-time group did remain relatively at a low error rate through the FOLLOWUP phase, indicating that the real-time feedback continuously reinforces the same (small) level of self-awareness of medication adherence mistakes.

#### 6.6.2.2 Self-awareness of medication correctness

In addition to measuring how accurately people reported their medication non-adherence (not taking pills), we also considered how accurately they reported opening the wrong door on the pillbox as another facet of self-awareness. However, the analysis of the results revealed no significant changes in the accuracy of self-reported medication correctness for the long-term and real-time groups across the phases. The group in the long-term condition went from a BASELINE of mean difference of 0.53 (SE=0.20) to 0.20 (SE=0.31) in the IMMEDIATE phase, and to 0.60 (SE=0.24) in the FOLLOWUP phase. Similarly, the group in the real-time condition also did not change much going

from a mean difference of 0.33 (SE=0.31) to 0.50 (SE=0.31) in the IMMEDIATE phase, and to 0.33 (SE=0.22) in the FOLLOWUP phase. There was no interaction effect between condition and phase ( $F[2,46]=0.222$ ,  $p=0.802$ ). These results show that neither the real-time display nor the long-term reflection sessions had any significant effect on how accurately people reported opening the wrong door on their pillboxes.



**Figure 6-21. Neither the real-time display nor the long-term reflection session had any significant effect on how accurately individuals reported opening the wrong door on their pillboxes.**

### 6.6.2.3 Summary of Support for Hypotheses 6.2a and 6.2b

The results show that the real-time feedback display did not directly change the individual’s self-awareness of their medication adherence mistakes (providing little support for H6.2a), but the long-term reflection did result in a significant increase in the accuracy of their self-awareness of their medication adherence mistakes (providing support for H6.2b). Moreover, neither the real-time nor long-term reflection interactions had any impact on how self-aware individuals were with about opening the wrong pillbox doors.

<b>H6.2a</b>	Interaction effect between condition and phase. (In other words, did real-time display have any effect on the accuracy of self-awareness?)
ADHERENCE	No, but given that individuals in the real-time improved in the task performance (by skipping fewer meds), the display may have supported a commensurate level of awareness since their error rate did not decrease.
CORRECTNESS	No
CONCLUSION	The real-time display had no significant effect on how accurate individuals were able to report about the medication routine.

**Table 6-7. Summary of support for Hypothesis 6.2a**

<b>H6.2b</b>	Interaction effect between condition and phase. (In other words, did long-term reflection have any effect on the accuracy of self-awareness?)
ADHERENCE	Yes ( $p=0.0036$ ), but the effect was temporary, as the error increased in 3 weeks after the reflection session.
CORRECTNESS	No
CONCLUSION	The long-term reflection session helped individuals increase the accuracy of their self-awareness (particularly about adherence) but the increase in accuracy did not last beyond 3 weeks after the reflection session.

**Table 6-6. Summary of support for Hypothesis 6.2b**



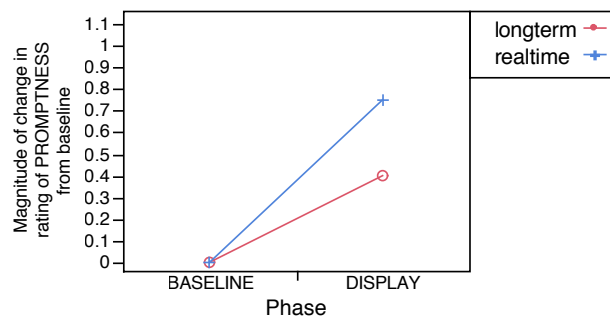
### 6.6.3 Self-Ratings of Abilities

In addition to behavior change and self-awareness, another important aspect of successful aging is knowing your own abilities to perform tasks important for independence. We asked participants to rate (subjectively) their own abilities to take their medications regularly (adherence) and promptly (promptness). After the introduction of either the real-time display or the long-term reflection session, changes (whether increases or decreases) in their ratings of the abilities can indicate that the objective sensor data had an influence on their perception of their abilities.

In this section, we quantify these changes to see whether the data had a subjective influence on the individual's perception of their abilities. Every month, individuals rated their ability to take medications promptly on a Likert scale with 7 response options and their ability to take medications everyday on a Likert scale with 6 responses. We consider ratings taken over the course of 6 months, with the first rating taken in month 2 of the study (just prior the introduction of the real-time display for the real-time group) and the last rating taken in month 7 of the study. The focus of the analysis in this section is in how individuals changed their ratings over time and as a result of the long-term or real-time reflection interactions, and thus, the normalized differences from the initial baseline (pre-intervention) rating are calculated for each month and the absolute values of the differences are analyzed. For example, if an individual rates his promptness over 6 months as: 3, 2, 3, 4, 5, and 3, then the absolute values from the initial rating of ( $r_0=3$ ) would be  $\text{abs}(r_0-3)=0$ ,  $\text{abs}(r_0-2)=1$ , 0, 1, 2, 0. Larger deviations from the initial rating indicate that the individual re-assessed her abilities to a greater degree. The direction of change is not as relevant as the magnitude, because the objective sensor data, depending on whether it shows behaviors that match the expectations of the individual, can trigger individuals to reassess themselves either as better or worse than they thought before. In the following sections, we examine how individuals changed their subjective ratings of their abilities to take their medications promptly and consistently. The pattern predicted by hypothesis H6.3 was found in the ratings for medication promptness but not in medication adherence. Individuals in the real-time condition did change their ratings from baseline as a result of the introduction of the real-time feedback display, but after the initial month of introduction, the ratings stayed relatively stable. However, individuals in the long-term condition, after the reflection session, tended to change their ratings of their ability to take their medications promptly to a much larger degree than individuals in the real-time group because the long-term view provided trend data useful for an individual to reassess their abilities.

#### 6.6.3.1 Ratings of ability to be prompt in medication taking

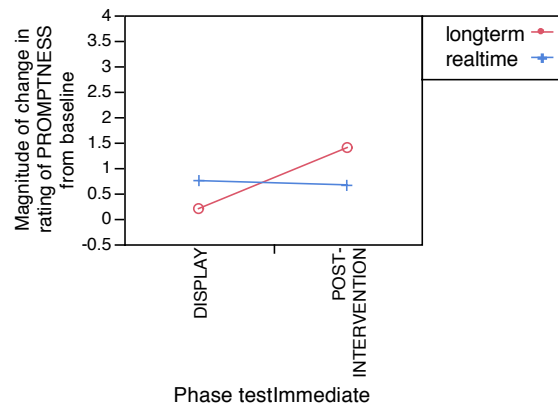
The real-time feedback did trigger individuals in the real-time condition to change their ratings. The results show an interaction effect between condition and phase ( $F[1,9]=5.22$ ,  $p=0.0480$ ), showing that the group in the real-time condition changed their rating significantly more than the group in the long-term condition when going from the



**Figure 6-22. The significant interaction effect between phase and condition shows that the real-time group changed their rating much more than the long-term group after the real-time feedback display was introduced.**

BASELINE to the DISPLAY phase (Figure 6-22). The real-time group went from a mean difference in rating of 0.0 (SE=0.080) to 0.70 (SE=0.073) and the long-term group went from a mean rating of 0.0 (SE=0.080) to 0.40 (SE=0.080). Both groups changed their rating from the BASELINE phase, but the real-time group rating in the DISPLAY phase is significantly higher than long-term group in the DISPLAY phase (contrast within the DISPLAY phase is significant  $F[1,9]=10.45$ ,  $p=0.0103$ ). This provides evidence that the real-time display motivates people to re-evaluate themselves, although the magnitude of change is rather small (0.75 on a scale ranging from 1 to 7). There is also a main effect of Phase ( $F[1,9]=56.45$ ,  $p<0.001$ ), which is not surprising because the baseline scores are all normalized to be zero, and thus the variance at the baseline is basically zero, making any change from baseline seemingly significant. Thus, the main effect on phase can be ignored, and the focus is on the interaction between condition and phase, which is not affected by the reduced variance from the normalization scheme.

The results also support H6.3 and show an effect of reflecting on long-term data for the group in the long-term condition. Analysis of the ratings reveals a significant interaction effect between condition and phase ( $F[1,27]=11.79$ ,  $p=0.002$ ) (Figure 6-23). The group in the long-term condition went from a mean of 0.20 (SE=0.16) in the

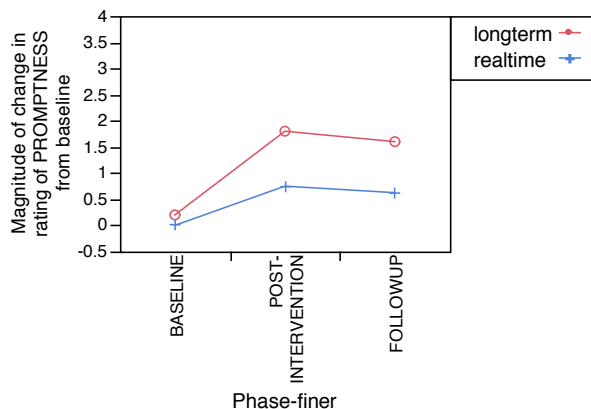


**Figure 6-23. A significant interaction effect for the phase and condition shows that after the reflection session (post-intervention), the long-term group changed their ratings, whereas the real-time group did not change their ratings as they continued with the real-time display.**

BASELINE phase to a mean of 1.4 (SE=0.16) in the POST-INTERVENTION phase, whereas the group in the real-time condition remained relatively stable from a mean of 0.75 (SE=0.21) to a mean of 0.67 (SE=0.14). This interaction effect shows the group in the long-term condition significantly changed their rating of their ability to take their medications promptly after reviewing the long-term data.

The results further support H6.3 in that analysis shows that the differences in magnitudes of the changes that the two groups make in their ratings are statistically different. For this analysis, we redefine the BASELINE phase to be the months before the intervention for each group (month 2 for the real-time group and months 2 and 3 for the long-term group), the POST-INTERVENTION phase to be the month just following the intervention for each group (month 3 for the real-time group, and month 4 for the long-term group), and the FOLLOWUP phase to be the months following (months 4,5, and 6 for the real-time group and months 5,6 for the long-term group). Analysis of these phases will reveal the effects of the real-time and long-term reflection interventions and will allow us to compare the magnitude of changes that each group made as a result of their intervention. There is a marginally significant interaction effect between condition and phase  $F[2,51]=2.45$ ,  $p=0.0964$ , with the long-term group making more changes in their ratings of their ability to take their medications promptly (Figure 6-24). In the POST-INTERVENTION phase, there is a significant contrast ( $F[1,47.6]=6.54$ ,  $p=0.0137$ ) between the deviation from baseline between the long-term (mean difference from baseline of 1.8 points, SE=0.30) and real-time (mean difference from baseline of 0.75, SE=0.28) groups. This difference continues into the FOLLOWUP phase as well

( $F[1,15.2]=12.97$ ,  $p=0.0026$ ). These results show that the long-term group, after reflecting on the long-term group, changed their subjective rating of their ability to take their medications promptly much more than the real-time did with the introduction of the display. The impact of reflecting on the trended long-term data had a greater subjective impact on the individuals in the long-term group than the real-time display had on individuals in the real-time group.



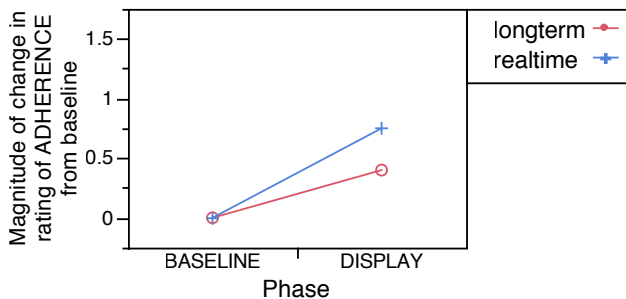
**Figure 6-24. The results show that the long-term group, after reflecting on the long-term group, changed their subjective rating of their ability to take their medications promptly in the POST-INTERVENTION and FOLLOWUP phases much more than the real-time did with the introduction of the display**

Trended long-term reflection can highlight the patterns of function (or dysfunction) that an individual can use to reassess and adjust their ratings of their own abilities. Reflecting on the long-term data seems to have a little more of a “shock” factor, showing individuals trends in their own behaviors, particularly trends that they did not expect. These unexpected trends can induce a feeling of cognitive dissonance that make individuals re-evaluate their abilities in a way that continuous real-time feedback does not.

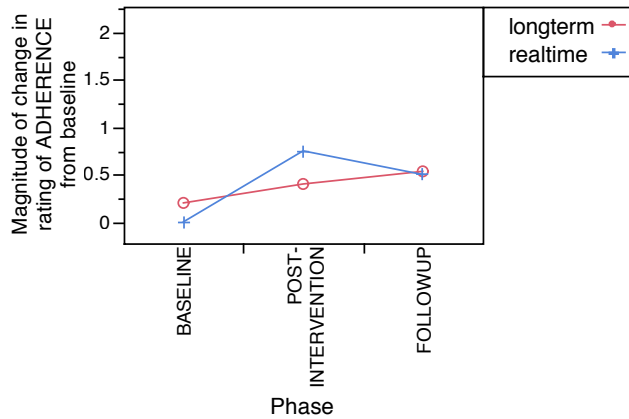
### 6.6.3.2 Ratings of ability to be adherent to medication routine

The real-time feedback did not trigger the real-time group to make any significant changes in their ratings of their ability to take their medications. The results show no significant interaction effect between condition and phase ( $F[1,10]=0.65$ ,  $p=0.438$ ) when comparing the BASELINE and DISPLAY phases (Figure 6-25). Again, the main effect of Phase is an artifact of the normalization routine where the baseline values are normalized to zero and thus can be ignored.

Likewise, there were no significant interaction effects when analyzing changes in ratings across the BASELINE, POST-INTERVENTION, and FOLLOWUP phases ( $F[2,51]=1.48$ ,  $p=0.2362$ ) (Figure 6-26).



**Figure 6-25. There was no significant interaction effect across phase and condition in the change in subject ratings of adherence. Even though the members of the real-time group did change their ratings more than the long-term group, as was predicted by hypothesis H6.3, the magnitude difference was not statistically significant. The real-time display did not have a strong effect in affect individual’s ratings of their medication adherence.**



**Figure 6-26. There was no significant interaction effect across phase and condition for ratings of adherence. Reflecting on the long-term data did not seem to contribute to any significant differences in how members of the long-term group rated their medication adherence.**

Therefore, when considering whether H6.3 can be supported by the data, the results about changes in ratings of medication adherence do not provide additional support above what was already found in the ratings of medication promptness.

### 6.6.3.3 Summary of Support for Hypothesis 6.3

As predicted by hypothesis 6.3, the results show that individuals in the long-term group did change their subjective ratings of how on time they took their medications after the reviewing 6-8 weeks of their medication taking data. Indeed, they changed it to a much larger degree than the changes that the real-time made to their ratings of how on-time they were. This difference in how individuals rated themselves can be explained by the fact that reviewing long-term trended data provides a “dramatic relief” by which the individual can notice how different their behaviors are from their own self-perception. The long-term reflection challenged their existing self-perception of their abilities and provided them with data to re-evaluate themselves. In contrast, the real-time display worked more subtly, helping individuals to improve their performance but doing so one day at a time, rather than showing a trend of dysfunction that contrasted with the individual’s current self-perception.

<b>H6.3</b>	Did the long-term group change their ratings of their abilities more than the real-time group?
PROMPTNESS	Yes
ADHERENCE	No
CONCLUSION	Hypothesis 6.3 is supported in ratings of promptness but not in ratings of adherence. The long-term group changed their ratings of promptness more from baseline than the real-time group because the long-term reflection session provided trend information that challenged their existing self-perception of their abilities.

### 6.6.4 Removing Real-Time Feedback and Behavior Change

The introduction of the real-time feedback display for the real-time group was shown to be effective in supporting individuals in their goal to take their medications more consistently over the long term. Individuals received the immediate feedback helpful for showing progress on the short-term sub-goals of taking their medications well for each day. However, the question arises as to whether the real-time feedback made any permanent changes in the individual’s

abilities or motivation to perform their medication taking tasks better. Without the support of the real-time feedback, individuals will not receive the incremental feedback about how well they are achieving their overall long-term goal of taking all their medications correctly, promptly, and consistently. The long-term goal can be broken down into a series of sub-goals, each of which is taking medications well for the current day. The real-time feedback provides progress reports on how well individuals are achieving their daily sub-goals. Frequent feedback on achieving sub-goals has been shown to support higher goal achievement (Latham, Mitchell, & Dossett, 1978). We consider the contrapositive claim about whether removing feedback, once individuals have grown accustomed to it, would result in a relative decrease in goal attainment (in this case, a reduced level of task performance). We consider the following hypothesis:

H6.4 Individuals will decrease their medication task performance (less adherent, less prompt, less correct, more late, and more variable in the time of day taken) after the real-time feedback display is removed when compared with their performance with the display.

In order to investigate the persistence of the effect of the real-time feedback, the real-time tablet display was removed from the homes of the individuals in the real-time display group after five months the display was initially deployed. Individuals continued to use the dwellSense sensors to log their task performance behaviors. In this section, the effect of removing the support of the real-time feedback is analyzed along the five dimensions of medication taking: adherence, promptness, correctness, lateness, and the variance in the time of day.

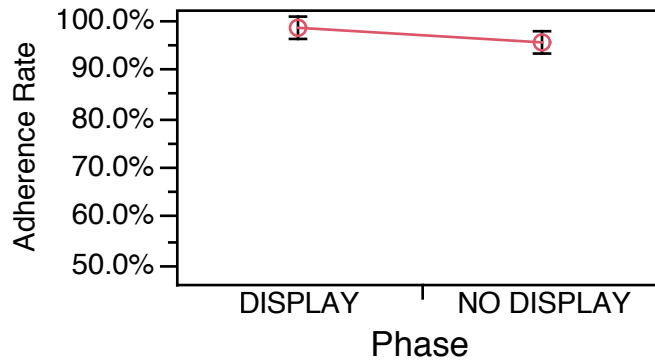
The analysis of behavior change associated with the removal of the real-time feedback display considers the 12 weeks of behaviors surrounding the date when the real-time display was removed from the individual's home. The "DISPLAY" phase consists of the six weeks of behaviors before the display was removed, and the "NO DISPLAY" phase consists of the six weeks after the display was removed. Averages were computer for two-week chunks of the data, resulting in a total of six repeated measures of each feature of the medication-taking task. Two-weeks units of analysis provided the best compromise between the number of repeated measures (because more measures are useful for more statistical power) and the noisiness of the data as the individual varies in performance from one week to another. A repeated measures within-subjects one-way ANOVA was used to identify whether individuals differed in their medication-taking behaviors with and without the support of the real-time feedback display. The presence of a main effect of Phase indicates a change in how the task is performed.

In contrast to the earlier analysis of behavior change, the long-term group is not used as a control group because individuals in the long-term group are undergoing changes as the result of their own informational intervention of the long-term feedback. Comparing the behaviors of groups at this point in the study would highlight the differences in the intervention type rather than highlighting specifically the within-group, within-subject effect of removing the real-time feedback, which is the focus of the current analysis.

The results of the analysis provides support for Hypothesis H6.4 by demonstrating that individuals decreased in the performance of their medication taking after the display was removed from the homes of the individuals. Individuals were significantly less adherent, less prompt, less correct, more late, and more varied in the time of day they took their medications.

#### 6.6.4.1 Medication Adherence without Real-Time Feedback

Hypothesis 6.4 predicts that individuals will decrease their medication adherence without the support of the real-time feedback. The analysis reveals a significant decrease in medication adherence when going from the DISPLAY (mean=98.2%, SE=1.0%) to the NO DISPLAY phase (mean=95.2%, SE=1.0%) ( $F[1,29]=4.85$ ,  $p=0.0355$ ). Even though the difference in adherence rate was rather small (averaging 3% across the six participants in the real-time group), but this difference was nonetheless statistically significant. The statistical significance of this small change in a relatively small sample size ( $n=6$ ) suggests a very clear pattern of change once the display was removed. Participant L02

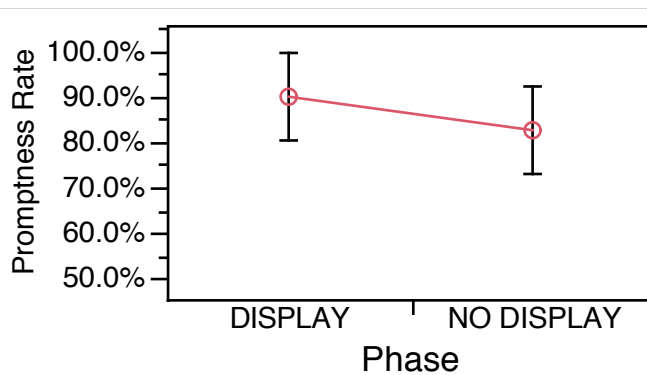


**Figure 6-27. Medication adherence decreased by approximately 3% after the real-time display was removed ( $p=0.0355$ ).**

provides a particularly apparent pattern of 100% adherence (missing zero medications) with the display, but without the display missed taking his medications seven times in the four weeks following the removal of the display. Individuals missed their medications more often after the real-time display was removed.

#### 6.6.4.2 Medication Promptness without Real-Time Feedback

Hypothesis 6.4 predicts that individuals will be less prompt without the support of the real-time feedback. Promptness is a binary measure for each medication-taking episode corresponding to whether the medications were taken before or after the pre-defined late time for each individual. The analysis reveals a significant decrease in medication promptness when going from the DISPLAY (mean=89.9%, SE=3.9%) to the NO DISPLAY phase (mean=82.5%, SE=3.9%) ( $F[1,29]=8.23$ ,  $p=0.0074$ ). Without the support of the real-time feedback individuals took their medications after their pre-defined late time more often.



**Figure 6-28. Medication promptness decreased by 7.5% after the real-time display was removed ( $p=0.0074$ ).**

#### 6.6.4.3 Medication Correctness without Real-Time Feedback

Hypothesis 6.4 predicts that individuals will be less correct in the medication-taking task without the support of the real-time feedback than with the real-time feedback. Correctness corresponds to whether the individual has opened the pillbox door that corresponds to the current day of the week. The analysis reveals a marginally significant decrease in correctness when going from the DISPLAY (mean=99.6%, SE=1.0%) to the NO DISPLAY phase (mean=97.3%, SE=1.0%) ( $F[1,29]=3.53$ ,  $p=0.0704$ ). Without the support of the real-time feedback individuals selected the wrong door on the pillbox more often.

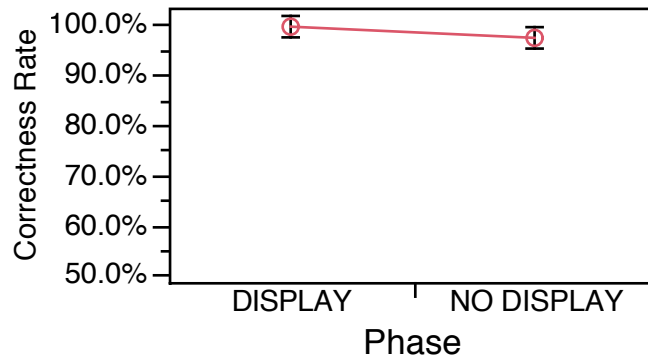


Figure 6-29. Medication correctness decreased by 2% after the real-time display was removed ( $p=0.0704$ ).

#### 6.6.4.4 Medication Lateness without Real-Time Feedback

Hypothesis 6.4 predicts that individuals will deviate from the normal time they take their pills (that is, increase in lateness) to a greater degree without the support of the real-time feedback than with the real-time feedback. In contrast to promptness, which is a binary measure of whether late or not late, the lateness measure quantifies the degree of difference from the normal time medications are normally taken. It is important for individuals to take their medications close to same time everyday so that the levels of medications in their body remain relatively stable. The analysis reveals a significant increase in lateness when going from the DISPLAY (mean=61.7minutes, SE=15.1) to the NO DISPLAY phase (mean=92.90 minutes, SE=15.1) ( $F[1,29]=35.09$ ,  $p<0.0001$ ). Without the support of the real-time feedback, individuals deviated from the time they said they should be taking their medications.

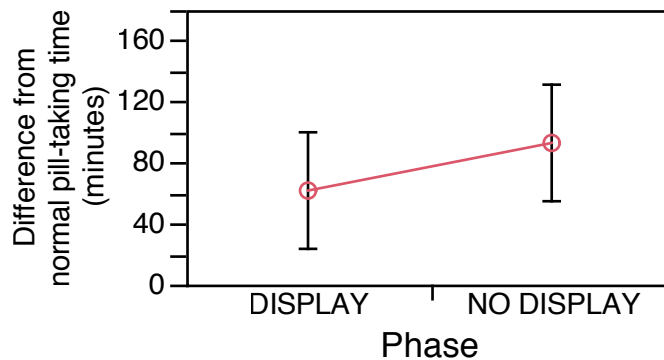
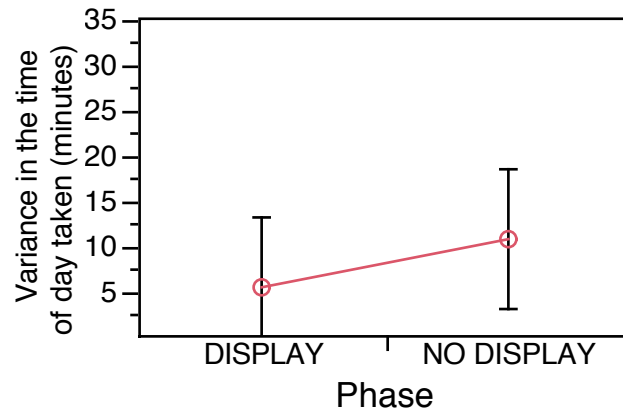


Figure 6-30. Individuals deviated from the time they normally take their medications to a greater degree after the real-time display was removed ( $p<0.0001$ ).

#### 6.6.4.5 Variance in the Time of Medication Taking without Real-Time Feedback

Hypothesis 6.4 predicts that individuals will vary more in the time of day they take their medications without the support of the real-time feedback than with the feedback. Similar to the lateness metric, the variance in the time of day is a measure of the “sloppiness” in the medication-taking routine. Individuals attempt to take their medications at roughly the same time everyday to maintain a steady level of medication in their system. The analysis reveals a significant increase in the variance in the time of day for medication taking when going from the DISPLAY (mean=5.48minutes, SE=3.24) to the NO DISPLAY phase (mean=10.78minutes, SE=3/24) ( $F[1,29]=4.44$ ,  $p=0.0438$ ). The times when the medications are taken are more spread out after the real-time display is removed.



**Figure 6-31. The time of day individuals took their medications was more spread out after the real-time display was removed ( $p=0.0438$ ).**

#### 6.6.4.6 Summary of Behavior Change after Removing Real-Time Feedback

Even though real-time feedback supported a higher level of performance in the medication-taking task, the higher level of performance was not sustained after the real-time feedback was taken away. Individuals in the real-time feedback ground significantly decreased their medication adherence, promptness, and lateness and significantly increased in lateness and in the variance in the time of day they took their medications. There was also a marginally significant decrease in correctness (choosing the correct pillbox door to open) after the real-time feedback was removed. The results provide a fairly strong support for hypothesis H6.4 and show that the real-time display provides the immediate feedback necessary for reinforcing and motivating individuals to perform their medication taking well. Without the frequent, immediate feedback from the display, individuals did not see progress on how well they were achieving the daily sub-goals of taking their medications well and thus were not able to see progress towards the longer-term goal of taking medications well over time. Without the reaffirming feedback of the display, individuals were not as aware of their mistakes and likely assumed they were performing rather well despite the mistakes they made after the display was taken away.

The magnitude of the changes in these measures were small yet statistically significant. It is also notable that individuals exhibited a very high level of performance with the display, in fact, the levels were very close to perfect (100%) adherence, promptness, and correctness. The decreases in performance from these very high levels to a level slightly less than perfect suggest that individuals went from making zero (or very nearly zero) number of mistakes to a significantly higher positive number of mistakes after the real-time display was removed. Even though statistically significant, it not clear (nor is it the focus of this particular analysis) whether the changes are significant from a personal or clinical perspective. Perhaps it is not important whether individuals miss zero pills per month, two pills per month, or seven pills per month. Based on qualitative feedback from the individuals themselves, it was their goal to never miss



taking their medications and thus skipping more than one or two was a concern, and thus they valued the support of the real-time display for helping them sustain a near-perfect level performance.

## **6.7 Discussion**

We examined the impact of reflecting on data in real-time and reflecting on long-term data on people's behaviors, accuracy of self-awareness, and their subjective ratings of their abilities.

### **6.7.1 Benefits of Real-time Feedback**

The display that provided real-time feedback helped individuals to be more adherent, more prompt, and less late in their medication taking routine, as well as reducing the variance in the time of day the medications were taken. The improvements in these behaviors occurred not only immediately after the introduction of real-time feedback display, but the improvements in adherence, promptness, lateness, and variance in the time of day persisted at least 3 months after the display was introduced. The real-time feedback provided individuals with the constant feedback loop that sustained the improvements in behavior over time. Individuals could check to see if they were meeting their goals on a daily basis and adjust their behaviors from one day to the next to meet their goals. These results provide a strong confirmation for hypothesis H6.1a.

Even though real-time feedback was effective at helping people improve their behaviors, the real-time feedback had little effect on making them more self-aware of the mistakes they make in their tasks. For example, they still under- or overestimated the number missed pills by the same amount before and after the introduction of the display. Thus, the real-time feedback enables behavior change without necessarily improving self-awareness of their abilities, or at least the type of retrospective self-awareness that we asked about in our questionnaires. The real-time dashboard style display encourages individuals to reflect in the moment and to use the immediate information to reinforce proper behaviors. However, it does not necessarily encourage individuals to remember the data values for future recall or comparisons. Thus, increases in the "in-the-moment" real-time awareness provided by the real-time display may not be directly measurable by asking them, as we did in our questionnaires, to recall the number of missed pills or incorrectly opened pillbox door in the past week. Based on these results, we did not find direct evidence for hypothesis H6.2a, which suggests that even though real-time feedback increases task performance, it does not influence the accuracy of the individual's self-awareness.

The real-time feedback also had a small effect on people's rating of their abilities to carry out tasks important for independence, such as taking medications promptly. In the month following the introduction of the real-time display, individuals in the real-time group changed their ratings more than the long-term group, which was acting as a control for the real-time group during this time period. Relative to the subsequent reassessment of their own abilities in the long-term group, the change in rating of medication promptness was rather subtle.

In summary real-time feedback was effective at behavior change to improve task performance but it achieved this rather subtly, without impacting the individual's self-awareness or the individual's subjective rating of their abilities.

### **6.7.2 Benefits of Long-Term Reflection**

Reflecting on 8 weeks of trended data helped individuals in the long-term group to be more prompt in their medication taking but had only very subtle effects on adherence, correctness, lateness, and the variance of time taken. Although, none of these latter task performance attributes were statistically significantly better than baseline, they all were higher directly after the reflection session than during baseline and follow up phases. A "v-shaped" curve was found for the long-term condition in the graphs where the immediate phase was usually higher or lower than both the baseline and follow-up phases. This repeated pattern across different attributes suggests that reflecting on long-term data may be

result in subtle (yet temporary) improvements in behavior that are not detected by statistics alone. This pattern as well as the statistically significant improvement in medication promptness provides some support for hypothesis H6.1b.

The long-term data view, with each data point contextualized within its neighbors, can highlight trends and make individuals aware of their own behaviors over an extended period of time. The data is also shown all at once, so that full effect of the data could be observed after the reflection session. The long-term feedback was effective at helping people be very accurate (to be almost without error) in their self-awareness of how often they took pills late but this accuracy did not persist beyond a month, as the accuracy started to decrease over time without the support of continued reflection or real-time feedback. These results support hypothesis H6.2a, suggesting that reflecting on long-term data is effective at supporting a greater self-awareness of behavior but the improvement in self-awareness is temporary.

Individuals in the long-term group also changed their subjective ratings of their abilities to take their medication promptly after reflecting on the long-term data. These results provide support for hypothesis H6.3. Moreover, they changed their ratings to a much greater degree than did the real-time group and they continued to adjust their ratings of their abilities months after the reflection session, whereas the real-time group made smaller changes to their ratings and did not change them much over time. This longer lasting pattern in the long-term group can be attributed to the “shock” factor that they experienced when reflecting on their data. That initial re-adjustment of their self-rating caused them to re-evaluate themselves frequently in the future. This suggests that the long-term reflection was more convincing and useful for people to re-assess their abilities and rate them differently than before, but not as effective as real-time reflection for supporting behavior change.

## 6.8 Summary

In summary real-time feedback was effective at behavior change to improve task performance but it achieved this rather subtly, without impacting the individual’s self-awareness or the individual’s subjective rating of their abilities. Individuals were likely to have been encouraged to perform better because the immediate feedback provided them with the progress reports on their sub-goals to take their medications every day. Long-term reflection, on the other hand, facilitated only limited amount of behavior change but instead the long-term data made the individuals more self-aware of their habits and motivated them to re-evaluate themselves much more often. The less frequent feedback did not give them frequent indications of progress on their sub-goals to take their meds everyday but did provide them with an opportunity to see how far their perceived behaviors differed from their actual behaviors and thus become more accurate in their self-awareness.

## 6.9 References

- Baer, J. S., Marlatt, G. A., Kivlahan, D. R., Fromme, K., Larimer, M. E., & Williams, E. (1992). An experimental test of three methods of alcohol risk reduction with young adults. *Journal of consulting and clinical psychology*, 60(6), 974.
- Bandura, A., & Simon, K. M. (1977). The role of proximal intentions in self-regulation of refractory behavior. *Cognitive Therapy and Research*, 1(3), 177–193.
- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of personality and social psychology*, 41(3), 586.
- Brug, J., Campbell, M., & van Assema, P. (1999). The application and impact of computer-generated personalized nutrition education: a review of the literature. *Patient education and counseling*, 36(2), 145–156.
- Carter, S., & Mankoff, J. (2005). When participants do the capturing: the role of media in diary studies. Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '05 (pp. 899–908). New York, NY, USA: ACM. doi:10.1145/1054972.1055098
- Choo, P. W., Rand, C. S., Inui, T. S., Lee, M. L. T., Cain, E., Cordeiro-Breault, M., Canning, C., et al. (1999). Validation of patient reports, automated pharmacy records, and pill counts with electronic monitoring of adherence to antihypertensive therapy. *Medical care*, 37(9), 846.

- Cochran, W., & Tesser, A. (1996). The “what the hell” effect: Some effects of goal proximity and goal framing on performance. *Striving and feeling: Interactions among goals, affect, and self-regulation*, 99–120.
- Conway, M., & Bekerian, D. (1987). Organization in autobiographical memory. *Memory & Cognition*, 15(2), 119–132. doi:10.3758/BF03197023
- Fischer, M., Stedman, M., Lii, J., Vogeli, C., Shrank, W., Brookhart, M., & Weissman, J. (2010). Primary Medication Non-Adherence: Analysis of 195,930 Electronic Prescriptions. *Journal of General Internal Medicine*, 25(4), 284–290. doi:10.1007/s11606-010-1253-9
- Hailpern, J., Jitkoff, N., Warr, A., Karahalios, K., Sesek, R., & Shkrob, N. (2011). YouPivot: improving recall with contextual search. *Proceedings of the 2011 annual conference on Human factors in computing systems, CHI '11* (pp. 1521–1530). New York, NY, USA: ACM. doi:10.1145/1978942.1979165
- Haynes, R. B., Ackloo, E., Sahota, N., McDonald, H. P., & Yao, X. (2008). Interventions for enhancing medication adherence. *Cochrane database syst Rev*, 2(2). Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000011.pub3/pdf/standard>
- Kalnikaite, V., Sellen, A., Whittaker, S., & Kirk, D. (2010). Now let me see where i was: understanding how lifelogs mediate memory. *Proceedings of the 28th international conference on Human factors in computing systems, CHI '10* (pp. 2045–2054). New York, NY, USA: ACM. doi:10.1145/1753326.1753638
- Latham, G. P., Mitchell, T. R., & Dossett, D. L. (1978). Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology; Journal of Applied Psychology*, 63(2), 163.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Prentice-Hall, Inc. Retrieved from <http://psycnet.apa.org/psycinfo/1990-97846-000>
- Millner, W. R., & Rollnick, S. (2002). *Motivational interviewing*. The Guilford Press, New York, London,.
- Olivieri, N. F., Matsui, D., Hermann, C., & Koren, G. (1991). Compliance Assessed by the Medication Event Monitoring System. *Archives of Disease in Childhood*, 66(12), 1399–1402. doi:10.1136/adc.66.12.1399
- Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of consulting and clinical psychology*, 51(3), 390.
- Prochaska, J. O., DiClemente, C. C., Velicer, W. F., & Rossi, J. S. (1993). Standardized, individualized, interactive, and personalized self-help programs for smoking cessation. *Health Psychology*, 12(5), 399.
- Sokol, M. C., McGuigan, K. A., Verbrugge, R. R., & Epstein, R. S. (2005). Impact of medication adherence on hospitalization risk and healthcare cost. *Medical care*, 43(6), 521.
- Tan, D. S., Pausch, R., Stefanucci, J. K., & Proffitt, D. R. (2002). Kinesthetic cues aid spatial memory. *CHI '02 extended abstracts on Human factors in computing systems, CHI EA '02* (pp. 806–807). New York, NY, USA: ACM. doi:10.1145/506443.506607

# 7

## Automatic Assessment with Sensors

Assessment of functional abilities is a standard part of clinical care. Assessments are carried out by having individuals self-report their abilities, asking informants such as family members to report on an individual's abilities, or having professionally trained clinicians to directly assess the individual's abilities. In the previous chapters, we have investigated how well individuals were able to self-report their behaviors and found that their self-reports could not match the accuracy or precision of the sensor data. By reflecting on the objectively collected sensor data, individuals were able to readjust their self-awareness of their behaviors and abilities and use the data to support their ability to carry out tasks important for independence.

Having demonstrated the usefulness of embedded assessment data on individuals themselves, now we investigate how embedded assessment can be useful for clinicians. We turn to the question of whether sensing embedded in the home can collect the relevant data and *automatically* assess how well individuals are carrying out their everyday tasks. One of the main qualities of embedded assessment data that makes it useful for supporting an individual's self-awareness is its objectivity. Performance testing, where a trained clinician observes and rates how an individual carries out a structured task, is currently the primary way of collecting objective data about an individual's functional abilities. Prior literature has compared performance testing with measures based on self-report of functional abilities. While in some cases, "subjective" self-reports and "objective" performance testing align fairly well (Myers 1992), there are many more cases, particularly in those cases when individuals are affected by cognitive or psychological conditions, where there is only a modest correlation between self-reports and performance testing (Reuben 1995; Hoeymans 1996). Performance testing often requires specialized equipment and training to be administered which makes it both time-intensive and costly (Sager et al., 1992). Embedded sensing technology in the home that can monitor and assess the performance of everyday activities has the potential to reduce the burden and costs of live performance testing while increasing its precision and validity.

This chapter investigates how embedded assessment compares with performance testing along dimensions beyond simply objectivity such as accuracy, precision, and representativeness. This chapter addresses research question RQ5 and describes how automated assessment based on only sensor data compares with assessments by a human clinician.

### 7.1 Performance Testing

Performance testing is used most commonly to provide objective clinical data about an individual's functional abilities when planning the discharge of hospital patients, screening for conditions like cognitive impairment, tracking of treatment efficacies such as rehabilitation following a stroke, and detailed assessments of wellness. In performance testing, a trained clinician assesses how an individual carries out a structured task common in everyday life that have well-structured steps that can be assessed relatively independent of each other. Many tools for performance testing have been developed to allow clinicians to observe and assess how an individual carries out a structured task with objective, repeatable, standardized metrics (Loewenstein & Mogosky, 1999). Performance testing was developed and is recommended to complement self-reports of ability that have been found to be inaccurate or biased due to low perceptions of physical competence or depressive symptoms (Wadley et al., 2003, Kempen et al., 1996).

Common tasks in performance testing include telephone use, handling money, meal preparation, dressing, medication management, eating, telling time, taking transportation, and grooming. Some instruments are targeted for individuals with particular conditions such as the Direct Assessment of Functional Status (Loewenstein et al., 1989) and the

Cognitive Performance Test (Burns et al., 1994) for dementia. Other instruments have been developed to test the tasks and skills more commonly performed by healthy older adults including the Everyday Problems Test (Willis 1996) and Observed Tasks of Daily Living (Diehl, Willis, & Schaie, 1995). We selected the Performance Assessment of Self-care Skills (PASS) (Holm & Rogers, 1999) as our benchmark performance-testing tool because 1) its assessment is based on decomposing the overall task into atomic steps and 2) it provided templates for authoring new tasks (such as medication taking) not already included in the standard battery. It should also be noted that the PASS was developed at the University of Pittsburgh, a collaborator on this research.

The PASS was developed as a performance-based, criterion-referenced tool for occupational therapists to assess how well individuals carry out activities important for independence. The PASS consists of 26 tasks within four functional categories: functional mobility, personal self-care, instrumental activities of daily living (IADLs) with a cognitive emphasis, and IADLs with a physical emphasis. Some tasks require standard props (such as pill bottles and beads for the medication management/sorting task), particularly when the PASS is administered in the clinic. The PASS can also be administered in the individual's home using the objects in their home as props. The therapist first explains the instructions of the test to the individual and asks if the individual understands the instructions. If the individual responds yes, then the therapist observes how the individual carries out the task and notes any safety issues or errors. In the study discussed in this chapter, the occupational therapist administered three PASS tasks: medication taking, phone use, and coffee making.

The medication-taking task (Appendix E) required individuals to demonstrate to the therapist how they used their pillbox to take their medications and how they followed through with consuming their pills. Some common errors include selecting the wrong pillbox (PM instead of AM) or opening a door on the pillbox that does not match the current day of the week. The phone use task was an updated version of the standard phone use task found in the standard PASS battery. The updated phone use task (Appendix F) tested whether individuals could dial a pharmacy, navigate the store's phone menu to hear the opening and closing times, and then report this to the therapist. Some common errors include misdialing the pharmacy number (even though it was written on a piece of paper) or selecting incorrect phone menu choices. The coffee-making task (Appendix G) assessed how well an individual performed the different steps in making a pot of coffee using the coffeemaker. Some of the steps assessed include measuring out the right amount of water, filling the reservoir, scooping coffee grounds into the filter, and turning on the machine. Individuals generally did not have problems performing this task during the PASS assessment.

The PASS is a criterion-based tool, not a normative tool. Individuals are assessed on their ability to carry out a task and whether they are able to meet pre-defined criteria while completing the task, rather than assessed relative to the normative ability of a particular population. PASS tasks are decomposed into atomic steps that can each be rated on independence, safety, and adequacy. Independence refers to how much assistance the individual needs to complete the step, with more assistance equating to a lower independence score. The PASS structures the types of assistance the therapist can provide starting with 1) verbal supportive encouragement, to 2) verbal non-directive, to 3) verbal directive, and all the way to 9) total assist (where the therapist completely performs that step for the individual). Each step is assigned an independence rating from 0 (total dependence) to 3 (completely independent) based on the number and level of assistance provided during that task step. Aggregating the independence scores at each step for a task-level independence score is done by averaging the independence scores for each step. Safety refers to whether the individual is taking risks to their personal safety while completing the step. Each step is assigned a safety score from 0 (step stopped by therapist to prevent personal injury) to 3 (completely safe practices). The aggregate safety score across all the steps is computed as the minimum of the safety scores for each step. Adequacy refers to essentially how well the task is performed. Adequacy has two components: the quality of the outcome of the task and the quality of the process of reaching that outcome. For example, in a phone-use task, when individuals are navigating a pharmacy phone menu, the quality of the outcome might be good (reached the menu that tells them the pharmacy opening and closing hours), but the quality of the process might be poor such as having difficulty remembering the menu options, pressing multiple

buttons at once, or having to repeat the options again and again. The overall adequacy score for a task is quantified subjectively by the therapist who considers the frequency and severity of the process and outcome issues encountered by the individual while performing the task. The therapist will consider which is lower, process or outcome, and assign an overall rating based on the lower of the two. The overall adequacy score ranges from a 0 (outcome standards not met, process so poor that it prevents the completion of the task) to a 3 (task performed relatively efficiently and with outcome standards met). By identifying the steps in which the individual encounters breakdowns in independence, safety, or task adequacy, the therapist can use the PASS data to know exactly which steps require intervention and also know what type and amount of assistance is necessary.

Performance testing, while providing objective data, requires a trained clinician to administer and can be burdensome to assessments can be rather lengthy (Moore et al., 2007). The field of performance testing is continually looking for ways to reduce the administration time (and thus cost) while maintaining the fidelity and objectivity of an assessment. In developing new tools for performance testing and also evaluating existing tools (like the PASS which, for the sake of clarity, is a test that measures performance, not a software tool), establishing predictive validity is an important criterion for a meaningfully useful tool. Predictive validity requires that the scores and measures resulting from a performance-based tool actually reflect the abilities or habits of the individual as they are in their everyday lives. Currently tools for performance testing are compared with each other to establish a certain level of agreement among the tools. Ideally, studies of tools would also involve extended observations and assessments of the individual's behavior in order to establish that the results of performance testing are actually directly related to how individuals perform tasks in their lives. The data collected by embedded sensors in the home that monitor everyday activities can be used as a helpful source of data to use in comparisons to establish predictive validity.

## 7.2 Research Questions

Performance testing by a trained clinician is analogous to how an embedded assessment system monitors and assesses the performance of an activity. In performance testing conducted with a clinician, each step of a structured task is first observed and then assessed in terms of its outcome and process according to pre-established criteria. Similarly, sensors can be embedded in specific places in home environments that correspond to critical steps of a task. A system using these sensors can first observe and recognize whether a step was started or completed and then assess how well that step was performed based on metrics in the sensor data (for instance, recovered errors or task completion time).

Even though the human clinician and automated system are both attempting to perform analogous tasks, they are limited by the inputs they have available. A human clinician can have a broad view of the task and can notice details that might not typically fit in the schema of completing the task, but the human clinician is limited in the frequency of her observations to only the duration of her visit. In contrast, a sensor-based system has a comparatively narrower view of the task, focusing on particular steps and leaving out the steps that may be too difficult to sense, but in the steps that the system does monitor, it can record details with a high level of precision (such as task completion time). Furthermore, automated sensor-based assessment is not limited to observations from a single visit but can be used continuously for an extended period of time, particularly if it is designed to monitor passively with a minimal amount of intrusion into the lives of the monitored individual.

To understand the relative benefits of automated sensor-based assessment versus traditional performance testing (conducted manually by a trained clinician), we consider the following research questions:

- RQ6.1 Can automated sensor-based assessment rate people's abilities as accurately as traditional performance testing?
- RQ6.2 Are behaviors recorded during traditional performance testing different than the actions taken during the individual's everyday performance? Can the reactivity of performance testing be quantified?

RQ7 What aspects of task performance is sensor-based assessment better suited for? What aspects of task performance is performance testing better suited for?

To address these research questions, both automated sensor-based assessments and traditional performance testing were applied to twelve older adults in a longitudinal study with sensors in their homes.

### 7.3 Data Collection

During the same 11-month sensor deployment described in Chapter 6, a trained occupational therapist administered performance testing (using the Performance Assessment of Self-Care Skills tool, the PASS) up to a total of three times for each of twelve older adults participating in the study. Each of these older adults received a suite of sensors appropriate for their routines to monitor their medication taking, phone use, and coffee making. Individuals were told at the beginning of the study that a trained occupational therapist would schedule a time to observe and test how they carried out common activities at home. In each of three visits, the occupational therapist would assess at least three different PASS tasks (medication taking, phone use, and coffee making). For those participants who made coffee with a coffeemaker, the occupational therapist would also administer the coffee-making PASS task and observe them make coffee.

The occupational therapist rated how each individual carried out their tasks in person and only provided assistance when necessary as outlined by the rules of PASS. The occupational therapist, contrary to normal practice, only assessed how the participant performed and refrained from fixing any problems they observed by suggesting modifications to participant's task routine. The occupational therapist was also careful not to provide any feedback to participants about how well they performed their task unless it was necessary for them to complete the task correctly. Thus, the visits from the occupational therapist were not intended (nor were they in practice) to be interventional visits, that is, with the goal to alter and improve the way the participants carry out their tasks. The occupational therapist in this study is acting as an objective and non-interventional rater, much in the same way as the sensors are passively monitoring, without providing explicit feedback (except for the individuals who received the real-time feedback tablet display).

The occupational therapist visited each participant and administered the PASS three times, except for participant L14 who withdrew from the study before the second round of PASS tests. In each round, the therapist administered the same three tasks: medication taking, phone use, and coffee making. The first round of the PASS was administered in the November 2011, which corresponded to either the 4<sup>th</sup>, 5<sup>th</sup>, 6<sup>th</sup>, or 7<sup>th</sup> month of the deployment, depending on when the participant was first enrolled in the study. The second round of PASS was administered in two months after the first, in January 2012, and the third and final round of PASS testing was administered in late February and early March 2012.

The PASS can be administered in the clinic with props or in the home with the individual's own objects as props. When administering the PASS at home, the ideal situation is to observe how individuals carry out their daily tasks as they actually perform their tasks as part of their routine. For example, to assess medication taking, the occupational therapist would ideally visit early in the morning when the individual is actually taking their morning medications. However, scheduling visits can be difficult (both in this study and in clinical practice) because therapists must visit many people in a single day and thus cannot always schedule visits to coincide with the individual's routines. In many cases during the visits in this study, the occupational therapist visited the individual after the time the individual would take her morning medications. In these situations, the therapist would ask the individual to simulate taking her medications usually with a prompt such as, "Imagine that you just realized you forgot to take your medications this morning, can you show me what you would do to take your morning medications now?" The individual would walk the therapist through by demonstrating what they would do including opening the pillbox to getting a glass of water to swallowing the pills. Similarly, the coffee-making task was often simulated because the time of the therapist's visit did not coincide with the normal time the individual would be making coffee (for example, right when they get up in the

morning). Nonetheless, the individual would demonstrate all the steps of making coffee from filling the machine with water, putting a new filter in, scooping in the coffee grounds, *etc.* In some cases, individuals refused to perform the coffee-making task because they just made a fresh pot prior to the therapist's visit or just did not want to make coffee at that time of day. Thus in practice, even though performance testing can be conducted in the home environment, it is limited in the frequency of assessment and in many cases also be limited to observing simulated behaviors rather than the actual behaviors as the individuals would naturally carry them out.

During these assessments, in addition to the occupational therapist, the sensors deployed in the participant's apartment also monitored the participant's task performance. The occupational therapist noted the start time for each task so that the corresponding sensor episode could be identified for comparison. The following sections describe how the sensor data was used for automatically generating ratings of task performance and how these ratings compare with the ratings from the PASS.

## 7.4 Automated Performance Testing using Sensors

Sensors embedded in the home environment can passively monitor how individuals actually carry out their tasks, particularly if they are designed to be unobtrusive and are deployed over an extended period to allow the novelty effect of monitoring to extinguish. Most prior research employing sensors in the home focus on recognizing what activities an individual is performing rather than assessing the quality of any particular activity. However, research conducted in relatively controlled laboratory settings have looked at assessing the quality of specific tasks such as hand-washing (Mihailidis et al., 2003), coffee making (Hodges et al., 2010), and meal preparation (Cook & Schmitter-Edgecombe, 2005). These approaches use rule-based criteria, such as requiring certain steps and the correct sequencing of steps, to rate the quality of tasks. For a more detailed discussion of these task-based assessments, please see Chapter 2. Another approach is to semi-automatically learn these rules using machine learning and applying these rules to sensor data sequences to determine the errors in task execution. For example, by representing sensor states over time using a Hidden Markov Model, sequences of recognized actions can be recognized as a task. Furthermore, there are efficient algorithms that can calculate the edit distance (the degree of error) from an incorrect sequence to a correct sequence, assuming that a human expert has trained the system and has labeled a subset of these sequences of recognized actions with a pass or fail rating (Wilson & Philipose, 2005). In order to align best with the approach used in non-automated performance testing and also removing the requirement for a clinician to take the extra step to train the system, the automatic error-detection approach used in this thesis uses pre-defined rules based on the Performance Assessment of Self-Care Skills (PASS) tool.

### 7.4.1 Rule-based Assessment with Sensor Data

Using the sensor data we had available for each task, rules for automatically rating the quality of how a task was performed (the task adequacy) were derived from the description of steps in description for each PASS task. The PASS task description included more steps that could be monitored using sensors, and thus naturally the rules for the sensor data were based only on the steps that could be monitored by the sensors. The system focuses on rating task adequacy instead of task safety or task independence because declines in task adequacy usually precede declines in safety or independence (Holm & Rogers, 1999) and thus can be earlier indicators for changes in abilities. The rules implemented by in the system and the results of the subsequent analysis provide a proof of concept for a more generalizable system that can take in a set of rules and a set of sensor events and generate a rating of how well the task was performed.

#### 7.4.1.1 Rules for Assessing Phone Use

The phone use task (Appendix F) is a prompted task where the therapist asks the individual to call a drug store and find out what time its pharmacy closes. The therapist first provides a sheet of paper with a list of pharmacies and phone



numbers, and the individual has to perform the rest of the task without assistance. The PASS breaks down the phone use task into seven distinct steps for evaluation: 1) reading handout and selecting a pharmacy to dial, 2) locating the telephone, 3) dialing the pharmacy number correctly, 4) holding the phone receiver correctly, 5) listening and navigating the pharmacy's phone menu by pressing the right sequence of numbers, 6) ending the call, and 7) reporting the pharmacy closing hour to the therapist. The pharmacies listed on the handout were all local Rite-Aid pharmacies with local telephone numbers and different closing times. Each pharmacy had a similar phone menu system with nearly identical options but announced in different orders. The phone menu consisted of six options, one of which, the number 6, was to hear the store hours. Other notable options were the number 3 (to speak with someone in pharmacy) and the number 9 (to repeat the menu options). The optimal path through the phone menu was to press the number 6 to hear the phone hours and then hang up. After the individual pressed 6 from the menu, the opening and closing times of the store were first read out and then the opening and closing times of the pharmacy were read out. Individuals had to retain the initial instructions in their heads to make sure they reported the correct time (closing time of the pharmacy) to the therapist. If the individual did not report the correct time, the therapist repeated the instructions and asked the individual to try again.

The phone sensor installed in the individual's home can decode the DTMF signal generated from pressing buttons on the phone and can directly detect the two most complex steps in the phone use as outlined in the PASS, step 3 (dialing the number correctly to reach the drug store) and step 5 (navigating the phone menu choices). The sensor data can also indirectly detect step 7 (reporting the correct pharmacy closing time to the therapist) by looking at the number of times the individual had to redial to listen to the information again. However, the system cannot the case when the individual gives up and refuses to dial the pharmacy again. An adequacy rating is automatically generated based on rules for each step ranging from 0 (inadequate performance) to 3 (perfect performance). The overall adequacy rating for the task is calculated by taking the minimum of the ratings for each step. For example, if the adequacy ratings in step 3 and step 5 are 2 and 3, respectively, the overall adequacy rating for the phone use task is calculated as  $\min(2,3)=2$ .

To detect whether the individual is dialing the drug store number correctly (step 3), the DTMF tones are decoded into numbers and the sequence of numbers is compared with the numbers in the list of drug stores. The highest rating of 3 is assigned if the number is dialed correctly on the first attempt with no missing, additional, or misplaced digits. A rating of 2 is assigned if the first attempt was incorrect and the individual had to hang up and make another (successful) attempt. A rating of 1 is assigned if the individual needs two or more redials to reach the drug store. A rating of 0 is assigned if the individual could not dial the correct number and reach the drug store before giving up.

To detect whether the individual is navigating the phone menu correctly to reach the selection that announces the store and pharmacy hours, the numbers dialed after drug store phone number are analyzed. In all cases, the menu option number 6 was the correct option to select to hear the pharmacy hours. The highest rating of 3 is assigned if the individual pressed the number 6 either immediately after hearing the option or at the end of the menu choices. If the individual needed to have the menu options or pharmacy hours repeated one extra time either by pressing '9' or '\*' or having to redial (correctly) to hear the hours a second time, a rating of 2 would be assigned. A rating of 2 would also be assigned if the navigation path through the phone menu was inefficient, for example, choosing '1' (refilling a prescription), then backing up by pressing '9', and then selecting the correct choice of '6' (to hear store hours). A rating of 2 was also assigned if the individual chose to speak with a live person to ask for the store hours by either selecting menu option '3', '5', or waiting on the line until an employee answered the phone. A rating of 1 would be assigned if the individual needed to make more than a total of two attempts at navigating the phone menu, either by hanging up and starting anew or by pressing the menu option to start over '9' or '\*' more than two times. The lowest rating of 0 would be assigned if the individual could not navigate far enough into the menu to hear the store hours.

To detect (indirectly) whether the individual was able to retain the task instructions during the final steps of the task and report the correct closing time of the pharmacy to the therapist, the number of calls that have successfully reached

the point of hearing the announcement of the pharmacy hours can be counted. Recall that if the individual successfully navigated the phone menu to hear the store and pharmacy hours, but if they report the wrong time (for example, the store's closing hour rather than the pharmacy's closing hour) then the therapists repeats the task instructions and asks them for the pharmacy's closing hour. If the individual is unable to recall that from the announcement, they are asked to try again by initiating another attempt at calling the pharmacy. The system can identify the first and repeated attempts by matching the pattern of a correctly dialed number followed by a sequence of phone menu selections that would lead to the announcement of the pharmacy hours. The highest rating of 3 would be assigned in the case that the individual used only one attempt to find out pharmacy closing hour. A rating of 2 would be assigned if the individual had to make a second call but remembered the pharmacy closing time before hearing the pharmacy hours announcement for the second time, essentially a spontaneous recall of the information from their subconscious memory, probably triggered by the act of dialing and navigating the phone menu. A rating of 1 would be assigned if the individual needed to hear the announcement a second time before reporting the correct pharmacy closing time. Normally a rating of 0 would be assigned if the individual wanted to quit the task and decide not to want to dial any more. However, it is difficult to disambiguate between the situations in which the individual decides to quit or just remembers the information before reaching the announcement a second time. However, it is clear that if the individual never reaches the hours announcement at all, then a rating of 0 would be assigned.

In summary, the rules for assigning a rating to the phone use task focus on three of the seven steps outlined in the PASS for phone use. The minimum rating across the three steps of dialing the correct number to the drug store, navigating the phone menu, and reporting the correct pharmacy hours to the therapist (as measured by the number of attempts needed to retrieve this information) provides the overall adequacy rating for the phone use task.

#### **7.4.1.2 Rules for Assessing Medication Taking**

In the medication-taking task (Appendix E), the individual retrieves their pills from the pillbox and swallows them. In contrast to the phone use task, the medication-taking PASS task was designed to be more observational, allowing the individual to carry out the task in her own way rather than following a set of instructions. The PASS decomposed the medication-taking task into 8 steps for evaluation: 1) locating pillbox, 2) opening the correct pillbox door, 3) removing pills from the pillbox, 4) handling pills between pillbox and consumption, 5) getting a beverage, 6) swallowing pills with beverage, 7) resetting the pillbox doors (so that it reminds them that they have taken their pills), and 8) putting away the pillbox in a place consistent with their routine.

The sensor-augmented pillbox can monitor three of the steps in the medication-taking task: step 2 (opening the correct pillbox door), step 3 (removing the pills from the pillbox), and step 7 (resetting the pillbox doors). Each of these steps was assigned an adequacy rating from 0 to 3. The overall adequacy rating for the task is calculated by taking the minimum of the ratings for each step. The dwellSense system also has the potential to monitor a proxy for step 5 (getting a beverage) using the motion detector in the kitchen to detect whether the individual has entered the kitchen to get a beverage. However, data from the motion sensors were excluded because the extra presence of the therapist in the apartment might have generated erroneous motion signals.

To detect whether the individual has opened the correct pillbox door, data from the augmented pillbox is analyzed to determine what open and close actions the individual has performed on the pillbox doors. The individual has opened the correct pillbox door when she has opened the door for the current day of the week. For example, if it Tuesday, then the individual should have opened the Tuesday door on the pillbox. The highest rating of 3 is assigned if the individual selects the correct pillbox door on the first attempt. A rating of 2 is assigned if the individual selects the correct pillbox door using more than one attempt. A rating of 1 is assigned if the individual opens one or more pillbox doors but none of them are correct. A rating of 0 is assigned if the individual could not open any of the pillbox doors.

To detect whether the individual has continued with the task and removed the pills from the pillbox, data from the accelerometer in the augmented pillbox is analyzed for instances of inverting the pillbox to pour out the pills. Tipping the pillbox on its side or upside down is a gesture commonly performed by the older adults in our study to pour out their pills, particularly if the individual has many pills in the same slot. However, two of our study participants L03 and L10 do not regularly invert the pillbox to retrieve their pills because they each keep only one large pill in each slot and they have the dexterity to reach in and grab that one large pill. Inspection of the data logs of another participant, L11, reveals that she occasionally inverts the pillbox and occasionally does not invert the pillbox. For these three participants, the automated assessment excludes the requirement of inverting the pillbox because they are not expected under normal circumstances to invert the pillbox. For these individuals, they automatically receive a rating of 3 for this step. For the remaining nine participants, their task performance is assessed based on whether they invert the pillbox or not. A rating of 3 is assigned if the individual inverts the pillbox (with any door open) and a rating of 2 is assigned if the individual does not invert the pillbox. A rating of 0 or 1 would be assigned by a therapist during the PASS if the individual had substantial difficulty taking out the pills from the pillbox or could not retrieve the pills at all, but the automated system cannot detect this substantial level of difficulty so it assigns a 2 as its lowest rating as a conservative estimate.

To detect whether the individual has concluded the medication-taking task correctly by leaving the pillbox doors in a state that reminds them that they have taken their pills for that time of the day, the data about which doors are opened or closed is analyzed for an acceptable pattern for each individual. Different individuals used their pillboxes differently. Most individuals closed the pillbox door after taking their pills. For this majority of individuals, a rating of 3 was assigned if they closed all the pillbox doors. A rating of 2 was assigned if they did not close the door they just opened. A rating of 1 was assigned if they left more than one door open. A rating of 0 was assigned if they left open any future door (except tomorrow's) that had pills in it (which is a hazard and improper cue). However, two study participants, L02 and L01, had slightly different habits. L02 opened up the pillbox door for the current day and left it open as a cue for tomorrow to open up the adjacent pillbox door. At the end of the week, all the pillbox doors would be open and he would close them after refilling the medications. L01 used just one pillbox for both her morning and evening pills, so she had the habit of opening up the pillbox in the morning when she took her morning pills and leaving it open to remind her that she has evening pills to take later. Once taking her evening pills (and inverting the pillbox to get the pills), she leaves that same pillbox door open until the next morning so that she has a visual reminder the next morning about whether she took her pills on the night before. She would close yesterday's pillbox door and open up the pillbox door for today and repeat the process. These habits were first self-reported by the individuals when a researcher asked them at the beginning of the study to walk through their pill-taking routine. These routines were also easily identified as a repeated pattern in the data. The pill-taking analysis code accommodated these habits via a parameter in a configuration file for these two study participants. For these individuals, a rating of 3 was assigned if they followed their normal pattern. A rating of 2 was assigned if they did not follow their pattern. In practice when the PASS is administered by the therapist, it is rare for them to assign a rating less than 2 because participants can usually rationalize why they might have deviated from their normal way of resetting the pillbox or that resetting the pillbox is not that important.

Other aspects of the pillbox door selection and opening step such as the number of "extra" doors opened before the individual selected the correct door could also be factored in the adequacy rating for this step. However, the concept of extra doors was not part of the original PASS task.

In summary, the rules for assigning adequacy rating for the medication-taking task focus on three of the eight steps outlined in the PASS. The minimum rating across the three steps of opening the correct pillbox door, removing/inverting the pills from the slot, and resetting the pillbox doors back to a state consistent with their routines provides the overall adequacy rating for the medication-taking task.

### 7.4.1.3 Rules of Assessing Coffee-Making

In the coffee-making task (Appendix G), the individual is asked to make a pot of coffee using the coffeemaker provided in the sensor deployment. By the first administration of the PASS, they have grown accustomed to using this coffeemaker for the past four months. Similar to the medication-taking task, the coffee-making task was designed to be observational, rather than prescribing specific instructions for the individual to follow. The therapist prompted the individual to make a pot of coffee with the coffeemaker and observed how they carried out the task, intervening only when required by the PASS. The PASS coffee-making task included a total of 12 steps for evaluation: 1) cleaning out the old coffee from the carafe, 2) measuring the water, 3) filling the water reservoir on the coffeemaker, 4) removing the old filter and grounds from the machine, 5) placing a new filter in the machine, 6) retrieving a can of coffee grounds from where it is kept, 7) measuring out the coffee grounds, 8) scooping coffee grounds into the filter, 9) putting away the coffee grounds, 10) closing the filter door on the coffeemaker, 11) placing the carafe into the machine, and 12) turning on the machine.

The sensor-augmented coffeemaker can sense six of the twelve steps in the coffee-making task, steps 2, 3, 4, 9, 10, 11, and 12. Step 2, measuring out the water, could be measured using the distance sensor mounted on the water reservoir door, but this data about this step was excluded from the automatic assessment because it was difficult for the therapist to know what was an unusual amount of water for that individual, and thus all individuals were scored a 3 for this step in the PASS. Step 9, putting away coffee grounds, could be detected using data from the cabinet sensor about whether the individual opened and closed the cabinet that contained the coffee. However, these data were not included in this analysis because some individuals were inconsistent in their storage of their coffee grounds.

Each of the remaining four steps (pouring in the water, closing the filter door, replacing the carafe, turning on the machine) was assessed individually to see whether the step was completed (in only one attempt), repeated more than once, mis-ordered (turning on the machine had to be the last step, but all other orders of the first three steps are acceptable), or missing (that is, either forgotten or not attempted). The overall adequacy score for the task was assigned a score based on how each step was performed. A top rating of 3 for the overall adequacy of the task was assigned if all steps were completed each in one attempt. A rating of 2 was assigned if there was at least one repeated or mis-ordered step. A rating of 1 was assigned if any of the steps were missing.

## 7.5 Results for Comparing Automatic Assessment with Performance Testing

The sensor data corresponding to each episode of PASS were used to generate adequacy ratings based on the rules for automatic assessment described above. To address research question RQ8.1, the automatically generated ratings based on the sensor data were compared to see how closely they aligned with expert ratings by an occupational therapist using the PASS tool. For each task, two statistics, Cohen's kappa ( $\kappa$ ) and Spearman's correlation coefficient ( $\rho$ ) were calculated to reveal how closely the sensor-based ratings and the PASS ratings aligned with each other. Cohen's kappa (Cohen 1960) represents agreement between two raters with *categorical* data and provides a conservative estimate of agreement because both the ratings based on the sensors and on the PASS are actually *ordinal* in nature (with 3 being greater than 2, which is greater than 1, which is greater than 0). Spearman's correlation coefficient (== CITE Spearman) also represents agreement between two raters but takes into account the ordering of the categories and thus provides a more powerful estimate of agreement between the sensor-based rating and PASS-based rating.

In the follow sections, the results of the analysis of the agreement between the sensor-based rating and PASS-based rating are described for each phone use, medication taking, and coffee making. There was a fairly high level of agreement in the phone use task, indicating that the phone sensor, even though it is limited in what it can sense, captured some of the most critical steps that determine performance. However, there was much poorer agreement in the medication-taking and coffee-making tasks due to the fact that individuals were performing simulations of the activity rather than actually carrying them out.

### 7.5.1 Agreement in ratings of phone use

Ten of the twelve participants in the study had a phone sensor installed in their home to monitor their phone use. However, in two cases, L07 and L14, both chose to use their cell phone to perform the phone use PASS task instead of their landline because they felt more comfortable using their cell phone. Others also elected to use their cell phone to perform this task because they reported that their landline was not working correctly at the moment. There were also other instances where the phone sensor was not working correctly or the individual just refused to perform the task. At the end of three attempts to administer the phone use PASS task to each study participant, a total of 19 episodes had sensor data available for analysis. For each of these 19 episodes, the corresponding phone sensor data was calculated and the rules described in Section 7.4.1.1 were applied to calculate an adequacy score for phone use.

The initial analysis of phone use will include the first six steps, leaving out the last step of reporting the correct closing time of the pharmacy to the therapist. As discussed in Section 7.4.1.1, the phone sensor can only indirectly detect this step. Nonetheless, the overall adequacy ratings for the phone use PASS tasks were calculated based on the first six steps and the automatically generated sensor-based ratings were calculated based on the two steps (dialing the number and navigating the phone menu). The agreement between the 19 PASS ratings and sensor-based ratings was perfect with a Cohen's kappa value of  $\kappa=1.0$  and Spearman's correlation of  $\rho=1.0$ . This perfect agreement shows that the variability in the overall adequacy scores for the phone use task (excluding the last step) in the sample population can be accounted by the two steps that sensors can easily monitor (dialing the phone and navigating the phone menu). This shows that sensors can automatically detect at least two of the three most important factors in the performance of the phone use task.

The PASS data shows that in 5 of the 19 instances of the phone use task, the individual had difficulty in step 7, that is, reporting the correct closing time of the pharmacy (for example, reported the opening time of the pharmacy or the closing time of the store instead). With step 7 included in the analysis, the agreement between the 19 PASS ratings and sensor-based ratings drops to a lower level, with a Cohen's kappa value of  $\kappa=0.677$  ( $SE=0.1366$ ). A closer examination of the agreement among the different values shows that good agreement in the categories for '3' rating ( $\kappa=0.8831$ ) and '1' rating ( $\kappa=1.0$ ), but poorer agreement in the intermediate category for a '2' rating ( $\kappa=0.5778$ ). Using  $\alpha=0.70$  as a rule of thumb, we see the agreement as calculated by the kappa value shows somewhat substantial level of agreement despite the fact the sensor-based score can only indirectly assess the quality of step 7. Furthermore, using a more powerful calculation that takes into account the relative order of the ratings, Spearman's correlation is calculated to be  $\rho=0.8144$ , which represents a fairly strong correlation (and thus agreement) between the expert rating and sensor-based rating. Overall, the sensor-based approach generated ratings that matched well with the ratings from performance testing by a trained therapist.

### 7.5.2 Agreement in ratings of medication taking

The medication-taking PASS task was administered to each participant a total of three times, except for L14 who withdrew between the first and second visits. There was one instance when the augmented pillbox was not functioning correctly (L05's third PASS assessment) and so there was no sensor data available for comparison. Of the remaining 33 instances of the medication-taking PASS task, the corresponding sensor data was analyzed and adequacy ratings were calculated based on the rules described in Section 8.4.1.2. The analysis shows that there was little agreement between the sensor-based ratings and the PASS ratings, with a Cohen's kappa of  $\kappa=-0.021$ . The correlation between these two sets of ratings also reveals poor agreement, with Spearman's correlation coefficient of  $\rho=-0.0231$ . The negative sign on both the kappa and correlation coefficient indicate that the observed agreement between the two ratings is actually lower than what is expected by chance. These results indicate that sensors are not able to assess how well individuals are able to carry out their tasks when demonstrating them in the presence of the therapist.

A closer examination of the PASS data and the sensor data reveal some interesting differences that can explain the poor agreement between the sensors-based rating and the PASS rating. The sensor-based assessment takes into account only three of the eight steps outlined in the medication-taking task. Either the discrepancy lies in a misinterpretation of the three steps or not taking into account the other five steps that affect the overall adequacy score for the task. In fact, it is the former (a misinterpretation of one of the steps) that accounts for a significant difference between the sensor-based and PASS ratings.

The discrepancy lies in whether the individual inverted the pillbox or not. In many cases during the PASS, the individual did not invert the pillbox when they normally would. The automated assessment appears to be classifying a number of instances rated as a "3" in the PASS as a "2" because these instances were missing an inversion. However, in these instances, the human expert administering the PASS did not mark the individuals off for not inverting their (empty) box but merely observed and listened to how individuals described how they would access their pills.

For the eight participants that regularly invert their pillbox, we administered the PASS medication taking task a total of 21 times. It is expected that participants invert their pillbox in each of these 21 instances because that is how they normally would carry out the task. However, in 8 of these instances, there were no inversions found in the sensor data during the PASS assessment. This discrepancy can be explained by the fact that most of the PASS administrations were simulated rather than purely observational, that is, participants were asked to perform the medication-taking task outside of their normal routine. For example, individuals were asked to imagine as if they had not taken their morning pills and to pretend to take their morning pills (even though they had already taken their morning pills and the slot was already empty). If the pillbox was already empty of their morning pills, then the individuals might not automatically invert the box to pour out the "invisible" pills in their simulation of their pill-taking routine. This discrepancy between their actions during the PASS assessment and the actions they normally do (and the actions they actually performed earlier that morning) illustrates some of the limitations of live expert assessments. A prompted request to simulate a task does not always provide the individual with the physical affordances typically present in the task (such as having pills in the pillbox) to carry it out in the same way as if they were actually performing it. Furthermore, when looking at the pill taking that the individuals self-initiated on their own (earlier or later in the day when they were not simulating their tasks for the therapist), in each of these 8 instances where they did not invert the pillbox in the presence of the therapist, they actually inverted the pillbox to take their morning or evening pills as part of their normal routines. If the rules were changed to use whether or not they inverted the pillbox earlier or later in the day as part of their normal routine, then the results provide a much stronger agreement between the sensor-based rating and the PASS rating. In fact, the Cohen's kappa value with the revised rules would be  $\kappa=0.5753$  and the Spearman's correlation coefficient would be a strong  $\rho=0.7185$ . This discrepancy between the pill-taking actions that the individuals performed during the PASS and the actions they do in regularly in their normal lives begins to reveal how different an individual might behave during a visit from a therapist and when they are on their own living their everyday lives. Overall, the rating generated using sensor-based approach for the medication-taking task matches fairly well with the rating from performance testing, but steps skipped because the individual was simulating the task can mislead the sensor-based approach into miscategorizing the rating.

### 7.5.3 Agreement in ratings of coffee making

Across the six participants who used the augmented coffeemaker to make their coffee, the coffee-making PASS tasks was administered a total of 17 times. There were two instances in which L09 had just made a fresh pot of coffee prior to the therapist's arrival so she did not want to make another pot. Another instance involved L07 simply deciding she did not want to perform that task at the moment. Thus, there were a total of 14 successfully administered PASS tasks. For each of these 14 coffee-making instances, the corresponding sensor data from the coffee maker was analyzed and a rating was calculated automatically according to the rules described in Section 8.1.4.3. Similar to the results found

in the analysis of medication taking, there was little agreement between the PASS rating and sensor-based rating, with Cohen's kappa of  $\kappa=0.1327$  and a Spearman's correlation coefficient of  $\rho=0.0542$ .

The low agreement, like in the case of the phone use task, points to either a misinterpretation of the existing steps that the sensors can monitor or a lack of accounting for the other steps of the coffee making that the sensors cannot monitor (such as how well the coffee grounds are scooped or cleaning out the old coffee from the pot). A closer look at the data shows that it is a combination of both the former and latter. In two instances when participant L05 was simulating the coffee-making task, she did not complete the task by pressing the power button to start the brewing process, instead she explained to the therapist what she would press to start and the therapist found that to be acceptable. The sensor data reveals that she did not press the button, but the PASS data says she did (or at least explained what she would do). Another example of a misinterpretation of the data is when participant L06 actually had to open and close the water reservoir door a few times (instead of just once) during the task. The rule-based system marked this as a repeated water-filling step and correspondingly assigned a rating of 2 to the overall task. However, the therapist either did not notice this repetition or deemed it was an acceptable repetition, and assigned an overall rating of 3.

Other instances of misclassification also reveal factors that affect task adequacy in the steps in coffee making that the sensors cannot monitor. The therapist noticed issues such as having difficulty placing the filter into the machine, being sloppy when scooping out the coffee grounds, or difficulty measuring out water when filling it from the faucet. These factors were unaccounted for in the sensor-based ratings, yielding a discrepancy between the sensor-based ratings and the PASS ratings. The analysis of the discrepancies between the sensor-based ratings and the PASS ratings show that 1) simulated tasks prompted by a therapist sometimes does not reflect what people actually do when they carry out the task in their everyday lives and 2) the sensing-based approach is limited in the factors it can monitor (but are actually easily noticed by a trained human observer) and this also affects how well it can classify coffee-making performance.

#### 7.5.4 Summary of Agreement between Sensor-based Ratings and Performance Testing Ratings

Addressing research question RQ6.1, the analyses of the agreement between the sensor-based ratings and the PASS ratings for the phone use, medication taking, and coffee making tasks provide evidence that the sensor-based approach captures many of the critical factors for evaluating task performance. In the phone use task, the sensors only consider two of the seven steps in the task and yet the ratings generated based solely on the sensor data can match the PASS ratings fairly well. Likewise, the medication-taking task considers only three of the eight steps in the task and yet there is still a fairly substantial amount of agreement and a strong correlation between the sensor-based ratings and the PASS ratings.

The analysis of the medication-taking and coffee-making tasks reveals that there are significant differences in how an automated system and human expert would assess how an individual takes their medications and make coffee. Individuals tended to act differently in the presence of an observer, in a way that they might not normally behave on their own. Thus, in situations where the individual is asked to simulate a task in the presence of a human evaluator, the human evaluator is better at assessing the quality of the simulated task because during a simulation, certain shortcuts are taken subconsciously, such as not inverting the pillbox or turning on the machine when making coffee, for which the human evaluator can reasonably assume that the individual can perform that step. However, the sensor-based assessments see these shortcuts as missing steps and knowing no better, assign a lower rating. In the medication-taking task, individuals who normally inverted their pillbox to retrieve their pills often did not invert the pillbox when demonstrating how they took their morning pills to the therapist. Likewise, in the coffee-making task, individuals might not actually press the button to start the machine when they are simulating how they make coffee in front of the therapist. This analysis provided concrete examples of how people behave differently when being assessed and when they performing the same tasks on their own.

The coffee-making task illustrates not only how performance during a test might differ from actual routine behaviors but also how other factors that sensors cannot detect such as how well the coffee grounds are scooped or how well the individual is able to measure out the water from the faucet that a human evaluator can notice and factor into the overall task adequacy score. Thus, to get a complete picture of functioning, human evaluators still must play an important role in assessing aspects of the task that technology is (currently) unable to monitor well. Nonetheless, automated rule-based assessment derived from an individual's typical patterns when performing the task alone is more accurate at assessing instances when the individual is alone. In the next section, we quantify the differences between how individuals perform tasks in the presence of a human evaluator and how they perform tasks on their own.

## 7.6 Representativeness of Behaviors During Performance Testing

The goal of performance testing is to understand how individuals carry out tasks important for independence. One of the drawbacks of performance testing is that it is performed infrequently and thus it might not be able to accurately capture how individuals carry out their tasks in their everyday lives. On the particularly day of testing, the individual might unconsciously change how they carry out the task. The individual may feel additional pressure to perform better (or worse) when tested by a human evaluator than when alone, in other words, a testing effect. In psychometrics, the term *reactivity* is used to describe the testing effect because the individual being tested reacts to the test itself and changes her performance (Portney & Watkins, 2008). The concern of a testing effect in performance testing for everyday functioning has been raised by researchers who develop new performance-based assessment techniques (Kapust & Weintraub 1988; Matheson et al., 2002), but the reactivity of performance testing has not yet been actually measured. It was not previously possible to monitor objectively how individuals typically behaved in the absence of a human observer. Comparisons of performance testing with self-reports (such as Myers et al., 1993) show differences but do not address the possibility of a testing effect for performance testing.

Moreover, the analysis of the results from the automatic assessment suggest that the human evaluator is often limited to evaluating only simulated, prompted tasks rather than observing how people carry out the tasks as they would carry them out in their normal routine. While simulating a task, the same affordances (such as the presence of pills in the pillbox or the physical weight of water when making coffee) as when the task is actually performed might not be present to guide the individual. The trend for performance testing is to assess individuals in their own context such as their home with the goal to be able to collect more ecologically valid data about task performance. However, the infrequency of assessments and common use of simulated tasks are still a threat to ecological validity. Thus, a natural question that follows is research question RQ6.2 about whether the actions people take during a visit from a human evaluator for performance testing differ from the actions individuals would normally perform when alone.

In this section, we quantify the testing effect of performance testing by using the sensor data to compare the actions individuals took during a visit from an occupational therapist with the actions taken on their own. We consider the medication-taking and the coffee-taking tasks because these tasks did not require the individual to follow specific instructions during performance testing but instead they were told to carry them out or demonstrate how they normally would carry them out. In contrast, the phone use task was a somewhat contrived task where the individual had to follow directions to call a pharmacy and ask for its closing time. Thus the phone use task was not a replication of a task that individuals performed regularly and does not lend itself for direct comparisons between how it was performed during performance testing and when the individual is alone. The following sections detail the differences in how the medication-taking and coffee-making tasks are performed when being tested and when alone, followed by a summary of these differences.



## 7.6.1 Differences in Medication Taking Actions

During performance testing, individuals demonstrated how they took their pills so that the occupational therapist could observe and assess how well they were able to take their medications using the pillbox. In only 2 of the total 34 times the medication-taking task was assessed by the occupational therapist did the individual actually take her pills instead of just simulating the task by walking and talking through the steps with the therapist. In this section, we investigate whether these 34 instances observed by the therapist are representative of the instances of the medication taking when the individual is alone. We consider differences along three metrics of the medication-taking task: task duration, the number of extra doors opened, and whether the pillbox was inverted or not. Task duration gives an indication of effort or attention paid to the task. The number of extra doors opened shows how inefficiently the task is performed. Whether the pillbox was inverted or not is important for showing that individuals actually carry through with taking their pills instead of just opening the box. The results show that the pill-taking behaviors of (not) inverting the pillbox and taking a longer time to carry out the task that were observed during performance testing with a human evaluator actually is statistically different than what individuals normally did on their own.

### 7.6.1.1 Task Duration

Task duration of the medication-taking task is a measure of how long it takes an individual to take her pills. By looking at the timestamps in the pillbox sensor data, the task duration is calculated as the number of seconds between picking up the pillbox (to initiate the pill taking task) to setting it down (to end the pill taking task) and not moving the pillbox again for at least five minutes. When the accelerometer in the pillbox detects a change in acceleration in any direction (that is, any sort of movement), the pillbox wakes up the wireless radio and begins to transmit the accelerometer and door state data to the laptop where the events are promptly timestamped. The timestamp of the first accelerometer event serves as the starting point of an instance of pill taking. The time for the action of setting down the pillbox to end an instance of pill taking is found by looking for the accelerometer event that precedes a series of steady-state accelerometer values that reflect no vertical (z-axis) movement. Task duration can be interpreted as the amount of effort the individual is devoting to the task, with greater time indicating greater effort or attention.

In this analysis, the natural logarithm is used to transform the number of seconds to minimize the correlation between the mean and the variance of the data and also to transform the data to align more closely with the normal distribution. A mixed-effect ANOVA model was used to find differences in task duration between the contexts of whether the instance was performed when being tested by a human evaluator during a PASS assessment or when alone as part the individual's normal routine. In this model, the participants were treated as a random factor and the context of the pill-taking instance (TEST or ROUTINE) as a two-level factor. The results show that the task duration for pill taking when being evaluated by the therapist (mean=3.60 [70.7 seconds on non-log scale], SE=0.31) was significantly higher than the duration when individuals performed the same task alone (mean=2.80 [58.4 seconds on a non-log scale], SE=0.25) ( $F[1,4822]=19.22$ ,  $p<0.001$ ). On the days when the therapist visited to observe their medication taking, individuals took on average 12 seconds (or 21%) longer than they normally took when performing the same task on other days when the therapist was not present. The mere effect of being tested seems to have resulted in individuals taking longer and being more careful when taking their pills. The difference in task duration suggests that individuals are paying closer attention to what they are doing and exerting greater effort to make sure they do it correctly in the presence of the therapist.

### 7.6.1.2 Opening Extra Doors

The number of extra doors opened is a metric that counts the number of pillbox doors opened in addition to the correct door, the door that matches the current day of the week. For example, if it is Thursday and the individual opens (and optionally closes) the Monday and Tuesday doors on the pillbox before opening the Thursday door, the number of extra doors is equal to two. If the individual does not open the correct door at all, then number of extra doors is not

considered in the analysis. Instances when the individual is refilling the pillbox, which is identified by the opening of at least five doors on the pillbox, are also excluded from this analysis. The number of extra doors opened can indicate how confused or unsure the individual is when selecting the correct pillbox door to open and thus it is a measure of the inefficiency in the pill-taking task. An optimal pill-taking instance would have zero extra doors opened.

The number of extra doors opened for each pill-taking instance was calculated and associated with a context, that is whether it was the pill taking was performed for the PASS under the observation of a therapist or performed alone in their normal routine. The majority of the data points for the number of extra doors opened was zero, indicating that individuals often chose the correct pillbox door without having to open any others before or after. Thus the variance in the data when individuals took their medications on their own was fairly low. Instead of treating each instance as a separate data point and using a mixed model ANOVA for analysis, the rates of extra doors opened was calculated in the context when the individual was performing the task for the PASS and when the individual was performing the task alone in their normal routine. The rates were calculated by summing up the number of extra doors opened and dividing it by the number of pill-taking instances. For example, L01 averaged 15 extra open doors per 100 pill-taking instances when taking pills on her own and 0 extra doors per 100 pill-taking instances when being assessed by a therapist. These two rates were calculated for each of the 12 study participants. A paired t-test was used to compare the mean rates across the two contexts while controlling for individual differences. The mean number of extra doors opened was 9.2 doors per 100 instances in the ROUTINE context and 8.3 doors per 100 instances in the TEST context. The results show that there was not a significant difference between the rates in the two contexts (TEST vs. ROUTINE)  $t[11]=-0.2029$ ,  $p=0.841$ . Given that the sample size in the PASS context was limited to at most three, it is not surprising to expect that there needs to be a much larger difference between the means to reach statistical significance. Nonetheless, it is interesting to note in the results that individuals tended to open more extra doors (which means they are being more “inefficient”) when they are alone than when they are being tested. Another interesting finding in the data was that 9 of the 12 individuals did not open any extra doors at all (and can be considered very efficient) when being tested, but when taking their medications on their own, they open extra doors at a non-trivial rate, averaging 8.5 doors per 100 pill-taking instances with a standard deviation of 4.9 doors. The remaining 3 of the 12 participants who did open an extra door when being evaluated each opened an extra door only once across the three times they were evaluated, and because there were only three PASS instances the statistical rate of extra doors opened for them was extrapolated to 33 per 100 pill-taking instances. The small sample size for the TEST context was likely the main factor for a lack of statistical significance for this metric. Nonetheless, the relative rates across the TEST and ROUTINE context still are useful for understanding whether merely being evaluated by a human evaluator affects how efficiently a task is carried out. In this case, even though the results are not statistically significant, the difference in the means suggests that individuals tend to be more inefficient (opening more extra doors) when taking their medications on their own than when being tested.

### 7.6.1.3 Inverting the pillbox

Some individuals turn the pillbox upside down when taking their pills to pour out the pills into their hands. Others, who keep only one large pill in the box, keep the box stationary on a surface and use their fingers to grab the pill rather than inverting the pillbox. For those individuals who routinely invert their pillbox, knowing that they invert their pillbox can indicate they are following through with the pill-taking task safely and not just gaming the sensor. Inversion is generally a fairly stable trait in the pill-taking routine. That is, individuals who invert the pillbox tend to always invert the pillbox, close to 90% of the time. Those who do not invert the pillbox do so very rarely at around 4% of the time. Thus, if there were no testing effect as a result of performance testing with a human evaluator, then it would be expected that the rate of inversion would be roughly the same when the individual is being tested and when the individual is performing the task alone. To identify whether the data supports this expectation, the inversion rates for each individual was calculated by summing the number of times the pillbox was inverted and dividing the sum by the number of pill-taking instances in each context: TEST (when being tested by the therapist) and ROUTINE (when the

individual is taking her pills on her own). The analysis approach is similar to the analysis for the rate of extra doors opened. A paired t-test was used to identify whether any differences existed in how often the pillboxes were inverted between the TEST and ROUTINE contexts, while controlling for individual differences. The results reveal that the inversion rate in the ROUTINE contexts (mean of 76 inversions per 100 pill-taking instances) was significantly higher than in the TEST context (mean of 44 inversion per 100 pill-taking instances)  $t[8]=-2.858, p=0.0212$ . These results quantify how individuals inverted the pillbox much less when being tested than what they would normally do when taking their pills on their own. As discussed in Section 1.5.2, individuals simulated the pill-taking task during the PASS assessment and may not have inverted the pillbox when demonstrating how they took the pills they already consumed earlier in the day.

## 7.6.2 Differences in Coffee Making Actions

Individuals were asked to make coffee using the coffeemaker in the PASS test so that the therapist could observe and assess how well they performed this multi-step task. Six of the twelve individuals made coffee using the instrumented coffeemaker on regular basis (at least 3 times a week). An occupational therapist administered the PASS coffee making three times for each individual at different times. On four of these visits, the individuals refused to perform this task because they just made a fresh pot of coffee or did not feel like performing this task, leaving a total of 14 instances of coffee making logged by the instrumented coffeemaker while the individuals were assessed by the therapist. In this section, we investigate whether these 14 coffee-making episodes differ from the how individuals make coffee on their own as part of their routines.

We consider three metrics of the coffee-making task that should remain the same if there is no testing effect: the number of missing steps, the task duration, and the amount of coffee. The number of missing steps can be considered a measure of forgetfulness when performing the task. The task duration can indicate the level of mental alertness, especially because coffee is usually made early in the morning. The amount of coffee made does not serve as a measure of abilities per se, but it is a feature of the task that should remain fairly stable because most individuals, who all live alone, make roughly the same amount of coffee every time. The results of the analysis show that there were some differences between how they made coffee when being tested and when they are alone. Individuals missed significantly fewer steps, that is, made fewer errors, when being evaluated in person by the therapist. Task duration was also found to be different, though not statistically significant, but the results still suggest a trend that individuals were more alert and faster when performing the task when being tested than when performing the task on their own. The amount of coffee brewed was found to be similar between the times individuals were being evaluated and when they made their coffee alone.

### 7.6.2.1 Missing Steps

A step is classified as a missing step in the coffee-making task if the individual did not complete a step before turning on the machine. For example, they may have forgotten to put water in the machine, turn on the machine and realize only after the machine beeps to warn the user that there is no water in the reservoir. Another common missing step is to forget to put the carafe back into the machine before turning on the machine. The number of missing steps is an indicator for forgetfulness. A perfectly executed instance of coffee making would have no missing steps.

The number of missing steps was calculated for each instance of coffee making. The number of missing steps can range from 0 to 3, though most commonly the number of missing steps in any given coffee making instance is 0 or next most commonly, 1. Instead of treating each instance as a separate data point, the rates of missing steps was calculated in under two contexts: 1) when the individual was performing the task for while being TESTed by the therapist and 2) when the individual was performing the task alone in their normal ROUTINE. The rates were calculated by summing up the number of missing steps and dividing it by the number of coffee-making instances. For example, L01 averaged 4.5 missing steps per 100 coffee-making instances when taking pills on her own and 0 missing steps per 100 pill-taking

instances when being assessed by a therapist. These two rates were calculated for each of the 6 study participants who used the instrumented coffeemaker. A paired t-test was used to compare the mean rates across the two contexts and to control for individual differences. The mean number of missing steps was 8.9 missing steps per 100 instances in the ROUTINE context was significantly higher than the mean of 0.0 missing steps per 100 instances in the TEST context. The results show that there was a significant difference between the rates in the two contexts (TEST vs. ROUTINE)  $t[5]=-3.883$ ,  $p=0.0116$ . Individuals made far fewer mistakes when making coffee when being evaluated (in fact they made NO mistakes at all) than when they are making coffee on their own. The testing effect seems to result in participants performing better (making fewer mistakes) than they normally would on their own.

### 7.6.2.2 Task Duration

Task duration is a measure of how quickly the individual starts and completes the steps in making a pot of coffee. Coffee making, unlike other tasks, is one of the first tasks performed at the beginning of the day right after the individual wakes up. Medications are also taken early in the morning but taken with breakfast (after the coffee is made). How long it takes an individual to make coffee is an indicator of the efficiency of their coffee-making process as well as an indicator for the alertness of the individual. The task duration is calculated as the number of seconds between the timestamp of the first action (for example opening the filter door or taking the carafe out) and the last action, which is typically pressing the power button to start the brewing process.

The task duration was calculated for each coffee-making instance. A mixed-effect ANOVA model was used to find differences in task duration between the contexts of whether the instance was performed when being tested by a human evaluator during a PASS assessment or when alone as part of the individual's normal routine. In this model, the participants were treated as a random factor and the context of the coffee-making instance (TEST or ROUTINE) as a two-level factor. The results show that the task duration for coffee making when being evaluated by the therapist (mean=119.2 seconds, SE=32.9) was lower than the duration when individuals performed the same task alone (mean=146.7 seconds, SE=20.4). The differences in the means are not statistically significant  $F[1,926]=1.46$ ,  $p=0.226$ , though the differences between the means do show that on the days when the therapist visited to observe their coffee making, individuals took on average 27 seconds (or 18%) less than they normally took when performing the same task on other days when the therapist was not present. Individuals appeared to have been more alert and more efficient when making coffee when the therapist was observing. A higher level of alertness is most likely due to the fact that the therapist visited either in the late morning or afternoon, typically after the individual had already had their morning cup of coffee. Individuals are likely slower when performing the task on their own because they normally perform the task right after getting up before they have had breakfast or have taken their medications. In contrast to the medication-taking task where individuals slowed down when being tested, individuals sped up their coffee making because they were more alert. The difference in the level of alertness does not apply as much to the medication-taking task in which they slowed down to be more careful. These results show that individuals tend to perform the coffee-making task more efficiently when being observed by a human evaluator.

### 7.6.2.3 Amount of Coffee Made

Individuals in the study typically made coffee for themselves because they lived alone and thus would make roughly the same amount of coffee every time they made coffee. The amount of coffee they made can be measured by how long the machine needs to run to brew the coffee. The brew time is linearly related to the amount (cups) of coffee made. The brew time is measured by looking at the pattern of electrical current used by the machine.

The brew time was calculated for each instance of coffee making. A mixed-effect ANOVA model was used to find differences in brew time between the contexts of whether the instance was performed when being tested by a human evaluator during a PASS assessment or when alone as part of the individual's normal routine. In this model, the participants were treated as a random factor and the context of the coffee-making instance (TEST or ROUTINE) as a

two-level factor. The results show that the brew time for coffee making when being evaluated by the therapist (mean=295.9 seconds, SE=46.45) was slightly higher than the duration when individuals performed the same task alone (mean=271.7 seconds, SE=42.07). The small difference in the means is not statistically significant  $F[1,929]=1.46$ ,  $p=0.227$ . It is also unlikely given the fairly large value of the standard errors for the means in the TEST and ROUTINE contexts that a larger TEST sample would increase the level of significance. Thus, the results show individuals tend to make the same amount of coffee when being tested by a human evaluator as when they are making it on their own. Individuals likely are relying almost entirely on their routines to determine the amount of coffee they make, and thus the amount of coffee made seems to be a fairly stable trait of coffee making across both contexts. There is little to be gained by changing the amount of coffee made, but in fact, changing the amount of coffee requires two changes in their routine (a different amount of water and a different amount of coffee grounds) which may require additional attention that the individual would rather devote to performing the task correctly. Performance testing seems to be able to capture the typical amount of coffee being made, which shows that some aspects of the task are representative of the individual's typical actions.

### 7.6.3 Summary of Representativeness of Evaluated Tasks

Performance testing to assess how individuals carry out IADLs is limited by the frequency of visits. Like many traditional forms of administered testing, there is also the potential for the test itself to affect how individuals behave. Given that one of the main benefits of in-home performance testing is to be able to measure how people function and behave in their own environments, it is important to address research question RQ6.2 and quantify any differences that might exist between the observed behaviors during testing and the typical behaviors found in an individual's everyday life. Below is a table that summarizes the results of differences between tasks performed when being tested in their homes by an occupational therapist and performed when alone.

The results show that individuals took significantly more time when taking their medications and also inverted the pillbox significantly less when being tested. Individuals also were on average more efficient and opened fewer extra pillbox doors when being observed. For the coffee-making task, individuals did not leave out any steps when making coffee when being tested but typically left out a small, but non-trivial, average number of steps when making coffee alone. Individuals were faster and more alert when making coffee when being observed due to the fact that coffee making typically happens early in the morning when the individual is not as alert and may be less efficient. However, the amounts of coffee they made when being watched and when alone were about the same because it is easy to follow the same routine and use a routine amount of water and coffee grounds. Across both the medication and coffee tasks, there were significant differences in how the task was carried out when the individual was being tested. These results highlight how performance testing by a human evaluator in the home, even though considered to be more accurate and more ecologically valid than other reporting techniques such as self-report and performance testing in the clinic, avoidably imposes a testing effect that changes the way individuals carry out tasks when being watched. In contrast, the sensor-based assessments are not limited to only observing behaviors during isolated visits but can unobtrusively observe how the individual naturally behaves in their own home when not performing for a test.

Task	Task Feature	Performance with human observer	Statistically significant?	Rationale for different performance when being tested
Medication taking	Task duration	Higher than typical	Yes	Being more careful when human observer is watching
	Number of extra pillbox doors opened	Lower than typical	No	Being more efficient when human observer is watching
	How often pillbox was inverted	Lower than typical	Yes	Did not perform all steps of the task when simulating pill taking

Coffee making	Number of missing steps	Lower than typical	Yes	Being more careful when human observer is watching
	Task duration	Lower than typical	No	Typically perform task early in the morning when alertness is low, and testing happens later in the day when they are more alert
	Amount of coffee made	No difference	n/a	Making the typical amount of coffee according to routines is easiest sequence to take

**Table 7-1. Summary of how individuals performed the medication-taking and coffee-making tasks differently when being tested and when alone.**

## 7.7 Comparing Strengths of Sensor-based Assessment and Performance Testing

To address research question RQ7, this section discusses the relative strengths of sensor-based assessment and performance testing. Automatic sensor-based assessment and traditional performance testing conducted by an occupational therapist have overlapping as well as complementary strengths.

### 7.7.1 Sensor-based assessments capture critical steps

Automatic sensor-based assessments were found to capture many of the steps in a task that contribute to most of the variance in the how tasks are performed. For example, in the medication-taking task, sensors were able to capture three of the most important steps—whether the correct door was opened, whether the pillbox was inverted, and whether the pillbox doors were reset. Likewise, the simple circuit plugged into the phone line could detect two of the three most important steps in the phone use task—whether the number was dialed correctly and how efficiently individuals navigated the store’s phone menu. As discussed in Section 7.5.4, monitoring only this limited number of steps, automated sensor-based ratings still corresponded fairly well with the ratings from the human evaluator.

### 7.7.2 Performance testing has a wider scope of evaluation

In contrast, in some tasks that involve more objects and interactions, a sensor-based approach is limited in the scope of its observations and can still be misled by unaccounted steps. Whereas the sensors captured most, if not all of the critical steps, in the phone use and medication-taking tasks, the instrumented coffeemaker in the coffee-making task was more limited in its ability to sense some of the steps that impact the overall task adequacy rating. For example, in the coffee-making task, the occupational therapist can observe if the individual spills coffee grounds when scooping them into the coffeemaker or how easily the individual aligns the paper filter into the machine. Assessing whether steps like putting in coffee or a filter is easy but assessing the quality of the process of these steps is difficult. In 3 of the 14 instances of coffee making observed, the therapist noticed that the individual was messy with scooping the coffee grounds into filter and subsequently reduced the adequacy rating for that instance. The sensor-based approach can only sense if the individual has put in water in the machine or opened/closed the filter drawer and cannot take into account the quality of the coffee grounds scooping and water filling steps. Performance testing with a human evaluator also can detect subtle and infrequent safety issues that a sensor-based system may not be able to detect. For instance, the therapist noticed that one individual (L11) dropped her pills on the ground when she inverted the pillbox. In another instance, the therapist noticed that the individual (L07) was wearing her oxygen tube, but when making coffee in the corner of the kitchen, the oxygen tube was stretched in a way that made it difficult for the individual to reach certain items of the kitchen to make coffee. Especially for tasks with many steps that require the individual to interact with multiple objects located in different parts of the room, the sensor-based assessment approach can be inaccurate due to being limited in the number and scope of steps it can track. In-person performance testing with a trained therapist can identify these issues much more easily.

Another source of error in sensor-based assessment is when individuals do not follow the same process for carrying out the task. For example, if individuals normally invert the pillbox to pour out their pills when taking their medication, but they decide on one day to not invert the pillbox, then the absence of an inversion can be construed by the system as an error and reduce the rating. These deviations from normal patterns were found to be particularly common when the individual was asked to simulate the task. When simulating the task, individuals tended to skip steps (like inverting the pillbox) or just talk about steps (like turning on the coffee machine) instead of actually physically performing every step they normally would do when actually performing the task on their own. Thus, sensor-based assessment, because of the rigidity of the rules it applies, can be misled by simulated (rather than performed) actions and interpret the missing step as an error. In the case of simulations of tasks, the human expert is a better observer of the task because the human can monitor both the actions performed as well as any descriptions of actions that the individual intended to perform but did not because of the circumstances of the simulation.

### 7.7.3 Sensor-based assessment can capture typical behaviors over time

Whereas traditional performance testing is better suited for isolated episodes of simulations of tasks, a sensor-based approach is ideally suited for monitoring routine task performance in the everyday lives of individuals. The choice of sensors and devices used in this study were designed to be unobtrusive and indeed individuals reported that they carried on with their normal routines without feeling like they were constantly reminded their behaviors were being monitored. As a result, individuals felt comfortable “just being themselves” after a few weeks, and the longer term deployment of the sensors motivated individuals to let the sensing fade into the background. Thus, the behaviors captured by the sensors were as close to the behaviors if there were no monitoring in place, essentially the actual behaviors individuals would normally perform.

A sensor-based approach is well suited for monitoring an individual’s true behavior over a long period of time. As the study described in this dissertation demonstrates, the value of sensor-based assessment is to be able to see how behaviors (in this case, task performance) change or stay the same over a long period of time (in this case, 10 months). Equipped with data points for each task from nearly everyday over long period, a system (or human interpreter) can quantify what behaviors are considered typical for the individual. For example, an individual with severe arthritis in his hands may report that he normally inverts their pillbox to pour out his pills but in performance testing with a simulated task, he does not invert the pillbox. To resolve this contrast, the sensor data can provide objective and frequent data about from the individual’s historical pill-taking actions. A therapist can based her judgment on both what she saw during the assessment, what the individual self-reported, and also the objective sensor data to determine whether the individual regularly inverts the pillbox when taking his pills. As a result, the therapist can assess whether the arthritis is impacting whether the individual is able to hold all the pills when inverting the box and provide adaptations to the task to improve the safety and efficiency of task. The long-term sensor-based approach for monitoring can also provide data that complements the measurements taken during performance testing. For example, it is common in performance testing for individuals to perform the task flawlessly without error because they want to demonstrate a high level of skill for the observer, but the sensor data from other instances of the same task performed alone by the individual may show some errors. Thus, performance testing usually assesses an individual’s abilities (what they *can* do if they really tried) but long-term sensor data about task performance can provide a window into an individual’s habits (what they *normally* do on their own). An individual’s abilities form the upper bound on how well (independently, safely, and adequately) a task is performed regularly. Individuals may not normally operate at their peak ability levels in their everyday lives but it is in their everyday lives that individuals encounter the hazards and consequences (such as taking the wrong pills, misdialing the phone, or spilling their hot coffee) of their typical, normal level of performance. Everyday assessment of task performance using sensors is critical for knowing how to provide the right interventions for an individual’s everyday lives to help them operate a level closer to their abilities (or indeed raise their abilities). Therefore, a sensor-based approach provides a more ecologically valid data stream than performance testing that can reveal how individuals

typically perform their tasks and potentially assist therapists in knowing how to support the individual in their everyday lives.

#### **7.7.4 Sensor-based assessment can capture more precise measures than performance testing**

A sensor-based approach naturally lends itself to using computation to quantify precise measures of performance such as precise timing and sequencing of individual actions as well as aggregating measures across steps or task instances for task duration or average task durations. The value of precision in measures of task performance lies in the fact that more precise measures can reveal subtle changes useful for tracking trajectories of change. Subtle patterns can emerge in the data with greater precision and can provide earlier warning signs of changes.

Precise timing of tasks are not commonly recorded in traditional performance testing because the therapist is often focused on observing the individual rather than focused on operating a stopwatch. Explicitly timing an individual can also put undue pressure on the individual performing the task. Another reason precise timings such as task duration are not collected in performance testing is these precise values only make sense if they are interpreted with respect to the norm for that individual, and performance testing is not performed frequently enough to be able to establish these individual-specific norms. A passive sensor-based approach can address the limitations encountered by the therapist. The system can easily record timings without losing focus on how the task is performed and do so passively and unobtrusively so that it does not place additional testing pressure on the individual. The system can also assess on frequent basis, providing enough data points from different time periods to be able to assess if an individual is performing at a level different from before. In fact, the dwellSense system can easily capture precise task durations for medication taking and coffee making over time or misdial percentages for phone use to identify trends of change in the data. These types of data at these levels of precision were not previously easily collected through traditional performance testing. Sensor-based assessment can quantify task performance at a level of precision that was not previous possible with traditional performance testing.

## **7.8 Summary**

In this chapter, automatic sensor-based assessment was compared with traditional performance testing by a human expert. An occupational therapist administered a traditional performance testing tool, the PASS, to assess and rate how well individuals performed medication taking, phone use, and coffee making. The ratings from the therapist were compared with the ratings generated automatically based on the sensor data that also recorded the actions the individuals took during the performance testing. The sensor-based assessment relied on rules based on the rules used in the traditional performance testing that assigned a rating for each task based on the data from sensors. The sensor-based approach was found to capture most of the critical steps in the phone use and medication taking tasks and thus generated a rating that matched well with the ratings from performance testing. In order to identify whether performance testing unintentionally imposed any significant testing effects, the tasks actions taken during performance testing was compared with the typical actions that individuals took when performing the task on their own as part of their normal routines. The results of the analysis show that there are significant differences (longer task duration) in how the medication-taking task is carried out between the contexts of being tested and when the individual is alone. Likewise individuals made no mistakes when being assessed by the therapist but made small but non-trivial amounts errors when making coffee on their own. Therefore, sensor-based assessments are well suited to assessing, with great precision, well-constrained tasks like medication taking but traditional performance testing still had a wider scope during assessment as the therapist can notice safety or process issues that sensors, in their more limited scope, were not designed to detect. Performance testing is well suited to assess an individual's abilities, but the long-term sensor-based approach for assessing everyday task performance can provide a window into the individual's typical habits. Sensor-based data about how individuals typically perform their tasks can provide an important piece of context for interpreting the results of performance testing, tracking trajectories of decline over time, and also to know how to best



support individuals in their everyday lives. Thus, the sensor-based approach for assessing task performance can match the ratings of traditional performance testing in many cases but can also provide long-term data about an individual's typical level of functioning.

## 7.9 References

- Burns, T., Mortimer, J. A., & Merchak, P. (1994). Cognitive Performance Test: a new approach to functional assessment in Alzheimer's disease. *Journal of geriatric psychiatry and neurology*, 7(1), 46.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37–46.
- Cook, D. J., & Schmitter-Edgecombe, M. (2009). Assessing the quality of activities in a smart environment. *Methods of information in medicine*, 48(5), 480.
- Diehl, M., Marsiske, M., Horgas, A. L., Rosenberg, A., Saczynski, J. S., & Willis, S. L. (2005). The revised observed tasks of daily living: A performance-based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology*, 24(3), 211.
- Hodges, M., Kirsch, N., Newman, M., & Pollack, M. (2010). Automatic Assessment of Cognitive Impairment Through Electronic Observation of Object Usage. *Pervasive Computing*, 192–209.
- Hoeymans, N., Feskens, E. J. M., van den Bos, G. A. M., & Kromhout, D. (1996). Measuring functional status: cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990–1993). *Journal of Clinical Epidemiology*, 49(10), 1103–1110.
- Holm, M., & Rogers, J. (1999). Performance assessment of self-care skills. *Assessment in occupational therapy mental health: an integrative approach*. Assessments in occupational therapy mental health: an integrative approach (pp. 117–124). Thorofare, NJ: B. Hemphill-Pearson.
- Kapust, L. R., & Weintraub, S. (1988). The Home Visit: Field Assessment of Mental Status Impairment in the Elderly. *The Gerontologist*, 28(1), 112–115. doi:10.1093/geront/28.1.112
- Kempen, G. I. J. M., Steverink, N., Ormel, J., & Deeg, D. J. H. (1996). The Assessment of ADL among Frail Elderly in an Interview Survey: Self-report versus Performance-Based Tests and Determinants of Discrepancies. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 51B(5), P254–P260. doi:10.1093/geronb/51B.5.P254
- Loewenstein, D. A., Amigo, E., Duara, R., Guterman, A., Hurwitz, D., Berkowitz, N., Wilkie, F., et al. (1989). A new scale for the assessment of functional status in Alzheimer's disease and related disorders. *Journal of Gerontology*, 44(4), P114–P121.
- Loewenstein, D. A., & Mogosky, B. J. (1999). The functional assessment of the older adult patient. *Handbook of assessment in clinical gerontology*, 529–554.
- Matheson, L. N., Rogers, L. C., Kaskutas, V., & Dakos, M. (2002). Reliability and reactivity of three new functional assessment measures. *WORK-ANDOVER MEDICAL PUBLISHERS INCORPORATED*, 18(1), 41–50.
- Mihailidis, A., Boger, J., Canido, M., & Hoey, J. (2007). The use of an intelligent prompting system for people with dementia. *interactions*, 14(4), 34–37.
- Moore, D. J., Palmer, B. W., Patterson, T. L., & Jeste, D. V. (2007). A review of performance-based measures of functional living skills. *Journal of psychiatric research*, 41(1), 97–118.
- Myers, A. M., Holliday, P. J., Harvey, K. A., & Hutchinson, K. S. (1993). Functional Performance Measures: Are They Superior to Self-Assessments? *Journal of Gerontology*, 48(5), M196–M206. doi:10.1093/geronj/48.5.M196
- Portney, L. G., & Watkins, M. P. (2008). *Foundations of Clinical Research: Applications to Practice* (3rd ed.). Prentice Hall.
- Reuben, D. B., & Siu, A. L. (1990). An objective measure of physical function of elderly outpatients. *J Am Geriatr Soc*, 38(10), 1105–1112.
- Sager, M. A., Dunham, N. C., Schwantes, A., Mecum, L., & others. (1992). Measurement of activities of daily living in hospitalized elderly: A comparison of self-report and performance-based methods. *Journal of the American Geriatrics Society*. Retrieved from <http://psycnet.apa.org/psycinfo/1992-37597-001>
- Wadley, V. G., Harrell, L. E., & Marson, D. C. (2003). Self-and Informant Report of Financial Abilities in Patients with Alzheimer's Disease: Reliable and Valid? *Journal of the American Geriatrics Society*, 51(11), 1621–1626.
- Willis, S. L. (1996). Everyday cognitive competence in elderly persons: Conceptual issues and empirical findings. *The Gerontologist*, 36(5), 595–601.

Wilson, D., & Philipose, M. (2005). Maximum a posteriori path estimation with input trace perturbation: Algorithms and application to credible rating of human routines. Proceedings of IJCAI. Retrieved from <http://www.seattle.intel-research.net/pubs/ijcai05.pdf>

# 8

## Conclusion

The need for an objective, frequent, and ecologically-valid record of how individuals carry out tasks important for independence has long been recognized by individuals who value their ability to live independently as well as by their caregivers and clinicians who support their wellbeing. These types of records have hitherto been estimated based only on a combination of infrequent performance testing and biased information sources including self-reports and caregiver-reports. One of the earliest applications of ubiquitous computing was to embed sensors in spaces to activate them as “smart” spaces that can track and understand what activities individuals are doing (Weiser 1991). This thesis examines a specific application of sensing technology to collect more objective, more frequent, and more ecologically valid information about how individuals carry out tasks around the home important for independence than existing information sources. In doing so, this thesis investigates the both the potential and actual utility of the collected sensor data, how to design sensors to track how tasks are performed, the usability aspects of interacting with the sensor data, and how well an computer-automated approach for assessing the quality of task performance matches with expert-based assessment by a human.

This thesis describes the process of developing a ubiquitous, embedded sensing system for recognizing, monitoring, and assessing Instrumental Activities of Daily Living. The process began with formative evaluations of the embedded assessment concept and provided the initial design parameters and validation to design a prototype. The prototype was deployed for over 18 months to evaluate its usefulness and usability in-depth using a case study approach. The prototype was redesigned based on the lessons learned from the initial pilot deployment and then deployed to a larger sample of users. The larger deployment provided evidence that the system can assess task performance nearly as well as a trained occupational therapist as well as demonstrated how reflecting on sensor data about task performance can support a more accurate awareness and improved task performance.

### 8.1 Support for Thesis

The goal of this work was to prove the following thesis statement (from Section 1.3):

Embedded assessment of wellness can provide ecologically valid assessments of task performance and reflecting on the generated information supports new opportunities for timely assessment of functional abilities for older adults.

The thesis statement can be decomposed into two sub-statements:

- A. Embedded assessment of wellness can provide ecologically valid assessments of task performance.
- B. Reflecting on the generated information supports new opportunities for timely assessment of functional abilities for older adults.

Support for thesis sub-statement B was supported initially from the results of the initial concept validation study (Chapter 3). This work first examined the potential usefulness of sensing systems embedded in the home to track how individuals carry out tasks important for independence using a concept validation study. Stakeholders found that sensing concepts and the data collected from the concepts would provide new opportunities to understand changes in functional abilities and to intervene to fix problems before they lead to disability. In particular, older adults found the information potentially useful for understanding the subtle changes in their abilities. Their caregivers as well as therapists found the information potentially useful for knowing when and how to intervene and provide care for the

older adult. Physicians found the information potentially useful for seeing the longer-term trends in an individual's functioning that they would not normally have available.

With the potential usefulness of using a sensing-based approach established and information needs of stakeholders identified, this thesis describes the design, development, and deployment of dwellSense (Chapter 4), a system that uses sensors embedded in the home that recognizes, records, and assesses the quality of Instrumental Activities of Daily Living, in particular, medication taking, phone use, and coffee making. The dwellSense system provides a concrete example of a system that uses embedded assessment of wellness technique, providing the initial step in proving thesis sub-statement A.

The initial concept validation provided support for thesis sub-statement B based on hypothetical scenarios. To provide concrete evidence for thesis sub-statement B, this work also used a case study approach in the 18-month pilot deployment of dwellSense (Chapter 5) to reveal how reflecting on objective sensor data helped support an accurate self-awareness of how well IADLs were carried out. The two individuals in the case studies were able to use the sensor data as an objective, frequent, and ecologically-valid data source to compare with their own impressions and learn how well or poorly they were carrying out tasks important for living on their own. By engaging users with their own personal data, this thesis identified the sensemaking process and the meaning attached to the sensor data values. Breakdowns in the sensemaking process revealed new opportunities for additional sensor data streams to aid the automated interpretation of the sensor data in addition to additional forms of feedback to support the individual's abilities to understand and explain the behaviors captured in the sensor data. dwellSense was revised to capitalize on these opportunities by adding motion sensing to be able to automatically interpret whether individuals are excused from taking their medications because they are not home and also by adding a real-time feedback display to support a more continuous awareness of their behaviors.

To find further support for thesis sub-statement B beyond the two individuals in the case study, a larger deployment of dwellSense (version 2.0) with twelve community-dwelling older adults for ten months (Chapter 6) was conducted to identify the differences between reflecting on short-term task performance in real time and reflecting on long-term task performance without real time support. Reflecting on real-time task performance via an in-home information display helped individuals improve their task performance by providing them with immediate feedback on their sub-goals to take their medications correctly every day. Upon removing the real-time display, individuals significantly decreased in their task performance because they no longer received the feedback about how they were achieving their daily goals to take their medications. Without the feedback, they could not continue the daily process of optimizing their medication taking routines to be adherent, prompt, correct, and consistent in the time of day they take their medications. Similarly, individuals who did not receive the real-time feedback but instead reflected only on visualizations of long-term data were able to improve their behaviors slightly for a short period (approximately three weeks) before they reverted back to their pre-reflection level of performance. However, reflecting on the trends and the large volume of data in long-term view of the data triggered individuals to re-evaluate their subjective ratings of their abilities and awareness of their abilities.

Thus, in both the pilot study and the larger deployment of the dwellSense system, individuals were able to reflect on the information from the sensing systems and find opportunities to become more self-aware of their actions as well as in some cases to even improve their task performance.

In addition to supporting greater awareness and improvements in task performance, the objective sensor data collected by dwellSense can be used to assess how well a task is performed. Support for thesis sub-statement A was found by analyzing how well a heuristic-based approach for automatically rating task performance using sensor data matches with the assessment approach used by a trained occupational therapist (Chapter 7). Heuristics for the automated

sensor-based assessment were developed based on the heuristics used in the PASS, a tool for performance testing used by occupational therapists. Ratings from the sensor-based assessment approach matched well with ratings from the occupational therapist. However, there were differences in the way tasks were performed when the individual was alone and when the individual was being tested by the therapists. dwellSense data provided the basis for understanding and quantifying just how differently individuals performed their tasks when being tested by a therapist. Individuals committed significantly fewer mistakes and were more intentional in their actions when being tested than when they carry out the same task on their own. A sensor-based approach for monitoring how an individual carries out their everyday activities can capture more ecologically-valid data than performance testing, the current most objective standard for assessment.

## **8.2 Contributions**

This work makes contributions in three different disciplines: Human-Computer Interaction/Design, Computer Science, and Health Science.

### **8.2.1 Contributions to HCI and Design**

The research described in this thesis has generated contributions in understanding users, how people interact with their own personal data, how to present data in a way that supports particular behaviors or attitudes, and also a methodology for designing sensing systems that record data with human-mediated meaning.

#### **8.2.1.1 User Reflective Design Process**

The User Reflective Design Process provides a methodology to leverage human insights when designing intelligent sensing systems. Many sensing systems record data about actions of the users. However, the recorded actions themselves may not reveal the meaning of why the user performed those actions. In fact, the meaning of the data is found only in the user's interpretation of the data. For example, a system that tracks physical activity can produce a chart of an individual's step counts over time. There may be some days that the individual was very active and other days that the individual was very inactive. The system cannot explain the reasons for high or low activity by looking only at the step count data. However, by engaging the individual with this log of their step counts, the individual can view the data and generate explanations for why she was active on certain days and inactive on others, for example, whether she was working from home (low activity) or working at the office (high activity), and thus her location is an important factor to explain her behavior. As a result, the designer can discover a new opportunity to add a new data stream (by adding a new sensor) to the system to enhance the ability of the system to reason about the data better and make better predictions of the user's actions. Thus, engaging individuals with their own personal data can lead to insights into how to make sensing systems more intelligent.

This methodology not only applies to the design of personal sensing systems like dwellSense or MemExerciser but also more broadly to any informatics system that collects or generates data that requires human interpretation. For example, analytics systems for business intelligence collect a wealth of data including business transactions, costs, revenue, sales, product features, and marketing schemes. However, ultimately, there is typically a human (an executive, marketing manager, salesperson, or analyst) who views dashboards of data to interpret the meaning and significance of how these data streams connect. The executive may think of a new explanatory factor that the system has not yet considered or find a new relationship between two extant data streams that the system did not yet compute. The User Reflective Design Process facilitates the capturing of the human insights like these so that the system designer can improve ability of the system to understand, reason about, and present the data.

### 8.2.1.2 Information Needs of Stakeholders

Understanding the needs of users is a common first step in many human-computer interaction innovations. This work follows the same path. The stakeholders involved in embedded assessment include the older adults being monitored, their family and other informal caregivers, and their clinicians (including physicians and therapists). Using an initial concept validation study followed up and validated by two long-term deployments, this work has uncovered the information needs of the stakeholders. All stakeholders found long-term data about their task performance because it enabled them to compare the individual's performance over time and identify changes in abilities. Only the individuals themselves wanted recent short-term views of their performance because it would help them feel more confident that they were carrying out their activities well. Family caregivers and occupational therapists wanted the low-level details of how the individual was carrying out each step of the task so that they would know exactly how to intervene and help fix the problem. Physicians, on the other hand, only had the resources to look at a quick high-level view of task performance. Understanding the differences in the information needs is critical for design sensing systems that aim to provide the stakeholder with the critical piece of information. The information needs identified in this work can help drive the design of future systems for the home health systems as well as provide a starting point for a larger class of personal sensing systems that support an individual's abilities to perform an action such as increasing the amount of physical ability or financial responsibility.

### 8.2.1.3 Sensemaking Process for Personal Sensor Data

This thesis provides insights into how people make sense of their own behaviors, as captured over a long term by sensing system. In contrast to diary studies, the data reflected upon is passively and objectively captured. Reflecting on this third-person account of behavior is a relatively new experience for many of the stakeholders. Individuals first identified the anomalies in the data, in other words, the particular data points that looked either different from the others or different from the individual's initial expectations. Individuals tried to find explanations for these anomalies, relying either on their memory, routines, or external memory aids such as a calendar or diary. Explaining anomalies was important because it was critical to know whether or not to excuse them or to acknowledge them as incongruous with their initial expectations. Looking into the low level details of the sensor data such as the timing was helpful for confirming explanations. Breakdowns can occur in any step in the sensemaking process and can provide opportunities for designers to support the sensemaking process as individuals reflect on the record of their own behaviors. As more technologies integrate more sensing and capture of user interactions, there will be more opportunities for users to engage directly with the data collected about themselves and about others. The sensemaking process identified in this work provides the framework for the study of how individuals make sense of new forms of data from future sensing system.

### 8.2.1.4 Real-time versus Long-term Reflection

This thesis investigated and quantified the differences between reflecting on real-time feedback and reflecting on long-term feedback. Real-time feedback about how individuals carried out their daily tasks reinforced the individual's motivation and ability to carry out their tasks at a higher level of adequacy. However, real-time feedback did not increase the accuracy of their self-perceptions of their actions. On the other hand, reflecting on long-term feedback such as a visualization of eight weeks of data did result in individuals adjusting a more accurate awareness of their abilities. The objective long-term account provided the evidence that enabled individuals to report their actions more accurately. After reflecting on the long-term feedback, individuals improved their task performance only temporarily whereas the individuals who had the real-time feedback continued at their increased level of task performance.

Designers of systems designed to provide feedback to users about their actions may have to choose between prioritizing real-time feedback or long-term feedback. Certain systems may be constrained in screen real estate (such as the notification area of a mobile phone or an ambient display with only a limited number of output dimensions), and thus

only certain pieces of information can be shown at one time. From a behavioral perspective, some systems might be designed to optimize behavior change whereas others may be designed more for changes in awareness. Differences in the way real-time and long-term feedback are presented also can influence the type of engagement the system has with the user. Frequent engagement with real-time feedback requires designing the system in a way that does not require or draw an undue amount of attention from the user. Motivating individuals to reflect on long-term feedback may also require the system to help the individual identify a trend in the visualization if one exists. If these challenges can be addressed, it is likely that providing long-term feedback first to reorient the individual and then providing the real-time feedback to support the individual's abilities to support their goals for changing their behaviors would result in the optimum combination of feedback to increase self-awareness, engagement, and behavior change.

## **8.2.2 Contributions to Computer Science**

The research in this thesis includes technical contributions to the fields of computer science and engineering.

### **8.2.2.1 Task-based Sensing System for Instrumental Activities of Daily Living**

This thesis describes the design, implementation, and evaluation of *dwellSense*, a task-based embedded assessment system that can recognize, record, and assess how well Instrumental Activities of Daily Living are performed by individuals in their own homes. Previous smart home systems use more generic sensors such as motion sensors and video cameras to monitor overall activity patterns. *dwellSense* uses a suite of sensors made from augmenting that objects normally used in the home to track how well particular tasks are performed. The Medication Monitor sensor tracks the steps in taking medication. The Telephone Tracker sensor tracks the numbers that are dialed on the phone. The Coffee Chronieler sensor tracks the different steps in making a pot of coffee. *dwellSense* (version 2.0) also added motion sensing to provide an important data stream for knowing when the individual was home or not. The *dwellSense* architecture uses a Zigbee-based wireless network for the sensors to transmit their data in real time to a local PC. The sensor data are automatically uploaded to a remote server where the sensor events are processed and the task performance is rated. *dwellSense* also generates web-based visualizations of the sensor data and higher-level task performance ratings. *dwellSense* was deployed for a total of 156 participant-months, including 36 participant-months of early pilot testing, and its common points of failure such as power outages and computer crashes provides future systems designers with opportunities to improve the system. The *dwellSense* architecture is extensible for the addition of other sensors and provides a tested proof-of-concept for future task-based embedded assessment and home monitoring systems.

### **8.2.2.2 Rule-based Assessment that Matches Human Ratings**

An important part of this thesis is to show that an automated approach can perform nearly as well as a trained clinician for assessing functional abilities. The *dwellSense* system uses rules manually derived from the PASS tool used by occupational therapists to assess the functional abilities. Using the simply derived rules, *dwellSense* assigns a rating to each step of a task that it can monitor and, following the convention in the PASS tool, uses the minimum rating across all the steps as the overall rating for the task. With this automated rule-based approach, *dwellSense* was able to generate ratings that have a high correlation with the therapist's ratings in the phone-use and medication-taking tasks. The sensors in the Coffee Chronieler sensor was unable to capture enough of the steps in coffee making to produce ratings reliably correlated with the ratings from the therapist. Nevertheless, in the medication-taking and phone use tasks, the sensors were able to capture the most important steps of the task that accounted for the majority of the variation in the ratings. Through the building and testing of the *dwellSense* system, this thesis demonstrates that it is possible use an automated approach based only on the sensor data to accurately assess how well individuals carry out Instrumental Activities of Daily Living.

### 8.2.3 Contributions to Health Sciences

In addition to contributions to human-computer interaction and computer science, this work also makes contributions to the health sciences, in particular, demonstrating how sensor data can be used to support an individual's self-awareness and the individual's ability to improve their behaviors. Furthermore, the dwellSense system enables researchers and clinicians to see how individuals function in their own home as well as to measure the effect of being watched by a therapist during performance testing.

#### 8.2.3.1 Supporting Self-Awareness Using Objective Data

The two long-term sensor deployment studies in this thesis provide the concrete evidence for how reflecting on embedded assessment data can lead to a more accurate self-awareness of an individual's abilities. The objectivity of the sensor data provides a trustable third-person account of an individual's task performance. The individual can compare her self-perception of her actions with the actions recorded in the sensor data. Often, the immediate reaction when reflecting on the sensor data includes mild surprise, as the individual is confronted with episodes of poor performance (such as missing or late medications), which contrasts with what they had originally thought about themselves. Individual typically trusted the account recorded by the sensors and as a result re-oriented their self-awareness to match what they saw in the sensor data. According to the Trans-Theoretical Model of Behavior Change (TTM), one mechanism that motivates individuals to be more self-aware and to change their behaviors is "dramatic relief." The objective sensor data are shown in dramatic relief to the individual's self-perceived level of task performance, particularly in the case when individuals initially believed they were doing a lot better than they actually were. After reflecting on the data, they were asked to report on different aspects of their behaviors such as missed dialed telephone calls or medications taken late, and these reports were more accurate after reflecting on the sensor data. Older adults themselves were able to reflect on the objective sensor data from dwellSense to support a more accurate self-awareness of their abilities.

#### 8.2.3.2 Supporting Behavior Change with Real-Time Feedback

Health behavior change has long been an object of study. This thesis provides evidence that frequent, real-time feedback helps individuals to improve the level of task adequacy when performing their tasks. Individuals who received the real-time feedback display in their homes significantly increased in the adherence, promptness, correctness, and consistency in the time of day they took their medications. The immediate feedback from the display provided individuals with feedback about how well they were meeting their daily sub-goal to take their medications well. The immediate positive feedback reinforced the individual's motivation to continue performing their tasks well. The immediate negative feedback highlighted opportunities for individuals to correct the problems the next time they carried out the task. Thus, frequent feedback on sub-goals helps individuals optimize their attainment of their overall long-term goal to take their medications correctly and consistently. This thesis provides concrete evidence of how ipsative (comparing to self, not with others) personalized feedback allows individuals to compare their current state with their desired goal state. Future interventions for health behavior change in other domains such as smoking cessation, alcohol use, or nutrition can adopt a similar strategy to foster behavior change by combining automatic assessment with frequent feedback.

#### 8.2.3.3 System for Tracking Typical Behaviors

The dwellSense system introduced in this thesis can monitor how individuals take their medications, use the phone, and make coffee using a coffeemaker in their own homes. These behaviors were hitherto unavailable for observation by researchers and clinicians. Existing smart pill bottle caps like the Vitality Glow Cap (==) can monitor how medications are taken but cannot measure time on task or whether the individual has actually followed through with taking pills out of the bottle. Another augmented pillbox MedTracker (Hayes et al., 2008) is static and is required to be plugged in, which makes it much less portable than the Medication Monitor used in the dwellSense system. The data is



collected passively in the home of individuals as they carry out their daily tasks, and thus the activities monitored by the sensors are exactly the activities that individuals are performing when on their own. Another key capability of dwellSense is that it can be deployed for an extended period of time so that it can record enough behaviors for patterns of typical behavior to emerge. In fact, dwellSense was deployed for over 18 months in a pilot study and for 10 months in a study with 12 individuals. Each individual's pattern of typical behaviors could be observed in the sensor data. The quantitative nature of the sensor data also allows for comparisons across individuals. Thus, dwellSense using sensor-based approach that can passively, unobtrusively, and objectively capture and create a log task performance, enables clinicians and researchers a window in the typical everyday behaviors of individuals.

#### 8.2.3.4 Quantified Testing Effect of Performance Testing

As a result of being able to track how an individual typically behaves in their own home, the data collected by the dwellSense system can also be used to compare typical behaviors with behaviors under particular circumstances such as when being evaluated by a therapist administering performance testing. The testing effect or reactivity of performance testing (Holm & Rogers, 1999) in the home by a visiting therapist can be identified and quantified using the dwellSense data. There were significant differences between the way the medication-taking and coffee-making tasks are typically performed when individuals are on their own and the way these tasks are performed when individuals are being tested. Individuals tended to make fewer mistakes but also skipped some steps they normally would have performed because tasks were often simulated (when the individual is asked to pretend to do the task instead of actually doing it as part of their normal routine) during performance testing. With this understanding, therapists who rely on performance testing can develop ways of adjusting the performance testing scores to match typical performance. With testing effects so strong in simulated tasks, therapists can also consider whether simulated tasks provide the correct target for observation in performance testing.

### 8.3 Limitations and Future Work

Evaluating embedded assessment technologies can be challenging for a number of reasons including the lengthy time span required for evaluations and the unpredictability of how individuals change over time. This section describes some of the limitations of this work as well as future directions that address some of these challenges and next steps for contributing to understanding how to design systems for embedded assessment.

The evaluations conducted in this thesis included a relatively small number of participants—two in the pilot study and twelve in the larger deployment. The small sample size makes it difficult to use statistical methods to find significant differences. To address this challenge, the study design used a within-subjects design with multiple repeated measures over many months. Individuals were compared with their own baselines so that individual differences would not unduly influence introduce confounding factors in the analysis. Another consequence of a small sample size is that having a true control group (that received absolutely no intervention during the entire evaluation period) would have reduced the treatment group sizes. Thus, in the latter deployment with twelve individuals, individuals were assigned to two groups, a real-time display group and the long-term reflection group. This design allowed the long-term reflection group to act as a strictly no-intervention control for the real-time display group for the first four months of the study. However, after month four, the long-term reflection group received their intervention, making them another treatment group. When comparing the two groups near the end of the study, for example, when the real-time display was removed, the long-term group no longer can be used as a strict control group for comparison because they have undergone changes as a result of the long-term reflection intervention. Another consequence of having no strict control group was that the normal trajectory of declines or changes in functional abilities could not be directly observed. Both groups in the deployment included an intervention that was designed to increase their awareness and find opportunities to improve their task performance. Thus, the trajectory that individuals might have normally followed was likely disrupted by the interventions. It is therefore difficult to extract from the results of this work the natural trajectory of functional changes

that individuals experience as they age in their own homes. However, future longer-term deployments of the technology without the interventions can be used to investigate these normal trajectories associated with aging. Furthermore, with the promising results from the evaluations in this thesis, future evaluations of embedded assessment technology and reflection on the data it produces should include a greater number of participants so that a strict control group can be used for comparisons.

Another limitation from a more technical perspective is that the dwellSense sensors, as well designed as they are, cannot always capture how tasks are performed if the individual decides to perform the task in an unsupported way. For example, individuals may decide to take their pills directly from the pill bottles instead of the pillbox when trying to finish off a bottle of medications. Even though individuals have taken their medications in that instance, the Medication Monitor pillbox sensor was not opened and thus the system records a missing medication taking episode. Even though individuals were told by researchers to take their medications using the pillbox (because it not only would help the study but also help them keep track of which pills they have taken), they would occasionally perform the pill-taking task in a way that cannot be tracked by the sensors. These deviations were more of the exception than the norm. Nonetheless, this highlights the limitations in the sensing technology. Sensing embedded in the home on particular objects to track how tasks are performed can unobtrusively collect objective and frequent data on how a task is performed. The sensor-based data are arguably much closer to the actual truth about what the individual is doing than either self-reports, caregiver-reports, or performance testing, but it still is subject to some degree of error as individuals perform tasks in alternative ways that the sensors cannot monitor. Future work to address this limitation can include developing and integrating new sensing modalities to accommodate alternate ways of carrying out a task.

The extent of the manually collected information was also limited in the studies in the presented in this work. Measures of awareness were collected using questionnaires only in the first eight months of the study, which was the originally planned length of the deployment. As a result, aside from measures of behavior change as captured by the sensors, the effects of removing the real-time display in ninth month of the study on the individual's awareness of their task performance was not measured. There were also limitations in the frequency of performance testing conducted during the deployment. Performance testing was administered three times for each participant. The activities of the study included visits from researchers to maintain the system and questionnaires to measure awareness and ratings of abilities. These visits grew in length to take approximately one hour, which was perceived by some participants as burdensome to their busy schedules. More frequent performance testing would have likely overburdened the study participants, particularly because the tasks assessed were the same each time. Moreover, the three performance testing sessions were administered within a four-month time span due to logistical issues with the IRB and contracts with the therapist, which meant that can only capture the changes only in the latter half of the study.

Measures of self-efficacy (Appendix H) were also included in the questionnaires administered monthly to each individual. However, all individuals tended to rate themselves as having high self-efficacy, even at the beginning of the study. This is not surprising given the community-dwelling, living-alone sample of individuals in the study. As a result, measures of self-efficacy hit a ceiling effect and thus only decreases in self-efficacy could have been detected. The scales used were also fairly coarse with only three levels of ratings (very confident, somewhat confident, not confident) and may not have been sensitive enough to measure the effect of reflection on embedded assessment data on self-efficacy. Future evaluations of embedded assessment would include more focused and sensitive questionnaires about awareness and self-efficacy as well as include distribute performance testing through the study so that it can capture the maximum range of abilities.

This thesis focused mostly on the usefulness of embedded assessment data for one particular stakeholder: older adults. Future opportunities include investigating how the data can be shared with caregivers and clinicians, as well as how these additional stakeholders would use the data to improve the way they provide care to the individual. Investigating how this new stream of objective sensor data can change the way individuals communicate with caregivers and doctors

is a fruitful area for investigation. Utilizing embedded assessment data may require a change the way information is generated and consumed in the clinical setting. Individuals may want to have the opportunity to review, annotate, or edit logs of their actions before sharing them with other stakeholders, and thus issues of privacy, control, and data integrity need to be explored. Moreover, this thesis also generated and iterated on some visualizations of task performance data. This thesis does not claim that the particular visualization designs are the optimal design but rather that they were useful for displaying information that enabled individuals to discuss their task performance. Further work to refine the visualizations for feedback to older adults, their caregivers, and clinicians is necessary to optimize the intuitiveness and clarity of the data visualizations. Addressing these issue can make it much more likely that new data streams with the potential for early signals for changes in health to be adopted into the home and clinical practice.

With the effectiveness of monitoring medication-taking, phone use, and coffee-making tasks established by this thesis, future applications of sensing technology in the home to track functional or cognitive decline can expand the suite of sensed tasks. For example, tracking an individual's sleep patterns can be a strong factor in how an individual performs other tasks. Changes in sleep patterns can also indicate changes in health status. Low-cost sensors for tracking the quality of sleep (such as the SleepCycle iPhone app) can be integrated with streams of embedded assessment data to find associations between sleep and other activities. Given the success of sensing how individuals interact with the electronic devices such as the phone and coffeemaker, future work can include other simple interactions with technology such as how individuals operate the television remote control, use a microwave, or measuring how correctly individuals learn to use a new digital device. Interactions with complex digital devices, particularly when learning how to use new devices, require a fair amount of cognitive abilities. Tracking how these abilities change over time can potentially provide early indicators for decline.

Inspired by the seminal work in embedded assessment (Morris et al., 2003; Morris et al., 2005), this thesis describes the design, development, and extended evaluation of a task-based embedded assessment system capable of monitoring and assessing how Instrumental Activities of Daily Living are carried out. The system and the information it collects is useful for reflecting on an individual's actions and providing opportunities to be more self-aware of how tasks are actually carried out and to maintain the ability to age in place.

## 8.4 References

- Hayes, T. L., Hunt, J. M., Adami, A., & Kaye, J. A. (2006). An electronic pillbox for continuous monitoring of medication adherence. *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE* (pp. 6400–6403). Retrieved from [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=4463275](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4463275)
- Holm, M., & Rogers, J. (1999). Performance assessment of self-care skills. *Assessment in occupational therapy mental health: an integrative approach. Assessments in occupational therapy mental health: an integrative approach* (pp. 117–124). Thorofare, NJ: B. Hemphill-Pearson.
- Morris, M., Lundell, J., Dishman, E., & Needham, B. (2003). New perspectives on ubiquitous computing from ethnographic study of elders with cognitive decline. *UbiComp 2003: Ubiquitous Computing* (pp. 227–242). Retrieved from <http://www.springerlink.com/index/2dmxcejn5c4hdhra.pdf>
- Morris, M., Intille, S. S., & Beaudin, J. S. (2005). Embedded assessment: Overcoming barriers to early detection with pervasive computing. *Pervasive Computing*, 333–346.
- Weiser, M. (1995). The computer for the 21st century. *Scientific American*, 272(3), 78–89.

