USABILITY ENGINEERING FRAMEWORK FOR PERSUASIVE MOBILE HEALTH APPS TO EFFECTIVELY INFLUENCE DIETARY DECISIONS OF OLDER ADULTS

by

Wen-Yu Chao

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering
West Lafayette, Indiana
December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zachary Hass, Co-Chair
Schools of Industrial Engineering and Nursing

Dr. Mark R. Lehto, Co-Chair
School of Industrial Engineering

Dr. Brandon J. Pitts
School of Industrial Engineering

Dr. Sandra S. Liu
Department of Public Health

Approved by:
Dr. Abhijit Deshmukh

Dedicated to My Beloved Family

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to thank my advisors, Professor Zachary Hass and Professor Mark R. Lehto, whose continuously support was invaluable in guide me through all the hard times I had during my Ph.D. journey. Your expertise and patience enlightening me to formulate right research questions and methodologies for this research work. And I appreciate my dissertation advisory committees, Professor Brandon J. Pitts and Professor Sandra S. Liu, your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to acknowledge my colleagues Dr. Le Zhang, Dr. QingXing Qu, Dr. Ya-Hsing Hung for their wonderful collaboration and inspiration for this topic. And also, I acknowledge the great help for my dissertation studies from Laura Warner and Mike Eldridge at Purdue Discovery Learning and Resource Center; Christine Rearick from Purdue Nursing Center for Family Health; Andrea Fisher and Bill Glick from Tippecanoe Senior Center; my participants referral, Mark Wuolle et al. and participants of this study; Faith West community center, Purdue Graduate Student Government, Purdue Writing Lab, and Purdue IE department & Purdue Nursing Department. I could not complete the studies in this dissertation without your kindly support. I also would like to give special thanks to Springer, Dr. Robert Proctor, and Dr. Ya-Hsing Hung for grant me the copyright to re-produce part of our previous academic works.

I would also like to thank my teachers and mentors during this Ph.D. journey, Dr. Vincent Duffy, Dr. Robert Proctor, Dr. Paul Parsons, Dr. Gregory Arling, and Dr. Yei-Chin Chao, for their valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my dissertation.

In addition, I would like to thank my parents and my sister for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed this dissertation without the support of my friends, Sally Wang, Trista Chen, Sabrina Huang, Kayle Tung who provided great mental support to rest my mind and happy distractions to enjoy my life outside of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Introduction: Mobile health applications (mHealth apps) have the potential to assist patients in adhering to their physician's advice in chronic disease management through the use of persuasive nudge. However, systematically developing the persuasive features of a mHealth app for the major user demographic of older adults is challenging. The current usability engineering framework could ensure the user-friendliness of the app but not the persuasiveness. It is necessary to extend the current framework with appropriate measures to better understand the effectiveness of persuasive design elements in an iterative design process.

Methods: A pilot design project was run, a persuasive mHealth app for dietary management was developed using the user-centered design approach (persona, use scenario, task analysis, and cognitive walkthrough), the pilot testing result showed high potential of technology acceptance of older adults. To further evaluate persuasiveness, a food choice experimental protocol and human decision performance metrics based on Signal Detection Theory (SDT) were proposed. A mixed-methods, full factorial user testing study was conducted with twenty older adults aged over 60 and twenty students age 18-35. Critical persuasive User Interface (UI) design variables included decision paradigm (digital nudge), nutrition information format (information nudge), and the system default pre-selection (default nudge). The proposed SDT metrics to evaluate persuasiveness were then compared with confusion matrix metrics which are frequently used to validate system decision-making performance. The relationship between the human performance, subjective workload, and perceived usability of the proposed mHealth app was also investigated.

Results: The 'Two Alternative Forced Choice' layout significantly increased the d-prime and accuracy (persuasiveness), the system default pre-selection decreased persuasiveness. The interpretative FSA Nutri-scores label reduced time of response and workload, and increased perceived ease of use, perceived ease of learning, and satisfaction. Among older adults, results differed by age, computer proficiency, and health literacy.

Conclusion: The findings of this study imply the proposed framework is a valid persuasive design research approach. And digital nudge is an effective persuasive design for mHealth app, while

default nudge may give rise to negative effects. A generalized human-centered digital nudge design framework along with ageing-centered guidelines were suggested for the similar research and design projects for persuasive technology performed in the future.

1. INTRODUCTION

Influencing patients' decisions for their adherence to a prescribed regimen such as medication, diet, exercise, alcohol and smoking cessation is critical in chronic disease management (Dunbar-Jacob et al., 2000). For example, type II diabetes is strongly associated with modifiable lifestyle choices such as overeating and low level of physical activities.

Several strategies for behavioral intervention were found useful such as motivation and social support. For example, Stephens et al. (2010) found the positive effect of social control if the spouse encourage the patient to adhere dietary recommendation guidelines with the cheering tone and positive influence attempts (Stephens, Rook, Franks, Khan, & Iida, 2010). As to self-management for many living alone cases, another strategy to be practically applied to induce health behavior change is "Nudge", which is a decision intervention by designing the choice architecture to influence an individual's decisions. An example of influencing individual's dietary behavior by nudge strategy would be to change the size of the available food plate at a food buffet restaurant to influence an individual's decision on portion size (Marteau, Ogilvie, Roland, Suhrcke, & Kelly, 2011).

As the usage of smartphones has grown, mobile Health applications (mHealth apps; "mHealth" is an abbreviation of the term "Mobile Health".) have become a great resource in patients' daily life to implement health behavior change techniques, such as social support, self-management, gamification, etc. For example, Carter et al. (2013) reported the efficacy of adopting a self-management mHealth app to support weight loss. In academics, more and more researchers and practitioners notice the potential and developed the mHealth apps to facilitate the behavioral interventions, for example, online dietary assessment tools which moves the 24-h dietary recall questionnaire to the digital sphere for tracking patient's eating behavior.

However, this kind of app development usually lacks the cross-disciplinary collaboration with the researcher in Human-Computers Interactions (HCI) domain area and considerations of usability engineering (Hingle & Patrick, 2016). As a result, most of the dietary assessment app are not tailored to the needs of the potential user group and/or have a lower user retention rate (J. Cho, 2016). For example, Cho (2016) investigated the post-adoption behavior of 343 smartphone health apps used in Korea and found low perceived usefulness of the app significantly reduces the intention of continued usage.

User Interface (UI) design is the key to success in an mHealth app since the human-computer interactions behavior is guided and User Experience (UX) is defined by UI design. In this context, researchers suggest the user-centered design of the mHealth apps to ensure the usability, of which the key aspects including the effectiveness, efficiency, and satisfaction based on ISO 9241-11 definition and the learnability based on pioneer research of usability engineering (Abran, Khelifi, Suryn, & Seffah, 2003; Dix, Finlay, Abowd, & Beale, 2003; Nielsen, 1994). Brown et al. (2013) and Schnall et. Al (2016) have further adapted the user-centered design and evaluation framework to the health information technologies design context for mHealth app (Brown, Yen, Rojas, & Schnall, 2013; Schnall, Cho, & Liu, 2018; Schnall et al., 2016).

However, there are few discussions about the other important aspects of the UI design of mHealth apps. First of all, the inclusive design of mHealth apps for older adults. Secondly, the UI design effect on the mHealth app as persuasive technology for health behavior change.

Older adult patients are one of the key potential user demographics of mHealth apps, since many chronic diseases, for example, type II diabetes, are prevalent in older adults. But currently, the UI designs of most mHealth apps seldom consider the needs and limitations of this specific user group during the product development stages. As a result, usability pitfalls are often found. For example, Whitlock & McLaughlin (2012) reported the usability problem of a blood glucose tracking app for older adults with limited numeracy to interpret the chart and suggest a decision aid (Whitlock & McLaughlin, 2012); Isakovic et al. (2016) also reported the usability issues of the European Union (EU) developed diabetes monitoring app for older adults (M. Isaković, Sedlar, Volk, & Bešter, 2016). Although later Wildenbos et al. (2015; 2018) suggest the user-centered evaluation framework based on the literature review to take aging barriers including physical ability, perceptions, cognitions, and motivations into design consideration. However, there’s still a research gap of empirical studies of the systematic design selections. Since in the real world, older adult patients are a unique group of people with a wide range of computer proficiency and technology acceptance, which add the complexity of the problem. In this context, how best to design the User Interface (UI) of mHealth apps to nudge older adults towards better health behavior remains undetermined.

Once the mHealth app is used as a persuasive technology, UI design could be the key to influence users' decisions, thereby altering an individual's health behavior. As Weinmann et al. (2016) defined, digital nudge is "the use of user-interface design elements to guide people's behavior in digital choice environments." (Weinmann, Schneider, & Brocke, 2016). However, limited studies reported the successful implementation of nudge approaches in mHealth apps, especially digital nudge. A possible reason is the research gap of measuring persuasiveness during the product development stage to develop persuasive UI design elements and systematically arrange the elements for building the digital choice architecture.

To narrow the research gap, this thesis proposed a human factors evaluation method for persuasiveness and integrated with Nielsen's usability engineering framework. The objective of this research is to extend the existing framework to systematically design the usability and persuasiveness in technologies for older adults. A persuasive dietary management app was developed based on the extended usability engineering framework and a user testing study with 40 subjects (20 older adults age above 60 and 20 students) was conducted to validate the proposed persuasive design research method. Three nudge approaches were selected and translated to the persuasive UI design elements, including decision paradigm (digital nudge), nutrition information format (information nudge), and the system default pre-selection (default nudge). Digital nudge approach is found the most effective nudge approach, while no significant effects of information nudge were found and negative effects of default nudge on discriminability and accuracy. The proposed mixed-methods persuasive design research method based on Signal Detection Theory (SDT) is discussed by comparing to confusion matrix metrics and examining the relationships between the measurements.

The structure of this dissertation is organized as follows: Chapter one introduces the background and identifies the specific research problem and research objectives. Key literature which serves as theoretical basis was also reviewed to illustrate the basic research structure. Chapter two narrows the scope to the persuasive mHealth app and reviews the related literature. Chapter three proposes an extended usability engineering framework for persuasive mHealth apps and presents the practice of the framework with a pilot design project of a diet management app. Chapter four proposes the research framework for evaluating the proposed app. Chapter five describes the research method in detail. Chapter six presents the study part one results about perceived usability and subjective workload of the proposed app. Chapter seven discusses the study part two results about the human performance and subjective workload of the food choice experiment for persuasiveness evaluation. Chapter eight discusses the generalized design framework and the measurement science of persuasive technology. Finally, this work concludes with final remarks around research limitations and the future research directions.

1.1 Significance of the Problem

Behavioral risk factors, such as tobacco use, poor diet and physical inactivity, and alcohol consumption, contribute to the leading causes of death in the U.S. including heart disease, cancer, diabetes, and stroke. Mokdad et al. (2004) attributed 15.2% of total US deaths in 2000 to poor diet and physical inactivity, which is even higher than that caused by motor vehicle crashes (Mokdad, Marks, Stroup, & Gerberding, 2004). And even as year have passed, Keeney (2008) and Putzer (2015) still found high percentage of premature mortality is associated with poor personal decisions toward modifiable lifestyle behaviors of smoking, diet, exercise, drinking alcohol, and illicit drug use. (Keeney, 2008; Putzer, Gavin; Jaramillo, 2015, 2017)

Behavior change, including dietary behavior change, is often difficult, but creating a properly designed smart system to support that change will remove some of that difficulty. Healthy eating is essential to prevention and self-management of diet-related chronic conditions such as Type II diabetes. Type II diabetes is prevalent in older adults in the US, and lifestyle change was considered as a critical factor to limit the progression of the disease. For example, Knowler et al. (2002) found the effect of lifestyle modification to limit the progression of type II diabetes is significantly stronger than medications (Knowler et al., 2002). Tuomilehto et al. (2001) have also found lifestyle change contributes to prevent the progression from pre-diabetes into diabetes. Eating is a major Activity of Daily Life (ADL) and diet is one of the most important lifestyle-related risk factors of type II diabetes (Tuomilehto et al., 2001). Nevertheless, it requires high level of self-regulation and adequate health and nutrition knowledge for an individual to change his/her dietary behavior (McLaughlin, Whitlock, Lester, & McGraw, 2017). For example, Klein and Meininger (2004) found there's a knowledge gap for older adults trying to conform to healthy eating guidelines, as they were reportedly using inaccurate heuristics of food selection and serving size for dietary control (Klein & Meininger, 2004).

As more and more information is moving to the digital sphere and systems are getting smarter with a broader knowledge base, two concerns and research gaps motivates the proposed studies in this dissertation:

1. The current UI/UX design of health technologies seldom considers the potential major user groups, for example, the dietary management apps which are designed without specifically considering the needs of older adult patients with type II diabetes and their caregivers. However, there's a research gap of empirical studies of ageing-center design for mHealth apps, so it's still unclear if the benefit would pay off the cost of designing for the specific user group.
2. As the health information technologies integrating with the Artificial Intelligence (AI) algorithms and getting smarter, it's potential to be used as the persuasive technologies for health behavior change. The UI/UX design of this kind of innovative systems would create a unique use experience and be expected to influence user's choices and behavior. However, so far, there's a research gap of practical approaches to facilitate the theoretical ideas of persuasive technologies.

However, the effectiveness of the persuasive technology remains undefined. When health information technologies are used for health behavior change, how to design the usability in the product and ensure the effectiveness of the persuasive features for health behavior change would be the core research problem to answer in this dissertation.

1.2 Theoretical Basis

1.2.1 Health Behavior Change Theories and Nudge

In order to plan a behavioral intervention to promote health, there is some applicable theories rooted from Health Belief Model (HBM) in the 1950s that helps to explain how an individual changes health-related behavior. These included Theories of Reasoned Action (TRA), Theory of Planned Behavior (TPB), and Social Cognitive Theory (SCT). According to the HBM, an individual's decision of taking the health actions would be based on the perceptions of their own risk of certain health condition or illness (perceived susceptibility), their own evaluation of the seriousness of the illness (perceived severity), and the balance between perceived benefits and the perceived barriers of the action. The decision could be triggered by either internal cues such as the symptom of diseases and external cues such as the health promotion message by the newspaper or the advertisement; or hearing that a friend has a certain disease. Health behavior change strategies were developed based on the above theories. Techniques which are frequently used in healthy eating interventions include providing information about the links between behaviors and health, physician's approval or disapproval of certain behaviors, and information on consequences; providing information for goal setting and barriers identification; encouraging specific goal setting, self-monitoring and review; providing instruction and general encouragement; setting graded tasks; modeling/demonstrating desired behavior (Abraham & Michie, 2008).

Nudge

Traditionally, the decision science is based on the assumption of classic economics that the decision agents are rational, who would have a stable preference and make decision to maximize their utilities subject to the constraints. However, behavioral decision scientists explained human decision maker seldom uses the optimization strategy but make a decision with the limitations to assess problem rationally, such as bounded rationality (Gigerenzer & Selten, 2002; Kahneman, 2003), heuristics (Gilovich, Griffin, & Kahneman, 2002; Tversky & Kahneman, 1974), and priming (Evans & Stanovich, 2013). The reason is because human have limited capability of evaluating the alternatives rationally and have cognitive and behavioral biases (M. R. Lehto, Nah, & Yi, 2012; Proctor & Zandt, 2018).

Thaler and Sunstein (2008) integrated the previous findings of behavioral decision theories to propose "Nudge" as the behavior change theory and make an influence to economics by changing the way people make a decision. They defined the term "Nudge" as the choice architecture surrounding the human decision-making behavior which could softly paternalize human behavior rather than strictly restricting the possible choices. They suggested that "if a particular unfortunate behavioral or decision-making pattern is the result of cognitive boundaries, biases, or habits, this pattern may be "nudged" toward a better option by integrating insights about the very same kind of boundaries, biases, and habits into the choice architecture surrounding the behavior" (Thaler & Sunstein, 2008).

Nudge has been a prominent topic since it changes the belief of paternalism in policy making and public health. In 2011, Kahneman (2011) published the popular science book, "Think Fast and Slow", which further summarized behavioral decision theories with the dual processing system theory about how people select responses using the heuristics system (automatic mind) or the deliberation system (reflective mind) to describe the automatic mechanisms of human decision-making (Kahneman, 2011). This book later was used as the theoretical basis for behavioral economists to design and implement "Nudge" approaches, for example, "Opt-in/Opt-out" nudge, which considered that most people employ the automatic mind and preferred the default option, so policy makers set the "Opt-in" as the default option for the social welfare plan such as retirement saving funds.

Persuasive Technologies and Digital Nudge

Running parallel to these developments in Behavioral Economics was the rise of personal computers and the related study of human-computer interactions. In 2003, Fogg proposed the idea of persuasive technology, suggesting that computing products could create a new type of interaction, becoming a source of motivation and persuasion. In order to achieve the desired result it's essential to study what elements of the design motivate or persuade people when they are interacting with the computing product (B. J. Fogg, 2002). Weinmann et al. (2016) extended the idea of persuasive technology calling it digital nudge, defined as "the use of user-interface design elements to guide people's behavior in digital choice environments." (Weinmann et al., 2016).

1.2.2 Usability Engineering Framework

What is Usability?

Based on the definition of ISO 9241-11-2018 (2018), usability is “the extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.” (International Organization for Standardization [ISO], 2018) However, ISO 9241-11 considered the usability in the dimensions of efficiency, effectiveness, and satisfaction only. Researchers including Dix et al. (2003), Abran et al.(2003), and Nielsen (2004) suggest “Learnability” is also another key dimension (Abran et al., 2003; Dix et al., 2003; Nielsen, 2004).

By Neilsen’s further definition (2004), usability could be considered as the quality attributes for the user interface to measure how well the user can use the functionality of the information system. It is not a single scale but multiple attributes and traditionally it should be considered in 5 dimensions: Learnability, Efficiency, Memorability, Errors, and Satisfaction (Nielsen, 2004). Learnability emphasize the first-time learning experience and it could be measured by perceived ease of learning with the subjective 5-point or 7-point Likert scale; efficiency indicates the ease of use once the task is learned, which could be measure by the perceived ease of use, or the task performance.

To evaluate the perceived usability, Lewis (1995) proposed 2 questionnaires based on research experience in IBM to collect subjective measurements to evaluate the perceived usability for computer system (Lewis, 1995). The first one is After-Scenario Questionnaire, which is a 3-item questionnaire to be used after the experimental scenario; the second one is Computer System Usability Questionnaire (CUSQ) used for the whole system. Another widely used questionnaire is Brooke’s System Usability Scale (SUS), which is a short version questionnaire with 10 questions and a standardize scoring system (Brooke, 1996). However, Lewis’ and Brooke’s questionnaires adopted the ISO 9241I-11 definition. So, for testing the learnability, instead of adopting Lewis and Brooke’s questionnaire to measure the perceived usability, Lund’s USE questionnaire could be used (as it is in this dissertation), since it is the only questionnaire considering the “perceived ease to learn” concept among the above questionnaires (Lund, 2001).

With the development of automated testing technology, more and more researchers also adopted the automated usability testing approach, such as remote user testing, or eye tracking.

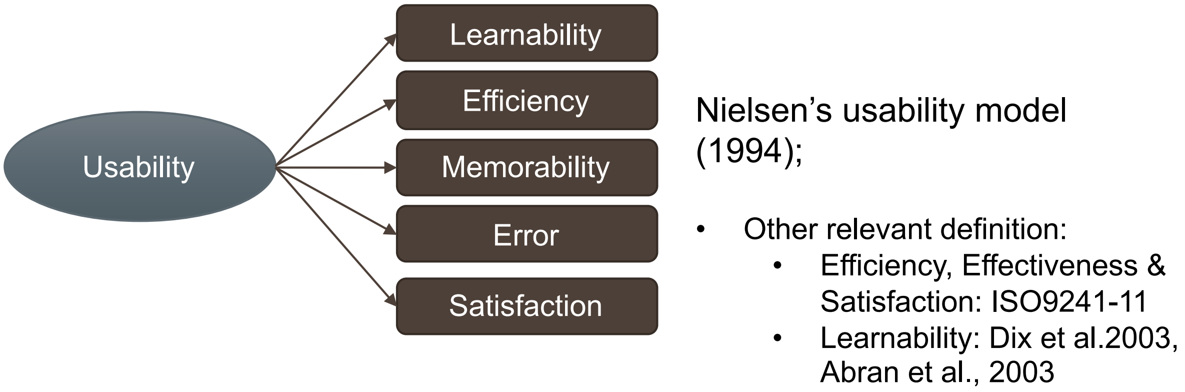

Usability
Learnability
Efficiency
Memorability
Error
Satisfaction
Nielsen’s usability model (1994);
• Other relevant definition:
• Efficiency, Effectiveness & Satisfaction: ISO9241-11
• Learnability: Dix et al.2003, Abran et al., 2003

Figure 1. Neilsen’s usability model

Usability Engineering Lifecycle

A key concept of Nielsen (1993)’s usability engineering framework is: ensuring usability are activities of continuous design improvement through the whole project development process, but not just a one-time visit of the pass/fail test for the final product quality. So, Nielsen has suggested 11 stages of usability engineering lifecycle, which could integrate with the early software product development process. The usability engineering life cycle starts with knowing about the users; conducting the competitive analysis with similar products; setting the usability goals in the financial viewpoint; conducting the parallel design or participatory design; applying guidelines and heuristics evaluation; prototyping; empirical user testing; iterative design; collecting the feedbacks when it’s online. It is not necessary to carry out all the proposed stages, as the typical software project development lifecycle has been shortened for the fast-changing environment (See the historical background of software engineering framework in section 2.3.3.). Currently, several stages of Nielsen’s usability engineering lifecycle have been combined and the methodologies have been integrated as the user-centered design framework.

User-Centered Design Framework

User-centered design (UCD) is a design philosophy to put users at the center of all design decisions during an iterative design process. The idea could be implemented by continuously probing the users' needs and modifying the design based on users' physical and psychological capabilities and recognized individual differences by their demographic. The tools in the design process include user research, prototyping, and user testing (Chao, Qu, Zhang, & Duffy, 2017; Norman & Draper, 1986).

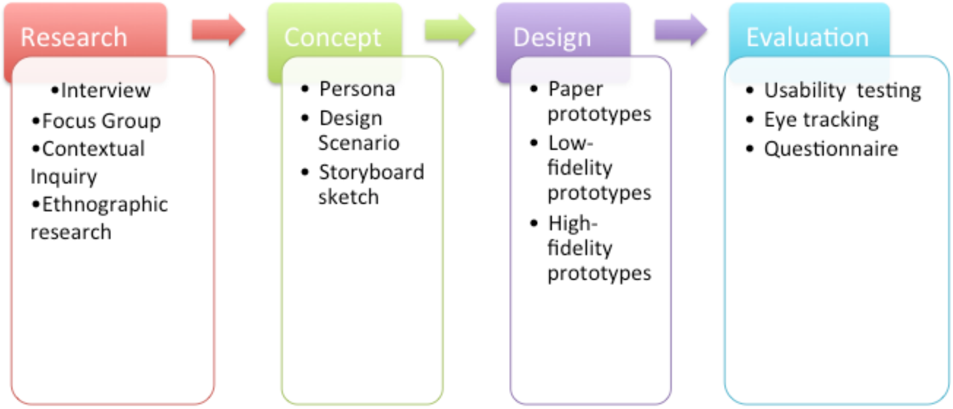


Figure 2. User-Centered Design process and related methods.

Usability Heuristics

Many researchers have generalized from the previous research findings about the characteristics of the usable interfaces as usability heuristics. The usability heuristics are rules of thumbs for design and also could be used as the rubrics of Heuristics Evaluation (HE), a usability evaluation method based on expert reviews. Human factors experts evaluate the usability of a product or an interface by comparing it with the given design principles. HE could be quickly conducted in the early concept and design stage since there is no need of a larger-scale user testing study.

Some well-recognized universal usability heuristics including Nielsen's ten usability heuristics (Nielsen, 1994), Norman's seven principles for design (D. Norman, 2013), and Shneiderman's eight rules for user interface design (Ben Shneiderman, 1997).

The most frequently used set of heuristics is Neilsen's ten usability heuristics, which are listed as followed:

1. **Visibility of system status**: Users could always learn what is going on with the system by the appropriate dialogue or other forms of the feedbacks, for example, providing the alert window showing the system upgrading progress.
2. **Match between system and the real world**: The system should align with the conventions and practices in the real world to deliver the information in user's viewpoint. For example, the understandable terms for general population rather than technical languages or jargon.
3. **User control and freedom**: Users always own the system control and they could freely leave an unwanted state. And it's better to have the undo and redo functions.
4. **Consistency and standards**: The whole system should follow the same conventions of naming, formatting, …etc.
5. **Error prevention**: Prevent error-prone conditions or consider poka-yoke strategy.
6. **Recognition rather than recall**: The procedures of the operations could be easily recognized from the interface, users don't need to recall from the memory to finish the task operation.
7. **Flexibility and efficiency of use**: Consider both groups of the novice users and the expert users. Allow the customized interface by tailoring the frequent-used functions to speed up the task operations.
8. **Aesthetic and minimalist design**: Simplify the user interface; deliver the precise information and eliminate the unnecessary design while keeping the aesthetics.
9. **Help users recognize, diagnose, and recover from errors**: Deliver the structured error messages by using plain languages to brief the problem and provide constructive solutions.
10. **Help and documentation**: It is necessary to provide easily accessible help and documentation.

Human Factors Measures

Table 1 lists some frequently used human factors methods for user research and usability evaluation. Two axes of the methods are genres (subjective / objective) and the data type (quantitative / qualitative) which can be divided visually into 4 quadrants, which may be helpful for readers to understand the features and the context of use based on this classification.

The subjective methods aim to describe the subjects' mental model based on the self-description or the self-ratings of their own subjective feelings and thoughts. Qualitative data could be collected by either one-on-one or focus group interview with structured or semi-structure questions for the further analysis. Quantitative data could be collected by subjective questionnaires. Likert scale is the most frequent-used measure to quantify the subjective feelings based on the participants' self-rating on a 3- , 5- , 7- , or 9-points-scale, such as Brooke's System Usability Scale (SUS). However, one should notice that, this measure could induce noise of between-subjects variance in a usability testing context since different individuals may adopt different self-rating strategies. The issue could be dealt with by employing a thoughtful experimental design; or the experimenter could consider employing the conjoint design. Conjoint analysis was firstly used in market research to analyze human's preference of products by systematically investigating how people value the related product attributes using questionnaires. And thus, the conjoint design breaks down the product attributes and specify the levels to design the structured questionnaire. For example, the choice-based conjoint design limits the responder's self-rating strategies by guiding the responders to systematically answer the pairwise comparison questions instead of rating questions, which could effectively collect reliable and valid answers. For example, NASA-TLX questionnaire measure the subjective workload based on five attributes of workload. The questionnaire firstly asks the responder about the subjective weighting of each attribute by pairwise comparison of the importance of each attribute.

Objective methods are mainly based on the observation of the subjects' reaction and behavior. The methods for collecting qualitative data included observations or contextual inquiry which are applied by observing human behavior or interview participants in the context of real situation. Data are usually collected by transcribing from think aloud methods, moderator's notes, video records…etc. The methods to collect quantitative data included biometrics (physiological measures), such as respiration rate, heart rate, eye tracking, EEG…etc., and human task performance measure. Time and accuracy are two major descriptors of human performance; however, the metrics are task specific. Some frequently seen task paradigms are derived from the classic human factors experiments which describe the particular cognitive stages of human information processing, such as visual search, respose to a stimulus, and signal detection. A special paradigm is primary and secondary task paradigm, which is frequently used in the driving context to simulate the multi-tasking or interruptive theme.

Table 1. Human Factors Measures

Human Factors Methods	**Quantitative Measures**	**Qualitative Measures**
Subjective	**Subjective Questionnaire**, e.g.: 1.Likert Scale 2.Conjoint Analysis 3.Kansei Engineering	**Interview**, (e.g. 1-on-1, focus group) **Think Aloud**
Objective	**Physiological Metrics** (e.g. Heart Rate, eye tracking, etc.) **Human Task Performance Measures** (Time of Response and Accuracy), e.g. : 1. Visual Search Task Paradigm 2. Stimulus-Response Task Paradigm 3. Primary and Secondary Task Paradigm	**Ethnographic Methods**, e.g.: 1.Observation 2.Contextual Inquiry; **Expert Reviews**, e.g.: 1.Cognitive Walkthrough 2.Heuristics Evaluation

1.2.3 Gap in Nielsen's Usability Model for Measuring Utility

Currently, Nielsen's usability engineering framework is regarded as the paradigm of UX Research and Design (R&D). However, it's unlike ISO-9241-11 in that effectiveness is a main concern. Nielsen's usability model focuses on efficiency, learnability, memorability, human error, and satisfaction. As a result, the scope of usability testing studies is often narrowed to the ease of use level. But in the product research and design viewpoint, assessing usefulness is also crucial since the product utility would directly define the core values and the market acceptance of the product.

To narrow the gap, researchers have often extended the user research scope based on the Technology Acceptance Model (TAM) to assess effectiveness and efficiency for Information Technology (IT) products. TAM is a theoretical human mental model explaining the consumers' technology acceptance behavior by perceived usefulness and perceived ease of use. And in many cases, perceived usefulness was found to be the major predictor of technology acceptance rather than perceived ease of use. For example, Lee and Lehto (2013) studied the use of YouTube by fitting survey data to an extended TAM model and found perceived ease of use was not a significant predictor of perceived usefulness and behavioral intention. They have also found perceived usefulness could be predicted by constructs of task-technology fit, content richness, vividness, and YouTube self-efficacy. (D. Y. Lee & Lehto, 2013)

Nevertheless, the integration with TAM implies the importance of considering utility as a part of the current usability engineering framework. In the following section, TAM and the extended TAM, Unified Theory of Acceptance and Use of Technology (UTAUT) are reviewed.

Technology Acceptance Model (TAM)

Technology Acceptance Model (TAM) could be the most important and widely accepted psychological model in information system research. TAM describes the human mental model of accept or reject an information technology (IT). It was proposed by Fred D. Davis in 1986 as his doctoral dissertation. The conceptual model of TAM is that the actual system use (use behavior) is a response that can be explained by user motivation, and the user motivation is influenced by an external stimulus. The motivation is further explained by three factors: perceived usefulness, perceived ease of use, and attitude toward using.

Perceived Usefulness (PU) is the major determinant of the TAM. It explains that the intention of use is based on the extent that people believe it would help them perform the job better. The definition is “the degree to which a person believes that using a particular system would enhance his or her job performance.” (Davis, 1985) In UTAUT, Venkatesh described the same concept as “Performance Expectancy” (PE). Most studies regarding to the technology acceptance of mHealth applications also find a significant influence of PU toward behavioral intention, including technology acceptance of medical education apps for medical students, physician rating apps for patients, fitness and obesity management apps, Electronic Health Record (EHR) portal, Home Telehealth Service for elderly, and Mobile Health Service, etc. (Bidmon, Terlutter, & Röttl, 2014; Briz-Ponce & García-Peñalvo, 2015; Jaehee Cho, Quinlan, Park, & Noh, 2014; Cimperman, Makovec Brencic, & Trkman, 2016; Deng, 2013; Jeon & Park, 2015; Yuan, Ma, Kanthawala, & Peng, 2015)

Both PU and PEOU are also frequently used as the measurements for system usability test in the scenario of utilization. Brown et al., proposed to measure the PU and the PEOU to illustrate the subjective satisfaction in the context of Health Information Technology (HIT) (Brown et al., 2013). Boland et al., use the questionnaire derived from UTAUT as the subjective measure of the mix method to evaluate usability (Boland et al., 2014).

Perceived Ease of Use (PEOU) is another important determinant of TAM, it was defined as “the degree to which a person believes that using a particular system would be free of effort.” (Davis, 1985) According to TAM, PEOU would also affect the PU.

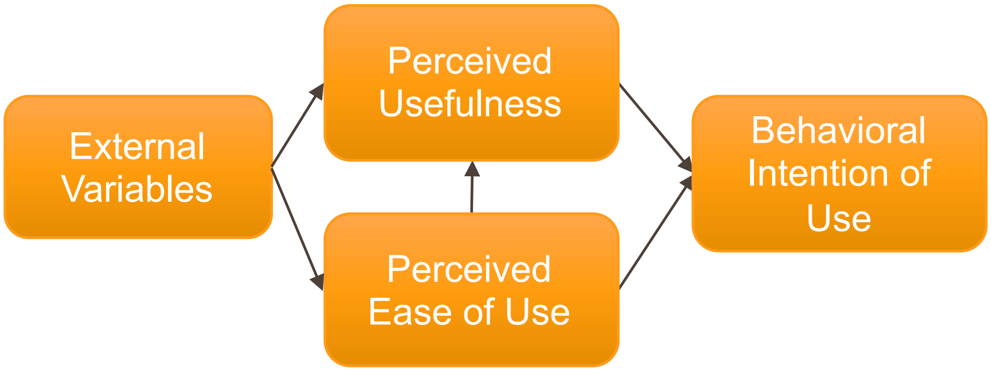


Figure 3 Davis' Technology Acceptance Model (Davis, 1985)

The attitude toward using is the determinant of the intention to use (Marangunić & Granić, 2015). Since the perceived usefulness would be the main determinant of attitude toward using, Venkatesh and Davis further proposed an extended TAM2 to identify the external variables that could influence the perceived usefulness, including subjective norm, image, job relevance, output quality and result demonstrability (Venkatesh & Davis, 2000). In 2003, Venkatesh et al. synthesize prior technology acceptance research and develop the unified theory of acceptance and use of technology (UTAUT). UTAUT theorized four constructs, including Performance Expectancy (PE), Effort Expectance (EE), Social Influence (SI), and Facilitating Condition (FC), would influence the Behavioral Intention and thus the Use Behavior. Venkatesh also firstly introduce four moderators including age, gender, experience, voluntariness of use to provide a theoretical justification of the hypotheses (Venkatesh, Morris, Davis, & Davis, 2003). UTAUT well-predicted the behavioral intention to use a technology and technology use in the organizational contexts. It explained about 70 percent of variance in behavioral intention to use a technology and about 50 percent of the variance in technology use. In 2016, Venkatesh extended the UTAUT with three more constructs, Hedonic Motivation (HM), Price Value (PV), and Habit (HT). They have also taken out the voluntariness as a moderator in a new model UTAUT2 in the context of consumer's technology acceptance (Venkatesh, Thong, & Xu, 2016).

A New Perspective of Utility for Smart Systems

Although TAM suggested the importance of considering utility, but the subjective measures of perceived usefulness and perceived ease of use may not be a holistic approach to assess effectiveness. In recent years, the revolution of smart systems has re-defined the IT product utility.

As the Artificial Intelligence (AI) technologies were introduced, the role of the computer has been changed from a tool of an individual to a smart agent in a socio-technical system. In this context, users are another human agent working with computer to achieve a higher systematic goal. So, the perceived usefulness of the user may not be the most critical measure of effectiveness. For example, to assess effectiveness of the next generation of surgical robot in an operating room, researchers may concern more about the overall system performance and its social impacts to human in the system.

1.2.4 Barriers for Older Adults to Use mHealth Technology

Aging is generally expected to lead to reduced physiological and psychological capabilities, however, not all capabilities decline with age. Fisk et al. (2009) summarized the reduced capabilities of older adults due to the aging process into three categories: sensation and perceptions, cognition, and movement control (Fisk, Czaja, Rogers, Charness, & Sharit, 2009).

For sensation and perceptions, there are age-related declines of auditory sense (especially for older men to hear high-frequency sounds), taste and smell, haptics with lower sensitivity of temperature and vibration (which can cause falls), vision acuity, dark adaption, color contrast, narrower vision field, lower visual processing speed and perceptual flexibility, and susceptibility to glare. In the context of barriers to mHealth technology design, the reduced vision and auditory sense affect the use of mobile technologies directly (Arning & Ziefle, 2009; Athilingam et al., 2016; Grindrod, Li, & Gates, 2014; Harte et al., 2014; Matthew-Maich et al., 2016); reduced haptics sense of vibration would also affect the mobile technology design (Chung, Kim, Na, & Lee, 2010).

For cognition, working memory often declines with the reduced ability of holding and manipulating the information. The decline of the long-term memory is not as much as working memory. Harms to the semantic memory is the least, then the procedural memory. However, it would be hard for older adults to quickly access information and acquire new procedure to inhibit the old one. The prospective memory regarding to do something in the future would decline, however the decline is not evident with the cue such as a reminder. Aligned with the summary from Fisk et al., Farage et al. (2012) conclude from literature that older adults would suffer from processing fewer information bits and reduced ability of recall, especially to the future-based time-based task (Farage, Miller, Ajayi, & Hutchins, 2012). Selective attention such as visual search on the display and reorientation of attentional focus would decline by age. Multitasking is nearly impossible since there are declines of both divided attention and switch attention. Age-related decline shows in spatial cognition, Ziefle and Bay (2005) reported the difficulty for older adults to navigate the hierarchical phone menus due to the declines of memory and spatial cognition (Ziefle & Bay, 2005). Language comprehension depends on the task, age-related declines occur when there's a need of inference and working memory. In the context of health information, age-related declines of semantic fluency, numeracy, and representational fluency influence the comprehension of health-related content (Maša Isaković, Sedlar, Volk, & Bešter, 2016; Kaufman et al., 2003; Morey, Barg-Walkow, & Rogers, 2017). The declines of cognition influence an older adult's performance of human-technology interactions with a longer response time and a higher error rate. For the movement control, the declines of dexterity and fine motor skills could be due to chronic disease such as arthritis. However, even healthy old adults move 1.5-2 times more slowly and less precise than younger adults (Joe & Demiris, 2013; Wallace, Graham, & Saraceno, 2013).

Fletcher and Jensen (2015) reviewed literature regarding to the barriers to use mHealth technologies for older adults who is above aged 65 years. They summarized three main barriers including physical barriers, acceptance barriers, and barriers related to technology design. Physical barriers refer to the physical and mental limitations due to age-related declines of motor, sensory, and cognitive performance, which is aligned with the summary by Fisk et al. (Fisk et al., 2009; Fletcher & Jensen, 2015). The acceptance barriers consider the psychological perspective of technology acceptance, including perceived ease of use, perceived ease to learn, confidence or self-efficacy of mHealth, privacy and security (Cimperman, Makovec Brenčič, & Trkman, 2016; E. Lee, Han, & Jo, 2017; B. R. Wang, Park, Chung, & Choi, 2014). Technology design barriers list the design issues of mobile phone applications for older adults which has been discussed in gerontechnology literature, including: (1) The design is neither ageing-centered nor universal design to consider the needs of older adults (Charness & Boot, 2009; Rodeschini, 2011; Wandke, Sengpiel, & Sönksen, 2012); (2) icons designed in a confusing way, it's difficult to learn the meaning (Santa-Rosa & Fernandes, 2012); (3) it's easy to get lost within the device menu (Arning & Ziefle, 2009; Zhou, Rau, & Salvendy, 2014; Ziefle & Bay, 2005); (4) it's difficult to use soft keys on the mobile phone and touch screens (Zhou et al., 2014); (5) a concern about the battery dying fast (Parker, Jessel, Richardson, & Reid, 2013).

Wildenbos et al.(2018) also summarized the evidence from the related literature and proposed a framework (MOLD-US) of the aging barriers which would affect mHealth usability (Wildenbos, Peute, & Jaspers, 2018). The basic constructs aligned with the key literature of Holzinger et al. (2007), Rogers & Fisk (2010), Czaja et al. (2013), including cognition, physical abilities, perceptions, and motivations (Czaja et al., 2013; Holzinger, Searle, & Witzer, 2007; Rogers & Fisk, 2010; Wildenbos et al., 2018). The construct, "motivations" is similar to the "acceptance barriers" of Fletchers' classification. Wildenbos et al. (2019) suggest, the proposed MOLD-US could serve as the heuristics to evaluate the usability (Wildenbos, Jaspers, Schijven, & Dusseljee-Peute, 2019).

1.3 Problem Statements

Since first personal computer was developed in 1980, information and communications technologies have officially stepped into our daily life and changed our behavior through human-computer interactions. In 1990s, researchers focused on the product efficiency and the impacts on individuals. For example, Neilsen proposed the usability engineering framework to design the usability in the technologies for ensuring the ease of use. Nowadays, this framework has become the paradigm of UX/UI design in software engineering. Since 2000, the discussion has been brought to the product effectiveness and social impact level, Fogg (2003) firstly talked about the concept of persuasive technology, which emphasizes the potential of influencing human attitude, decisions and behaviors and making a broader social impact through human-computer interactions. However, there's still a gap to facilitate the concept in the product design and development perspective, there is a need of integrating the existing frameworks with the proper design research method to establish the paradigm of the persuasive design framework.

A successful UI design for the persuasive mHealth app would increase users' intention of use and nudge users' health behavior. To maximize the market value, the UI design of a new information technology should take the major group of potential users into account. However, currently, most of the mHealth app designs have not considered the older adult users' needs for (or barriers against) using the apps.

To improve the usability, the information technology should ensure the effectiveness of delivering the information to the end user. In the context of dietary control type mobile health apps, effectively delivering nutrition information is a main concern. Furthermore, when the information technology is used for health behavior change, the UI design of the persuasive technology, which has constructed a virtual choice architecture in the digital environment, should effectively influence the end users' decision making and use behavior and thus guide the user to adopt a healthier lifestyle. However, it's still unclear which kind of design elements and how best to deploy them to achieve maximum effects.

So, to summarize, there are three purposes of conducting this research:

1. Systematically design and develop the UI of the mHealth app to increase the usability of the proposed smart systems for the specific user group, such as older adults.
2. Propose the objective metrics and user testing protocol to evaluate the persuasiveness of the app based on human information processing theories.
3. Systematically design the UX of the mHealth app to ensure the effectiveness of the persuasive technology.

1.3.1 Research Questions and Hypothesis

In the following chapters, the project scope will be further narrowed to the persuasive mHealth app design. And a dietary management app will be proposed. In this section, the research framework is briefly described. Two research questions were proposed to better guide a research for the proposed app to serve the above objectives.

1. Which UI design elements of a persuasive dietary management app would improve usability?

2. What are the critical UI design elements of a dietary control mHealth app to nudge user’s choice, which infers the dietary behavior change?

To facilitate the persuasive design, three nudge approaches were proposed based on literature review (see section 2.4) and brainstorming. The approach was translated to the UI design element to form the research hypothesis to answer the research questions.
The first consideration is the implementation of creating the choice structure on the UI. We considered the use scenario of meal planning and visualized the cognitive processing stage of identifying the healthiness for a food option as a searching-based UI; and the stage of selecting between 2 alternatives as a choice-based UI. And based on the findings about multi-dimensional signal detection theory (the theoretical basis of the signal detection analysis for the two-alternative forced choice (2AFC) task the), human performance under a 2AFC testing paradigm should be better than the former yes-no paradigm when discriminating the signal. So, we assumed choice-based UI would be a better design which would help people identify healthy food when considering healthy-related information as signal.

H1.1: The choice-based search result layout would significantly improve usability comparing to the browsing-based search results layout.

H2.1: Choice-based UI is significantly better than the searching-based UI to ”nudge” users to select the system defined ”truth”.

Based on Bauer and Reisch’s taxonomy of nudge approaches for healthy eating (2018) and considering the difficulty levels of implementation, the information nudge and the default nudge (Bauer & Reisch, 2018) were chosen to be developed as another two UI design elements.

For information nudge, dual process theory (Kahneman, 2011; Sanjari, Jahn, & Boztug, 2017) and was considered and two kinds of nutrition information was selected: text list-based nutrition label, Nutrition Facts Panel; and a symbolic interpretative nutrition label, Nutri-scores. Nutri-Scores label was assumed to be more effective for human information processing of nutrition information based on Rasmussen’s Skill-Rule-Knowledge based model and its application of signal, sign, symbols design (Rasmussen, 1983).

H2.1: The FSA Nutri-scores label would significantly improve usability comparing to the FDA Nutrition Facts Panel label.

H2.2: The specificity of nutrition information has a significant effect to ”nudge” users to select the system defined ”truth”.

According to Thaler and Sunstein’s nudge theory (2008) and the best practice of “Opt-In/Opt-Out” paradigm of public policy making, default nudge was assumed to be powerful to influence human’s decision making. We implemented the default nudge by the pre-selected button, and a hypothesis was established:

H2.3: The default nudge (existence of pre-selection) has a significant effect to ”nudge” users to select the system defined ”truth”.

2. LITERATURE REVIEW

In this chapter, the scope is narrowed down to the persuasive mHealth app for dietary management. Section 2.1 reviews the historical background of related persuasive technologies for the chosen case. Section 2.2 reviews the related works of dietary management. Section 2.3 reviews the related theoretical framework for the proposed extended usability framework for persuasive mHealth apps. Section 2.4 reviews the related works of the proposed persuasive nudge design elements.

2.1 Historical Background

Since 1980s, development of the first personal computer has changed our daily life behavior through human-computer interactions. Researchers firstly discussed human-computer interactions from the usability perspective, which is focused on efficiency to ensure the ease of use of technologies. As the systems are getting smarter by integrating ubiquitous computing, data analytics, and machine learning algorithms to solve more complicated problems, the discussion about human computer interactions has also been brought to a higher level. For example, the utility of computers and the effectiveness; and the social role of computers and their impacts. Fogg (2003) firstly mentioned the concept of persuasive technology to describe the idea of influencing human attitude, decisions and behaviors and making a broader social impact through human-computer interactions, which defines a new utility of technologies.

2.1.1 Mobile Health Technology

There's an expectation that the applications of connected smart things would bring a revolution to the healthcare service sector, by accelerating the service innovation and providing a patient-centered, technology-enabled smart service such as mobile health. Mobile health (mHealth) is the application of mobile and related wireless technology to medical or public health service. It has become a hot issue since the proliferation of mobile technology, in the beginning, the discussion was about adopting the Personal Data Assistant (PDA) for service providers in the medical and healthcare setting, and nowadays the smartphone has become the mainstream in the mHealth research.

With the growing pace of connected smart things, such as smartphones, tablets, and smart watches, more and more researchers are focused on developing a smart service system by explicating the ubiquitous computing, context-awareness, and direct manipulation system features of these devices. When designing a smart service system to deliver healthcare service, collecting user's personal health behavior related data is a key to success. First of all, smart things must learn from user's past data to train with algorithms to adapt and provide a better customized service. The system should also allow users to acquire information from the smart things. Thirdly, in the healthcare delivery process, value could be co-created by the patient's activities of collating information (i.e., retrieving information, managing daily activities.), and co-learning by seeking the related information from other resource (McColl-Kennedy, Vargo, Dagger, Sweeney, & van Kasteren, 2012).

There's a positive impact of individual's health outcomes with the patient's engagement of collating information and co-learning, especially in the dietary self-management. Evidence-based research found that mHealth apps service innovation with connected smart things may improve the efficiency and effectiveness of health service and thus improve health outcomes for an individual. For example, using the smartphone dietary apps could encourage patients adhering to the self-management intervention (of recording food intakes for calories control) and get better health outcomes in a weight control program (Burke, Wang, & Sevick, 2011; Michelle Clare Carter, Burley, Nykjaer, & Cade, 2013).

With more and more successful controlled experiment trials, it's possible in the future that the healthcare practitioner would prescribe patients with a smartphone mHealth app (Martin, Vicente, Vicente, Ballesteros, & Maynar, 2014). However, there's a research gap between the optimal design of system feature allocation and successful outcomes. It's unknown what kind of system features and user interface design attributes of the smartphone mHealth app makes the success of user's preference to adopt the technology and adherence of using the app to change their health behavior. As a result, there are huge number of dietary management related health apps appearing in the app store but just few of them are successful across the mobile health market (J. Cho, 2016). Most mHealth apps are neither following clear design guidelines to attract the consumer's attention nor strategies to support health behavior change (Azar et al., 2013). In this context, patients may waste time searching and they could make a wrong decision of selecting a related app (Eng & Lee, 2013). In the viewpoint of practitioner, even knowing that mHealth may help, they are still hesitant to make a recommendation of an app due to lack of the awareness of the best recommendation (J. Chen, Lieffers, Bauman, Hanning, & Allman-Farinelli, 2017).

Project Scope

Previous mHealth research has shown the trend of development and identified classification of mHealth app by its usage type. For example, Ali et al. (2016) reviewed 3277 articles on PubMed from 1993-2015, and made a taxonomy of mHealth apps based on the purpose of the mHealth intervention: health promotion, disease prevention, diagnosis, treatment, monitoring and support for health services (Ali, Chew, & Yap, 2016).

Liu et al. (2011) conducted a developer survey in 2010, they found out that tracking tools are the most popular feature among surveyed patients. In the developer's viewpoint, tracking tools for the diabetes have the highest business potential among the therapeutic apps for the chronic condition (C. Liu, Zhu, Holroyd, & Seng, 2011). Among the tracking tools in general, diet tracking is important since healthy eating is one of the key constructs of the healthy lifestyle (Michie et al., 2011). In this work, the research objective targets the dietary management app context which assists with a diet self-management intervention.

To better understand the potential competitors, their products on the market, and the context of use scenario. The keyword "mHealth app" and "design" were used to search in two peer-reviewed journal databases, Scopus and Web of Science. I found 98 related articles in Scopus and 45 related articles in Web of Science from 2011-2018, removed 13 duplicate records, and did the text mining to cluster the relevant keywords from the title and abstract fields. There were five clusters are created by this method, the visualization of the keyword's clusters are shown on Figure 4.

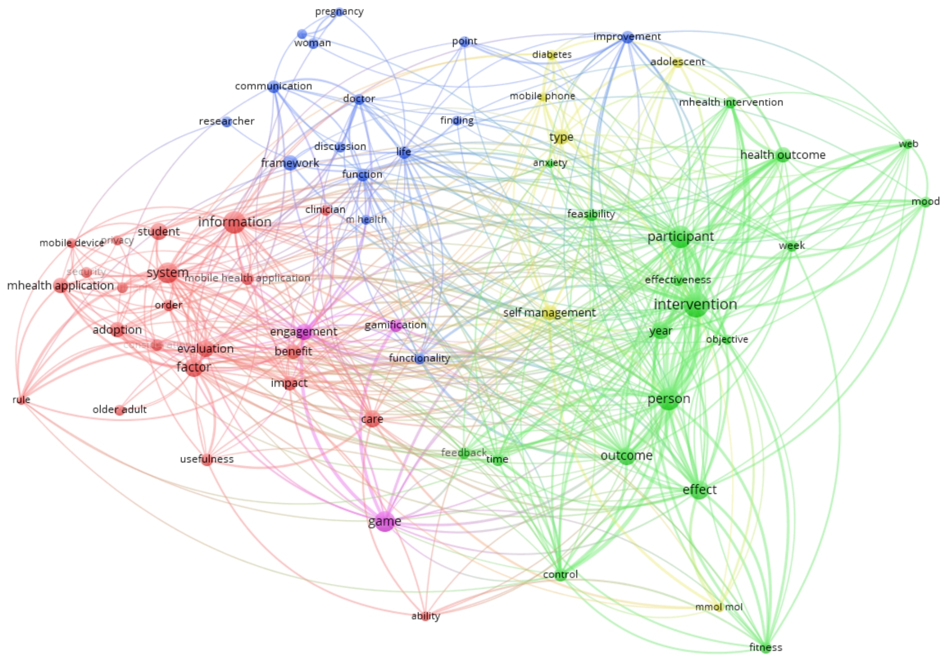


Figure 4. Cluster Analysis for the 130 literature with the keyword "mHealth app design".

We used the most related and frequently seen keywords to refine the search result and get 5 article lists, by reviewing from the top highly cited articles on each list, we found 5 trends of mHealth app design research directions:

1. Technology acceptance of mHealth app as an information system.
2. Use of mHealth app intervention to improve health outcomes
3. Pregnancy mHealth app design
4. Diabetes, self-management app design
5. Gamification to increase the user engagement with mHealth apps

Self-management app design for Diabetes is selected as the developing themes since it has appeared to be one of the important research trends, and is related to health behavior change purpose of enhancing a healthier lifestyle. Healthy eating leads to a healthier life as evidence-based research has shown that diet control interventions significantly improve the health outcomes in weight control for both healthy and chronic disease conditions. Additionally, there are several controlled-experimental studies that have reported a better health outcome of weight control by using smartphone dietary app instead of traditional paper diary to record personal food intake data in self-management intervention (Michelle Clare Carter et al., 2013; Castelnuovo et al., 2017; W. Lee, Chae, Kim, Ho, & Choi, 2010). As a result, in this article I focus my research objectives on the design of food tracking apps to develop a more effective dietary self-management system.

mHealth app Design Issues

There are an increasing number of mHealth apps appearing in the market, but few of them make a successful penetration to the customer (J. Cho, 2016). And according to Chen's study in 2015, in general, the user retention rate of smartphone apps is low: there were 77% of users delete the smartphone app in the first 3 days after they downloaded the app (A. Chen, 2015). However, in the context of healthcare service, patient's adherence to the intervention is the key to changing an individual's health behavior and get a successfully improving health outcomes. In this context, low user retention rate to the mHealth app could be a problem. Additionally, research has reported prevalent usability issues for the existing commercial mHealth apps. From the human factors engineer's viewpoint, these two issues could be due to several issues including:

1. The current design framework for most of the commercial mHealth apps cannot meet users' needs and does not take users' limitations into consideration.
2. The current usability evaluation framework for most of the commercial mHealth apps is not suitable for testing the UX quality for a smart healthcare service system.

Inclusive Design of mHealth Apps for Older Adults

In addition to the current design issues of mHealth apps for general publics, inclusive design is another concern since the potential demographics of the major mHealth app users is older adults.

The world population is getting older, the United Nations reported that 9 percent of the global population was aged 65 or above in 2019, and predicted a growth to 25 percent of populations in North America by 2050 because of the longevity of the baby-boomer generation (United Nations, 2019). The prevalence of chronic disease and multimorbidity in older adults is also increasing (Freid, Bernstein, & Bush, 2012). Aging is a complex process of the physiological declines and psychological changes linked with the social condition changes. It is associated with health threats such as frailty (Fried et al., 2001), malnutrition (Guigoz, Lauque, & Vellas, 2002; Hickson, 2006), and chronic conditions and multimorbidity such as hypertension, cardiovascular disease, and type II diabetes (Freid et al., 2012; Go et al., 2013; Mancia et al., 2007; van Oostrom et al., 2016).

According to Farber et al. (2011), nearly 90% of Americans aged over 65 want to stay at home if possible, which indicates a substantial potential demand for solutions that support aging in place (Farber, Shinkle, Lynott, Fox-Grage, & Harrell, 2011). In 2018, AARP still reported nearly 80 percent of adults age 50 and older indicate this same desire. And 50-60% of younger adults aged 18-49 want to stay at their homes or communities as they age. Despite the fact that most adults have an intention to grow old in their communities, many of them are concerned about age-related disability and morbidity and believe they will need home healthcare support to live independently. In the viewpoint of healthcare providers, home support may be a burden as they try to allocate human resources offsite while trying to ensure patient-centered service with respect to the rights and dignity of older adults (Byrne, Frazee, Sims-Gould, & Martin-Matthews, 2012; Gregory, Mackintosh, Kumar, & Grech, 2017; Hu, Chau, Liu Sheng, & Tam, 1999; McCormack, Roberts, Meyer, Morgan, & Boscart, 2012).

The potential of technology to meet this demand is promising. Smartphones have the ubiquitous computing and context-aware features and had been adopted by more than three quarters of the US population in 2015 (Poushter, 2016). Additionally, an increasing number of Behavior Change Techniques can be easily implemented via smartphone applications, including self-monitoring, goal setting, social support, gamification, etc. (Kankanhalli, Shin, & Oh, 2019). For example, the connected network of the smartphone offers a chance to strengthen social support and promote the health behavior change based on Social Cognitive Theory. According to Bandura (2001), "human health is a social matter, not just an individual one.", especially in the context of health promotion (Bandura, 2001). Currently, most of the evidence-based research has already found the efficacy of adopting smartphone applications as a self-management support tool for chronic conditions including epilepsy, diabetes, hypertension, and obesity (Free et al., 2013; Kelly, Reidlinger, Hoffmann, & Campbell, 2016; Shegog & Begley, 2017; Sorgente et al., 2017; Whitehead & Seaton, 2016; Wu et al., 2017).

However, there has been a concern about delivering service via mobile technology to older adults. In the past, there was a stereotype that older adults may not want to use new technology. So, there may be no need to change the traditional healthcare service delivery model for older adults. The statement may be based on the reason that there was a digital divide between older adults and the younger generation, partially due to older adults growing up with limited or no access to technologies (Wallace et al., 2013).

The statistics shows the situation has been changed. According to Pew Research Center (2018), in 2018, 85% of adults above 65 years own a cell phone, up from 69% in 2012; and 53% of those 65 and older has a smart phone, up from 47% in 2012 reported by AARP and Pew Research Center (Pew Reserach Center, 2018; Zickuhr & Madden, 2012). Some researchers have also found evidence showing that most older adults are willing to learn new technologies as long as the design fits their needs and requirements and some older adults can even be identified as tech savvy (Brauner, Calero Valdez, Schroeder, & Ziefle, 2013; Hanson, 2011; Olson, O'Brien, Rogers, & Charness, 2011). In conclusion, it's possible for older adults to adopt mHealth technology. However, there's a need to design for older adults specifically, based on the insights of their needs and limitations.

2.1.2 Persuasive Technology

Persuasive technology is defined as an interactive system which is designed to implement "the attempt to change one's attitude and behaviors rather than coercion", based on Fogg's definition of persuasion (B. J. Fogg, 2002). Hamari et al. (2014) have further pointed out that persuasion is intentional and contextual. The persuasive design aim to guide users towards desired attitude and behavior change, and the timing, events, and strategy of persuasion should be considered (Hamari, Koivisto, & Pakkanen, 2014).

Fogg (2003) suggested seven types of persuasive technologies when computers are used as tools, including Reduction, Tunneling, Tailoring, Suggestion, Self-monitoring, Surveillance, Conditioning (B. J. Fogg, 2003). Reduction technology persuade users by simplifying the complex behavior with simpler tasks and thus the user is encouraged to practice by doable steps (Oinas-Kukkonen & Harjumaa, 2009); tunneling technology guide users through process of changing the behavior; tailoring is a concept of providing information based on user's need; conditioning is based on behaviorism to provide rewards or give punishments. Reduction, tunneling, tailoring, and conditioning are more like "strategies" rather than practical steps to take. Suggestion is a practical approach to provide suggestions to users; self-monitoring and surveillance track the progress of behavior change by self or by others.

Although this work provided some persuasive design directions, however, this classification is not based on the solid foundations of behavioral change theory. Many important behavior change techniques (e.g. social support, providing information, personalization, …etc.) are not mentioned in this classification as well. Some categories are merely conceptual persuasive strategies with no clear boundary to classify such technologies since most persuasive technologies applied multiple persuasive strategies at a time. In this context, it is difficult for designers to make systematical design decisions. For example, a persuasive pedometer is designed for self-monitoring the progress, but it's also possible to provide suggestions and its user interface could also follow the design principles of reduction and tailoring. It would be difficult to classify this pedometer and compare it to the similar products in the product design viewpoint.

To better tailor to product design needs, Oinas-Kukkonen and Harjumaa (2009) further collected more theoretical-based persuasive features such as reminders, rewards, social support,…etc., listing definitions and design principles for 27 persuasive features. They also classified the features and the related design principles based on use scenarios including primary task support, dialogue support, system credibility support, and social support. This framework provides foundation to design and evaluate persuasive technology, more details is discussed in section 2.3.4.

Most persuasive technologies apply persuasive strategies as product features but are not designed for persuasion. Recently, recommender systems which are designed to provide suggestions, appear to be a strictly defined persuasive technology. The technology could also be integrated to other systems as a persuasive feature. In this dissertation, recommender system is suggested to be adopted in the final design to provide the persuasive features on a self-monitoring tool for dietary management. For this reason, it is reviewed as the background knowledge in the following section.

There are growing numbers of strictly defined persuasive technologies appearing on the market as the advanced computing technologies are developing at a fast pace, such as Social Assistive Robotics (SAR). SAR implements the social support with the humanoid robots and thus attracts researchers' focus. However, the related discussions about technical issues, emotional design, and use cases of computers as social actors, etc. are beyond the project scope. Therefore, SAR is introduced in section 8.4.2 as one of the future research directions.

Recommender System

A recommender system is a software tool to predict user preference of items from a huge amount of available options on the web. Recommender systems could be regarded as a type of decision support system, since it's developed to support the users' daily life decisions. However, it's unlike the enterprise decision support system, which aims to solve a super complex enterprise level decision-making problem under the uncertain circumstances with involving multiple parties involved, multiple objectives, and many constraints of enterprise resources. The recommender system is more focused on the single activity of a consumer. It could be as simple as recommending few options to help an individual under the scenario of choosing a movie for Saturday night from a movie database; or finding the related book which is similar to the user's most recent book purchase from the online bookstore; or suggesting a new song based on user's profile from an online music streaming service. Of course, this kind of prediction problem of the recommender system could also be quite complicated as there may be a seemly unlimited number of options and the user's mental model of preference is vague. Considering the food recommender system design as an example in this context. The food recommender system could be very difficult to design as the number of food ingredients is in the thousands and all the possible combinations of ingredients produces a huge number of recipes options. Besides, user's food preference is usually fuzzy, for example, a user may like oatmeal for breakfast but not for dinner.

There are four frequently-seen types of recommender systems based on the difference of the algorithms used, collaborative filtering (CF), content-based (CB), hybrid system, and knowledge-based system. The collaborative-filtering recommender system is based on the crowd's opinion, which is the most popular recommender algorithm since the winner of the 2006 Netflix competition using CF. There are several kinds of CF algorithms including user-based, content-based, and matrix factorization. The most common user-based CF predicts the user's rating of a specific item by classifying the user to a similar user group. The CB recommender system is based on the user profile and the description of the item. There's a need for CB to pre-process user's profile or the description of the information retrieved; the knowledge-based recommender system is useful when the rating of each item is low or when there's a special requirement. The frequently-seen algorithms for the knowledge-based system include the case-based, constraint based, or critique -based.

Interestingly, there is some evidence that recommender systems may be preferred by older adults. According to the study by Beel et al. (2013), the older adult users (age 50-54) click on the recommendation more than younger adult users (age 20-24) in a large-scale study of the click-through rate of 1028 users for nearly 38,000 research paper recommendations (Beel et al., 2013).

2.1.3 Historical Background of Self-Monitoring Tool for Dietary Management

This dissertation is focused on developing a persuasive mobile dietary management app. Currently, most mHealth apps for dietary management are developed as a self-monitoring tool. For this reason, the historical development backgrounds of the tool are reviewed in this section.

Traditional Dietary Assessment Tools

For the dietary management purpose, it's necessary to understand an individual's eating habits but the idea is difficult to facilitate. Most people are unaware of what and how much food they have eaten, so it makes the food tracking and dietary assessment difficult. Traditionally, it requires skillful and knowledgeable personnel to moderate an interview or to design a well-defined questionnaire to collect this information in order to help study participants to recall their daily eating behavior and assess the food choice quality and quantity. Some practical subjective dietary assessment methods including the food frequency questionnaire (FFQ), Dietary Record (DR), 24-hr dietary recall food record method (Shim, Oh, & Kim, 2014). Those tools are some existing instruments developed by experts to collect detail information of consumed food and beverage in 24 hours for a respondent in the real-world healthcare setting. To serve the dietary assessment purpose, it guides the respondent not only to recall the eaten item and the time of consumption but also to think through the detail of the item including the ingredients and the serving size. Although those tools are cost-effective approaches to collect precise data, they usually impose a heavy burden on respondents. Furthermore, due to the limitation of memory capacity, human errors of omission often occur when performing the recall task.

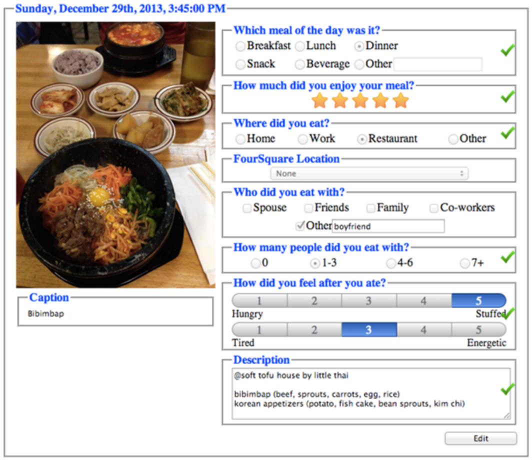
To improve the human factors issues in existing methods, currently, several innovative smart food technologies for dietary assessment have been developed in academics and industry to reduce human efforts to recall from memory and manually record data.

Web-Based Dietary Assessment: Food Journaling

Food journaling is an effective feature of tracking daily food choice and food intakes to facilitate the self-monitoring of healthy eating behavior. In the medical discipline, recall and paper diary is the key method to investigate and monitor an individual's eating behavior. In recent years, it is getting more convenient for an individual to record the dietary diary using the web-based platform or on easy access mobile devices such as smartphones or tablets which can make the interventions more effective. For example, Hollis et al. (2008) established the Weight Loss Maintenance (WLM) paradigm, which is a 6-month educational program and compared the WML trials with the regularly personal contacts intervention and the usage of the interactive self-monitoring website intervention. They have found better health outcomes improvement with the internet usage arm of WML trials (Hollis et al., 2008). Funk et al. (2010) did the secondary analysis with the internet usage group arm of WML trials of the study from Hollis et al. They found less weight regain is strongly associated with those website usage variables such as number of log-ins, minutes on the website, number of weight entries, number of exercise entries , and sessions with additional use of website features after weight entry (Funk et al., 2010). Mark et al. (2009) also made the same conclusion from another large-scale web-based weight loss maintenance program (23000 users, and 4400 participants to the study). They have further found, compared to the younger group, participants aged over 65 are more actively participated in the program, lost more weight, and were more likely to stay in the program (van der Mark et al., 2009).

More recently, with a focus on the increased smartphone usage in the U.S., Carter et al. (2013) did a controlled experimental trial in a 6-month weight loss program to compare the effectiveness of self-monitoring intervention for food intakes between the traditional paper diary measure and the smartphone app measure. They found a higher retention and a better health outcome of weight loss in the smartphone app group (Michelle Clare Carter et al., 2013). Laing et al. (2014) performed a randomized control trial to compare the effectiveness of a weight loss program between the intervention group using a smartphone application and the control group. The experimental app was a commercial app, MyFitnessPal, which is currently the most popular and the well-known app in the weight control category of mHealth apps. It offers the functions of goal setting, food intakes tracking (by user manual input into the dietary diary),and physical activity tracking (by connecting to the tracking devices or user manual input). However, the study did not found significant change in the health outcomes in terms of weight loss and SBP or in other self-reported changes in health behavior. Interestingly, however, they did observe that the intervention group used the goal setting function the most. Although most users are highly satisfied with MyFitnessPal, few of them adhere to using the app after the first month (Laing et al., 2014).

Solving the user retention problem is an active area of research. To solve the usability problems in weight control apps, Cordeiro et al. (2015) did interviews and surveys to better understand consumer's viewpoint of the existing food journaling methods included paper diary, mHealth apps (MyFitnessPal, Weight Watchers, and others), Fitbit or other physical activities tracker, or Desktop/Website based solutions. At the end of their study, they proposed a DECAF (Diary of Emotions, Context, and Food) project. In the DECAF project, the participants would install an mHealth app with the feature of "lightweight, photo-based food journaling" and a website for users to provide more detail and reflection. The mHealth app is designed to be used at the moment before the users eat (to remove the recall problem). The food logging page is shown as Figure 5, in which the user logs the food journal by taking a picture and finishing the questionnaire of their eating experience. In the contrast of the existing method, DECAF is designed to release users' burden of recalling food intakes and assessing a diet. The project has been evaluated with a 4-8 weeks field study for 14 healthy users and 13 obese and overweight users. Most of the participants thought the proposed photo journal lower the barrier of food tracking comparing to the existing journaling methods. However, 93% of them still reported forgetting to log the food journal (Cordeiro, Bales, Cherry, & Fogarty, 2015).



Sunday, December 29th, 2013, 3:45:00 PM

Caption: Bibimbap

Which meal of the day was it?
○ Breakfast ○ Lunch ● Dinner
○ Snack ○ Beverage ○ Other ✓

How much did you enjoy your meal?
★★★★★ ✓

Where did you eat?
○ Home ○ Work ● Restaurant ○ Other ✓

FourSquare Location: None

Who did you eat with?
□ Spouse □ Friends □ Family □ Co-workers
☑ Other boyfriend

How many people did you eat with?
○ 0 ● 1-3 ○ 4-6 ○ 7+ ✓

How did you feel after you ate?
Hungry 1 2 3 4 [✓ 5] Stuffed ✓
Tired 1 2 [✓ 3] 4 5 Energetic

Description:
@soft tofu house by little thai

bibimbap (beef, sprouts, carrots, egg, rice)
korean appetizers (potato, fish cake, bean sprouts, kim chi) ✓

Edit

Figure 5. Dietay intakes data input page of Cordeiro's DECAF food journal project

(Source: Cordeiro, F., Bales, E., Cherry, E., & Fogarty, J. (2015). Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture.)

Image-Based Dietary Assessment: Food Recognition

The breakthrough and continuous progress of optical technology has led to alternative forms of image-based dietary assessment. The camera on the smartphone has the potential to function as the machine vision of the mobile phone. The widespread use of smartphone cameras contributes huge image databases for image recognition. And based on the pioneering image processing techniques including image features extraction and the deep learning algorithms, the system has the potential to accurately recognize different kinds of objects from the images and learn to interpret by its features, for examples human faces and emotions (Bettadapura, Thomaz, Parnami, Abowd, & Essa, 2015; White, Dunn, Schmid, & Kemp, 2015).

Food images recognition is relatively difficult compared to other kinds of objects because there are many food ingredients for each food category, which could form a huge amount of possible combinations and recipes. In 2009, Chen et al. published the first visual dataset of standardized fast foods videos and images from 11 fast food chains listed in USDA Food and Nutrient Database for Dietary Studies and National Nutrient Databases. There are 101 different foods selected from those chains including burgers, pizza, salads, etc. Their long-term goal is to connect the food image database with the FDA food composition and nutrition database (M. Chen et al., 2009). Yang et al. worked predicting nutrition data from image classification using Support Vector Machine (SVM) method utilizing four statistics pairwise local feature recognition methods. However, the accuracy of classification of 61 food categories was lower than 10%. Not until they reduced the classification down to 7 main categories, could the accuracy be raised to 80% with the OM feature recognition method (Yang, Chen, Pomerleau, & Sukthankar, 2010).

Since 2010, a research group at Purdue University has proposed the TADA project (Technology-Assistance Dietary Assessment) with several publications over the past 10 years. They have built up a standardized smartphone food image database and developed advanced food recognition technology to train a system to recognize an ordinary meal image. They have also successfully implemented the image-based dietary assessment system to automatically recognize food and serving size on the smartphone, with the integration of the standardized smartphone images dataset, food recognition, and classification using the machine learning algorithm, Supporting Vector Machine (SVM) (Bosch, Zhu, Khanna, Boushey, & Delp, 2011; F. Zhu, Bosch, Khanna, Boushey, & Delp, 2015). In recent years, there are more researchers using the advanced learning algorithms such as convolutional neural network to recognize more food items and the serving size with a higher accuracy (Kawano & Yanai, 2014; Chang Liu et al., 2016).

In 2018, Jiang et al. (2018) used the google glasses to implement the real-time nutrition information provision on Augmented Reality (AR) in the real-world grocery shopping scenario (Jiang, Starkman, Liu, & Huang, 2018). The food images recognition technology is also getting proliferation. Recently, Huawei, a Chinese mobile phone manufacturer, debuted a new smart phone with the food recognition features embedded. The user could use the smartphone camera to focus on fresh produce such as an apple, which is a grocery item and the smartphone could automatically search for the food composition and nutrition database to show the calories in real-time.

Healthy Food Recommender System

As section 2.1.2 has mentioned, recommender system could also be integrated with other tools to provide persuasive features. This paragraph reviewed the developing backgrounds of healthy food recommender system as a technical foundation to support the potential design ideas of persuasive dietary management app.

Mika (2011) has defined 2 types of healthy food recommender systems, the first one is the recommendation of the healthier food or recipe which is similar to the food user preferred (Mika, 2011). For example, Freyne & Berkovsky (2010) proposed a food recommender system by providing the recommendation based on predicting user's rating of a new (healthier) recipe or food item by their previous rating. To implement the prediction, they have tried the content-based algorithm, collaborative filtering algorithm, and some hybrid algorithms. They found CB algorithm is robust among different off-line evaluation metrics including overall accuracy, classification precision, recall, and F1-scores . They suggested a CB food recommender system by firstly decomposing the target recipes into ingredients and assigning the rating to each ingredient, and then predicting the unknown rating of a new recipe (Freyne & Berkovsky, 2010).

The second type of Mika's classification is based on the nutritional needs of the user. Such systems mainly used the CB algorithm to implement the prediction. For example, Agapito et al. developed a health-oriented food recommender system on the smart phone app. The user needs to fill in the questionnaire to identify their possible chronic conditions, and then the system would recommend the dietitian recommended recipes for the patient with thir particular chronic disease (Agapito et al., 2018). Tran et al. (2018) further added 2 other types into Mika's classification (Trang Tran, Atas, Felfernig, & Stettinger, 2018). The third type is the hybrid system, which aims to balance the user preference and nutrition needs of the user, for example, Elsweiler et al. (2015) proposed to add the nutrition error item into the predictive formula of user's preference (Elsweiler, Harvey, Ludwig, & Said, 2015). The last type is food recommender for social groups, such as family groups, friend groups, since group dining is a frequently-seen theme in many cultures. Some group decision-making theories could be applied in this context. However, since the group dinning setting is beyond the project scope, it is not further discussed in this dissertation.

2.2 Related works of Persuasive Technologies for Dietary Behavior Change

2.2.1 Related Works of mHealth apps for Dietary Control

Diet management is a popular form of mHealth app. A recent content analysis of diabetes managements apps from the Apple store found 36 that feature diet management features such as food journaling and intake tracking (C. Gao, Zhou, Liu, Wang, & Bowers, 2017). Most of these (67%) provided access to a searchable food database, some allowed picture uploading and a degree of social networking (22%) and a few required recording similar to that of a paper diary (8%).

In the U.S., commercialized web-based dietary assessment apps have been developed based on USDA food composition database and other commercial crowdsourcing database, such as MyFitnessPal and LoseIt. MyFitnessPal has more than 140,000,000 users, and it's a top-rated weight management app in the Health & Fitness category of Apple store and Google Play. It offers the dietary assessment function with the barcode scanning technology and a crowdsourcing database. However, some negative feedbacks is levied at the app, mainly due to the data quality from crowdsourcing (inaccurate nutrition information) and the difficulty of learning and continuing to use the current text-list based UI.

In academic literature, more attention is paid to the research of advanced technology such as photo-based dietary assessment, or to developing a dietary app tailored to the patient with specific symptoms. For example, Hongu et al. (2015) developed a photo-based 24-h recall questionnaire on the smartphone app platform to help individuals easily record their food intakes by taking the pictures (Hongu et al., 2015). Astell et al., (2014) considered the fact that Alzheimer's dementia is usually associated with malnutrition. They developed the Novel Assessment of Nutrition and Ageing (NANA) toolkit system on the tablet to support the multidimensional assessment of nutrition, cognition, and physical activities. The system allows older adults to record daily food intakes by selecting from the database (Astell et al., 2014). Hakobyan et al. (2016), used a participatory design method to design a novel user interface of dietary diary for Older Adults with Age-Related Macular Degeneration (Hakobyan, Lumsden, Shaw, & O'Sullivan, 2016). Elbert et al. (2016) proposed sending the health information via mobile phone to promote fruit and vegetables consumption. They tested the difference between text-format and audio format and found the effect was moderated by health literacy (Elbert, Dijkstra, & Oenema, 2016). Eyles et al. (2017) proposed an innovative smartphone application that enabled shoppers to scan the barcode of a packaged food and see a real-time nutrition label with the traffic light format, along with suggestions of lower sodium alternatives on the screen (Eyles et al., 2017).

A couple of studies in Europe focused on updating the official food composition and nutrition database in their redspedctive countries and developed an online nutrition assessment platform. For example, Carter et al. (2015) developed a dietary assessment app by re-designing the database in the UK to include several frequently-seen commercial foods and used the 24-h dietary recall questionnaire format to track a user's daily food intakes (Michelle C Carter et al., 2015). Svensson and Larsson (2015) developed an innovative dietary assessment app to assess Energy Intake (EI) and Total Energy Expenditure (TEE) based on a complete Swedish food database which was developed during 2010-2011. At that time, the Swedish National Food Agency had developed a web-based dietary intake tracking tool and conducted a national food survey to develop the database. Svensson and Larsson further developed a mobile-based dietary assessment tool for adolescent in Sweden (Svensson & Larsson, 2015). They further conducted a large-scale interview with 75 users and analyzed the result based on Self Determination Theory, and found the app motivated participants in changing their dietary behavior and lowered the perceived barrier of facilitation (Svensson et al., 2016). Elsewhere, Wellard-Cole et al. (2018) developed the "Eating and Tracking" app in Australia, which provides dietary assessment based on a revised set of guidelines for Australian for younger adults (age 18-30) with the 2011-2013 Australian food database (Wellard-Cole et al., 2018).

There are also a few studies that considered the interactive graphical user interface design of the dietary assessment in the Europe. For example, Franco et al. (2018) developed a graphical food frequency assessment app (eNutri) in the UK, which provides a graphical format of food frequency questionnaire (FFQ) for users to record daily food intakes easily and provide the personalized nutritional advice (Franco, Fallaize, Lovegrove, & Hwang, 2018). However, the current commercial dietary assessment apps in the U.S. are seldom focused on integrating the advanced technologies and the interactive graphical user interfaces to improve the user experience, and thus, it results in lower user retention rate (J. Cho, 2016). In addition, currently, most of the app designs seldom consider the barriers of older adults, who are the potential beneficial users of mHealth apps (Wildenbos, Peute, & Jaspers, 2017).

2.2.2 Persuasive mHealth apps which incorporated the Health Behavior Change techniques

In addition to evidence that the commercial apps should have offered a better user experience to users, there is also a threat of the limited effectiveness of commercial apps (Hingle & Patrick, 2016). Most of commercial mHealth apps are reportedly not using any theoretical strategies to support health behavior change (Azar et al., 2013). Behavior change techniques based on Abraham and Michie (2008)'s taxonomy approved by health behavior change theories, have been recommended for use in mHealth apps design (Bardus, van Beurden, Smith, & Abraham, 2016; Flaherty, McCarthy, Collins, & McAuliffe, 2018). Techniques which are frequently used in healthy eating interventions including providing information about behavioral health link, other's approval, and information on consequences; prompt intention information and barriers identification; prompt specific goal setting, self-monitoring and review; provide instruction and general encouragement; set the graded tasks; model demonstrate the behavior (Abraham & Michie, 2008). Hales et al. did a content analysis for the commercial picture-based diet tracking apps, they found that few commercial apps incorporate an evidence-based health behavior change strategy (Hales, Dunn, Wilcox, & Turner-McGrievy, 2016).

In the realm of academically designed apps, Kankanhalli et al. (2019) reviewed the HCI literature before 2018 of dietary behavior interventions incorporating with behavior change techniques. They found 30 studies, all of which facilitated the self-monitoring feature; 18 studies provided personalized feedback (reminder, or recommendation); 10 studies used gamification goal reviews in 5 studies, social support in 3 studies, and educational information in 2 studies. When evaluating the effectiveness of the app, 13 studies evaluated health outcomes and 12 studies evaluated dietary behavior change, although the measures of dietary behavior change were inconsistent (Kankanhalli et al., 2019).

However, there's a limited discussion about the practical health behavior change techniques incorporating a nudge design. Nudge has also been considered as a health behavior change theory ever since Thaler and Sunstein published a book discussing its impact in health policy. Researcher have also proposed nudge as a promising public health strategy towards fighting against obesity by changing an individual's dietary choice and behavior (Arno & Thomas, 2016). But there's limited empirical studies and practical applications in the real-world setting, especially the applications of nudge approaches to persuasive technology.

2.3 Theoretical Basis of Usability Engineering Frameworks for Persuasive mHealth apps

To propose a framework for developing an effective and efficient persuasive mHealth app, concepts from different theories were integrated with Neilsen's usability engineering framework. This section reviews those theories that were included: decision-based design framework, which is the theoretical basis of adapting a systematical engineering design approach to make data-driven design decisions in this study; service engineering framework brought in the value of the co-creation idea as the theoretical basis of participatory design; software engineering framework guided the software product (the persuasive mHealth app) development process; persuasive design framework is the theoretical basis for developing persuasive features; signal detection theory served as the theoretical basis of the proposed human factors evaluation method for persuasiveness.

2.3.1 Decision-Based Design Framework

Decision-Based Design (DBD) is a terminology to describe the parental framework of the engineering design approaches which are popular in 1980s such as Taguchi's Robust Design, Quality Function Deployment...etc. This unifying framework was proposed by Dr. Hazerlrigg from NSF in 1998, The key idea is to view the engineering design as a decision-making process, allowing the theories and the methods from economics, operation research, and decision science to be used to optimize the design in an enterprise resource allocation viewpoint. According to Hazerligg (1998), "*It forced the process of engineering design into a total system context and demands design decisions account for product's total life cycle.*"

2.3.2 Service Engineering Framework

Service is one of the basic constructs of exchange and thus it plays an important role in economies (Vargo & Lusch, 2004). According to Maglio (2015), Human Centered Service Systems (HCSS), which focus on human interaction and personal service, translate the economic relevance to the individual level while covering the essential areas for our society such as education or healthcare (Kleinschmidt, Peters, & Leimeister, 2016; Maglio, 2015).

The modern service system is evolving from goods-dominant logic to service-dominant logic. Under the S-D logic, the consumers are not only recipients of the service, but they would also co-create the value to the system. In a Human-Centered Service System (HCSS), the personal interaction between the actors is the essential component of value co-creation. With the prevalence of service-dominant (S-D) logic and the proliferation of connected Smart Things, there would be a "smart" service innovation by reconfiguring the current HCSS with the human computer interaction as a Human-Centered Smart Service System.

2.3.3 Software Engineering Framework

System Development Life Cycle

System Development Life Cycle (SDLC) is a critical system engineering framework to describe the planning and creating process for an information system. It's also used as the "Application Development Life Cycle" in software engineering discipline. Basically, there are 5 working stages in the SDLC, planning, analysis, design, implementation and maintenance. The modern computer system could be complex while integrating many different traditional systems. Based on the complexity of computer systems, there are two different SDLC models developed to manage different level of system complexity, the waterfall model and the Agile software development model.

Waterfall is the oldest SDLC model, firstly defined by Royce in 1970, and refined for the system with a higher-level complexity by Boehm in 1976 (Boehm, 1976; A. M. Davis, Bersoff, & Comer, 1988; Royce, 1987). The classic waterfall model is as shown in the Figure 6. Companies vary in what they consider standard methodology and in naming conventions for various stages. The first stage would be the user needs analysis, typical names as requirement analysis, system analysis or specifications, to see the system requirement and then break down to the software configuration. Second is the preliminary design stage, which is often called high-level design, top-level design, software architectural definition, or specifications. Next is the detailed design stage which is often called program design, module design, lower-level design, algorithmic design, or just plain design. When the design phase is done, the next stage is implementation with code, debug, testing, and operation. And the final phase is maintenance.

The disadvantage of traditional waterfall model is the rigid boundaries of each stage in a sequential structure, and the documentation-driven standard makes the transition stiffened. And thus, the overall SDLC of waterfall model would in average take more than 9 months and the outcome of the product is usually not that user-friendly. With the growing need of shortening the SDLC, in 1980s, researcher proposed the iterative model. Don Norman et al. combined the user-centered design concept with an iterative design model to kick off the design process based on users' needs and to create the prototypes and to evaluate the UX quality based on the user testing.

Bohem further (1988) developed an iterative model of Spiral software development product cycle. In recent year, a new SDLC model, agile software development was proposed to shorten the system development life cycle. With the development of the object-oriented language, the software product with higher complexity could be broken down into smaller builds, and thus the Agile SDLC model is a combination of iterative model and incremental process.

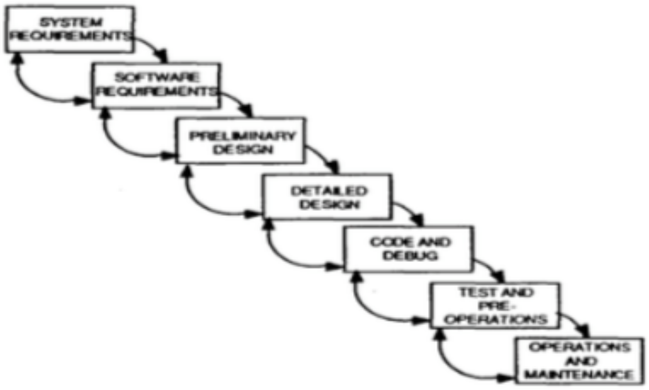


Figure 6. Waterfall SDLC model

Source: Boehm, B. W. (1976). Software Lifecycle Model. Software Engineering.

Requirement Analysis

Requirement analysis is an important phase of the software development life cycle, it studies, determines, and documents the user expectation and needs (Catanio, 2006). Traditionally, the practitioner would employ the market research methods such as personal or focus group interview, which is a subjective research method that needs a skillful interviewer to design the questions and conduct the interview, and it would be difficult to translate the data without the human bias. Further, the customer may not really know what they need, especially for innovation products (Von Hippel, 1986). The modern user research methods including the observation in the natural environment such as ethnographic study and contextual inquiry, however, the natural observation has its disadvantages such as Hawthorn effect, difficulty to access users under the specific environment, and thus it could cost time, money and huge labor effort to finish an observation in the natural environment. As a result, requirement analysis is often skipped by many software organizations to save time and money. Currently, most commercial apps have a low user retention rate, the reason could due to the lack of the consideration of user's need (J. Cho, 2016).

Object-Oriented Analysis and Design

Object-Oriented Analysis and Design (OOAD) is a method for objective-oriented programing to develop the conceptual model by use cases, which are the documentation of the target user groups and how would users use the functionalities in the specific circumstances. The most common output document of the OOA is the use cases diagram drawn by the Unified Modeling Language (UML) form.

To digitalize a human-centered service system, the interactions between the agents should be analyzed by human factors evaluation methods to draw the use cases diagram. There are two kind human factors evaluation methods of task analysis methods that are frequently used in the Human Computer Interaction (HCI) context, Hierarchical Task Analysis (HTA) and Cognitive Task Analysis (CTA). HTA follows the theoretical human cognitive model to break down a complicated job into tasks and lower level subtasks, and put them into a hierarchy, and thus the human behavior could be systematically analyzed by the task performance and error rate. The observation of human error is also helpful to redesign a reasonable task. However, HTA is simply focused on the tasks, which is task oriented. And CTA is more user oriented, it focused on analyzing human's cognitive model. To perform the CTA, the human-centered testing data is usually collected by think aloud protocols including the concurrent think aloud protocol and Retrospective Think Aloud (RTA) protocol. The former protocol asks the participants to talk about what's in their mind when they are performing the task. The second protocol records the video while the participants are performing the task, and after the task has been finished, the experimenter replays the tape and asks the participant talk about what's in their mind at that time based on the video record. There are many kinds of concurrent think aloud methods, the easiest one to implemented is the cognitive walkthrough, which is a combination of Hierarchical Task Analysis and the concurrent think aloud protocol.

2.3.4 Persuasive Design Framework

Since Fogg's persuasive technology concept were proposed in 2003, several persuasive design frameworks and persuasive design principles have been suggested to facilitate the concept. Fogg (2009) proposed the Fogg Behavior Model (FBM) to explain that human behavior change is driven by three endogenous psychological factors of motivation, ability, and trigger together. He further suggested the analysis and design should be based on this model (B. Fogg, 2009). Oinas-Kukkonen & Harjumaa (2009) suggested the practical heuristics to analyze persuasion based on the context of intention, event, and strategy. The critical design principles were categorized by the design context of the behavior change support system quality, including the primary task support, dialogue support, system credibility support, and social support (Oinas-Kukkonen & Harjumaa, 2009). Murillo-Munoz (2018) integrated Fogg's and Oinas-Kukkonen & Harjumaa's work and extended the application to the mobile systems as a design framework with three stages: (1) understand the key issues behind to identify the purpose of persuasive design; (2) Analyze the context; (3) Design of the system quality based on Oinas-Kukkonen & Harjumaa's heuristics (Murillo-Munoz, Vazquez-Briseno, Cota, & Nieto-Hipolito, 2018). Taype and Calani (2020) summarized persuasion principles from the literature and translated them to the design language and conducted cluster analysis to classify the design principles into four groups including "Simplifying tasks", "Usability", "Credibility", and "Social Influence" (Taype & Calani, 2020).

However, although the above frameworks have already defined the guidelines of persuasive design, there is still a gap between design and engineering. For example, although design principles could be used to conduct heuristics evaluation in the very beginning concept and design stages, but there still could be a gap between expert and user's viewpoint. There remains a need for human factor evaluation methods and quantitative metrics for user testing study to support systematic design decision-making during the product development stages.

2.3.5 Human Factors Evaluation for Persuasiveness

Currently, there are limited human factors methods that were proposed for evaluating persuasiveness of behavior change support systems. The effectiveness of such systems are mostly evaluated by health outcomes which are indirect measures of persuasiveness (Kankanhalli et al., 2019).

Lehto et al. (2012) developed a survey instrument to measure perceived persuasiveness by the self-rating questions on a Likert scale. A survey about persuasiveness of a web-based behavior change support system was conducted with 172 users using the instrument. It was found that perceived persuasiveness was positively affected by primary task support, dialogue task support, and system credibility. It is also positively influenced the intention of use but not actual usage (T. Lehto, Oinas-Kukkonen, & Drozd, 2012). De Jong et al. (2014) conducted expert reviewed heuristics evaluation for a Nurse Antibiotic Information App project based on Oinas-Kukkonen and Harjumma's framework and a scenario-based user testing study with 62 nurses using Lehto et al.'s instrument. They found the expert-assessment aligned with the user testing results (De Jong, Wentzel, Kelders, Oinas-Kukkonen, & Van Gemert-Pijnen, 2014). Still, all of the above studies are mainly based on subjective measures which may be biased by individual differences of rating strategies in many research setting. For example, older adults respondents who tend to respond in a socially desirable way (Dijkstra, Smit, & Comijs, 2001; Fastame & Penna, 2012; Ray, 1988). There is a gap of research about objective measures of persuasiveness for user testing studies.

Signal Detection Theory

Currently, persuasiveness is usually evaluated based on the proportion of human correct responses to the information interpretation questions. This metric is subject to the human bias of guessing yes/no with a liberal or a conservative criterion. To narrow the research gap, I proposed to evaluate the persuasiveness based on human performance in a decision-making scenario (healthy food choice in the dietary behavior change context). Human decision performance was measured based on Signal Detection Theory (SDT). SDT was proposed in 1960s and widely accepted as an experimental psychology method of analyzing human responses of detecting a signal from a noisy environment. Signal Detection Theory (SDT) provides a framework to interpret the detection experiment result. It assumes the sensory evidence of the signal could be represented as a continuum. The amount of evidence varies at each trial with the presence of noise. Even when the noise is presented alone, there is a little amount of evidence of the signal presence. When the amount of evidence reaches the observer's decision criterion, it triggers the observer's response of "yes". The response criterion is the key metric to evaluate the tendency of an observer to answer "yes" or "No". And the general approach of analyzing human responses is also applicable to other decision problems with uncertainty, such as discriminating between two different type of stimulus (Green & Swets, 1988).

In a typical signal detection experiment, the presence of the target stimulus is a "signal" trial, otherwise is a "noise" trial. The observer responds to a series of trials with the random presence of signals by either saying "yes, it is a signal" or "no, it is a noise". The responses were counted in four conditional categories and analyzed based on statistical decision theory with the terms: hit (true positive; TP), false alarm (false positive; FP), miss (false negative; FN), and correct rejection (true negative; TN). The discriminability between signal and noise is then defined as the sensitivity of a signal detection experiment. The higher discriminability means the better this signal is detected (Chao, Lehto, Pitts, & Hass, 2021).

Signal detection analysis assumes the distributions of signal in the presence of noise and noise trials follow the Gaussian distribution. So, human response criterion could be drawn on the diagram and the areas on the right-hand side of criterion are human responses of "yes", on the left-hand side of criterion are human responses of "no". So the areas under the normal distribution curves and criterion represent hit rate, FA rate, correct rejection rate, and miss rate. And then, the discriminability, d-prime (d') = Z(Hit Rate) – Z(False Alarm Rate), could be calculated by hit rate and false alarm rate.

When the healthy food is regarded as the signal in a signal detection experiment, human response information quality could also be measured by discriminability, d', which gives us an objective measure of UI design quality.

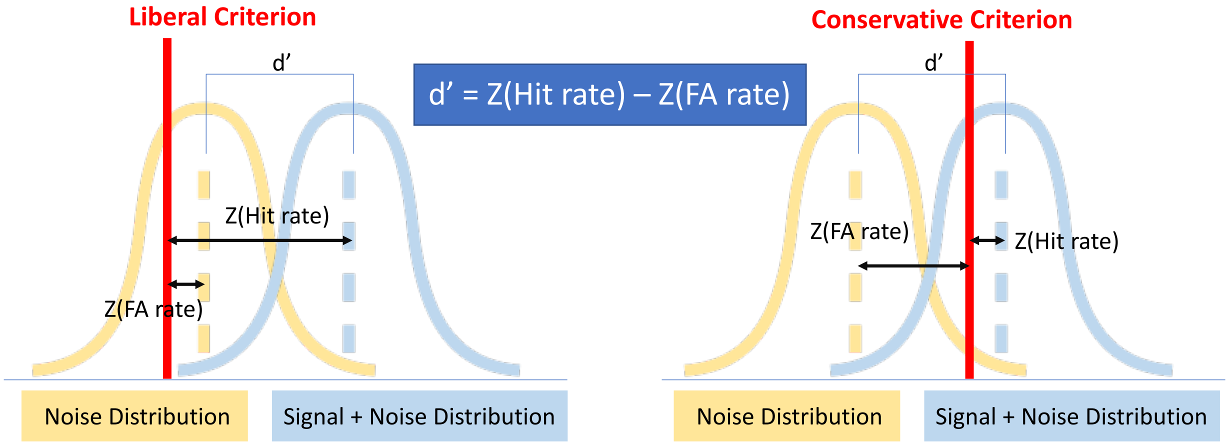


Figure 7. Signal Detection Theory

In this article, I proposed modeling the choice of healthy food based on Signal Detection measures. The healthier food alternative (out of 2 options) is regarded as "signal", and the other alternative is regarded as "noise". When the decision maker selects a healthy food choice which follows the FDA 2015-2020 Dietary Recommendation guidelines, it is regarded as a "hit", otherwise it's a "miss. Preposition of the food choice experiment: If the intervention is successful, the user's sensitivity would get higher and the response criterion will be closer to the neutral.

2.4 Persuasive Nudge Design Elements for Dietary Management Apps

In this dissertation, the persuasive design framework is followed to develop strategies and practice nudge approaches. Context analysis was conducted based on the user task flow chart developed from hierarchical task analysis (See the detail in chapter 4.) Opportunities to practice nudge approaches revealed from that analysis included 1. Manipulating the layout design of the choice-based interface, which facilitates the digital nudge idea; 2. Manipulating the provided information, which practices information nudge; 3. Manipulating the provision of default selection, which practices default nudge. To better implement nudge approaches, below sections review the literature for the related design elements potentially affecting user experience.

2.4.1 Theoretical Basis of Nudge Approaches

As previous sections have mentioned, practical nudge approaches for persuasive mHealth apps are underdiscussed, despite the fact that they are potentially powerful and effective approaches with a strong behavioral science theoretical basis. In consequence, this section reviews the related works of nudge approaches to better weave the ideas into the persuasive design.

Nudge Design for Dietary Behavior

Bauer and Reisch (2018) systematically reviewed the literature of changing dietary behavior using nudge from 2011-2017. They categorized the article into 5 kinds of nudge, adopting the themes from Perry et al. (2015) : (1) Information Nudge: providing nutrition information when people make decisions to choose food, e.g. caloric information of the meal on the restaurant menu, or nutrition label on the packaged food (Bleich et al., 2017; Campos, Doxey, & Hammond, 2011); (2) Social Norm Nudge: social support, peer pressure, or the social norm to help with decision making according to Bandura's social cognitive theory (Bandura, 2001; Robinson, Thomas, Aveyard, & Higgs, 2014); (3) Default Nudge: making healthier options as the default choice.; (4)Physical Environment Nudge: changing the location, presentation, and composition of the physical environment to make the healthier alternative salient to the decision maker. e.g. change the healthy food location in supermarket; change the plate size or color of dinning environment etc... (Bucher et al., 2016; Holden, Zlatevska, & Dubelaar, 2016); and (5) Incentives Nudge: give rewards for healthier food choice (Harden, Peersman, Oliver, Mauthner, & Oakley, 1999; Hillier-Brown et al., 2017). However, there is a research gap around default nudge since it's difficult to implement in the real-life scenario for the general population (Bauer & Reisch, 2018; Perry, Chhatralia, Damesick, Hobden, & Volpe, 2015).

Nudge in Digital World

As an increasing number of healthcare researchers and practitioners recognize mHealth apps as a persuasive technology to facilitate the health behavior change. Nudge design and implementation on the smart agents have also grasped the eyes of the HCI practitioners.

Caraban et al. (2019) systematic reviewed 71 articles regarding to the technology-mediated nudges in HCI domain areas between 2008 and 2017. They summarized 23 mechanisms (what approaches) and their basis in human cognitive bias (how to nudge) to provide nudge design insights for designers. However, there's still a gap of an overall knowhow about the effectiveness for each nudge approaches. Currently, most proposed nudge approaches are transparent methods or based on human's automatic mind according to Caraban et al.'s classifications based on Hansen and Jespersen's typology (2013) (Caraban, Karapanos, Gonçalves, & Campos, 2019; Hansen & Jespersen, 2013), Most of them are the external techniques adapted from some other empirical paradigm. There's a research gap of technology-mediated nudge approaches designed based on the implicit redesign of the choice architecture in the reflective mind, which is labeled by Hansen and Jespersen as "choice manipulation" and thought as the essential of nudge by Thaler and Sunstein. In other words, so far, there's a research gap of the empirical approaches of Weinmann' et al.'s digital nudge.

2.4.2 Digital Nudge Design

Digital layout of decision-making is the main concern of the digital nudge design in this study since it directly structures the decision architecture. Based on the task analysis of dietary decision-making process in a digital world, searching UI and decision paradigm are potentially critical design elements, and thus the related works are reviewed.

Search UI design

Search User Interface (SUI) design determines how users interact with the Information Retrieval (IR) and Recommender System (RS) or the database. And according to the task analysis of making online food choices, users must firstly perform the searching task to inquire the related web contents and needed information to support decisions, and the system would return the results on

SUI (See the detail in chapter 3.) for users' selection. So, the presentation of searching results (e.g. the search results layout, formats, information architectures, …etc.) may make an effect of nudge. In the following section, related works about SUI design are reviewed to analyze the critical nudge design elements.

Since early 2000s, the evolution of the search engine has changed the way humans interact with the information retrieval system. Nowadays, people are used to searching for a keyword on a simple search box rather than performing the query search on a command-line system. (Chao & Hass, 2020). Ever since Google dominated the search engine market as the most widely used landing homepage of internet browsers, its UI has become the SUI paradigm. Relatively better usability and learnability of the Google SUI made it stand out from competitors. Sayago & Blat (2007) have reported the Google search engine UI is easier to use and learn for older adults and novice users compared to other search engine UIs at that time such as Yahoo! and MySpace (Sayago & Blat, 2007). Despite this, Aula (2005) argued that the information architecture was confusing to older adults and caused some usability problems that still occur due to the information architecture may be a new idea to some older adults (Aula, 2005).

The Google SUI has defined the modern SUI, and so nowadays, we could see most SUI designs follows the same layout and features including search box input, interactive query change control (e.g. auto-fills and auto-corrections), standard results list, and personalization features (Wilson, 2011). However, the evolution of SUIs should still be continued to accommodate the fast-paced development of new technologies. For example, to adapt to the limitation of smaller screen size on mobile devices, designer could consider the design of facet search (Karlson, Robertson, Robbins, Czerwinski, & Smith, 2006; Kleinen, Scherp, & Staab, 2014). Faceted metadata search defined the search space by allowing users to apply filters to the metadata of searching results (Wagner, Tran, & Ladwig, 2011; Wilson, André, & Schraefel, 2008). Compared to the keywords search, this kind of dynamic interactive system has been found to improve the search experience of exploratory search type tasks (Stoica & Hearst, 2004). What's more, the facet search system facilitates Bates' "Berrypicking" browsing model. The Berrypicking browsing model describes searchers' natural behavior based on the elimination by heuristics to effectively collect needed information from various resources (Bates, 1989; Chao & Hass, 2020).

Decision-Making Paradigms

Two kinds of decision-making paradigms, which describe the common human decision-making scenarios are frequently used in human factors experiments. The first one is the yes/no paradigm, which is also the experiment paradigm of SDT. The respondent decides the existence of target stimuli and makes an yes/no response. Another paradigm is the Two-Alternatives Forced Choice (2AFC) comparison paradigm. 2AFC comparison paradigm is a classic cognitive psychological experimental design. It randomly presents both alternatives on every trial in spatial or temporal order. The sensitivity & the bias could still be estimated by signal detection analysis. In empirical studies, accuracy in yes-no test paradigm tends to be lower than that in 2AFC test paradigm, based on the prediction of detection theory. (MacMillan & Creelman,1991; 2004)

2.4.3 Information Nudge Design

Interpretive Nutrition Label Design

Providing nutrition information is a widely accepted cost-effective method to nudge consumer's food choices and subsequent dietary behavior. The nutrition label is one of the vehicles provided on the packaged food to support a consumer's decision (Bauer & Reisch, 2018; Bleich et al., 2017; Campos et al., 2011).

In the United States, the Nutrition Facts Panel (NFP) is the standardized format Back-of-Package (BOP) nutrition label regulated by the government (Food and Drug Administration; FDA) for packaged foods through the Nutrition Labeling and Education Act of 1990. The NFP displays information including serving size, number of servings, total energy, and a selection of nutrients such as energy from fat, total fat, saturated fat, trans fat, cholesterol, sodium, carbohydrates, dietary fiber, sugar, protein, vitamin A, vitamin C, Vitamin D, calcium, and iron (Chao et al., 2021). As the risks of transfat and sugar were found in the related evidence-based research, the transfat information was included in a 2006 amendment to the legislation. In a 2016 revision, added sugar was included in alignment with the 2015-2020 Dietary Guidelines for Americans: added sugar intake to less than 10% of daily calories (Malik, Willett, & Hu, 2016).

With the above revisions, nowadays, the NFP is not only a label for packaged food but also a widely used tool to support healthy behavior change. Practitioners train patients with certain health conditions, including type II diabetes, hypertension, etc. to use the NFP to guide their healthy food decision making. However, researchers doubt the effectiveness of providing the NFP to nudge, since studies often found BOP labels are ignored by consumers and the NFP is incorrectly interpreted (Talati et al., 2016). In recent years, many food manufacturers, marketers, and policy makers lean toward the idea of designing a comprehensible Front-of-Package (FOP) label. A system review on the impact of FOP labels found evidence to support this viewpoint. Cecchini and Warin (2016) conducted meta-analysis of randomized studies between 2008 and April 2015 and reported FOP labels increased nearly 18% of the subjects choosing a healthier food product and decreased 3.59% of caloric choices for an individual. (Cecchini & Warin, 2016).

FOP labels have already been developed in many European countries for years, including health claims, Guidelines Daily Amounts (GDA), traffic light system, and Nutri-scores (Hodgkins et al., 2012). The GDA label is a nutrition label developed in the UK around 1998, which could be viewed as a simplified NFP with the healthy guidelines for the general population. It lists only five key nutrients: calories, fat, saturated fat, sugar, salt with the absolute amount per serving and the percentage of daily value. Another kind of FOP label presents interpretative information on the symbolic display, for example, the traffic light system. The traffic light label used traffic light signal colors to code the expert rated healthiness by the high-, medium-, and low-level. A symbolic interpretative FOP label, Nutri-scores, has recently been selected as the official nutrition label for packaged food in France and widely accepted in other countries since 2017. The Nutri-scores label combines the weighted sum of the GDA (which is calculated by UK Food Standards Agency as the FSA scores) with a 5-point system using the letters A, B, C, D, E (from the best to the worst) and the traffic light style color-coding (Chao & Hass, 2020).

Among those labels, interpretative FOP labels such as traffic light system or Nutri-scores which directly presents the judgements on food healthiness to users were preferred by authorities. Those labels were regarded as a cost-effective information nudge approach to promote healthier food products. For example, in 2017, the World Health Organization (WHO) recommend to provide interpretative FOP labels as the official "best buys" suggestion to prevent dietary-related noncommunicable disease (World Health Organization, 2017). Egnell et al. (2018) further found that human subjects can better understand symbolic Nutri-score label than the monochromic GDA labels (Egnell et al., 2018). In this work, the FSA Nutri-scores label was chosen in comparison with the FDA Nutrition Facts Panel as the baseline to verify WHO's recommendation of using interpretative FOP as information nudge.

2.4.4 Default Nudge Design

Default nudge is viewed as the paradigm of nudge since it was found effective in many contexts, for example, to persuade the general public to adopt a healthy policy by using a strategy that defaulted individuals into the participating option (Thaler & Sunstein, 2009). The approach has also been found useful in influencing dietary behavior. Several studies have proposed the default nudge approach and confirmed the effectiveness of default nudge in dietary control context. For example, Friis et al. (2017) pre-proportioned the salad with the recommended amount of vegetables in the self-serving buffet setting to promote the vegetable consumption, and found the default nudge is more effective than priming or visual presentation of more varieties (Friis et al., 2017). Van Kleef e al. (2018) defaulted the bread type of sandwich in a sandwich choice experiment and found more than 80% of the study participant stick to the default choice (van Kleef, Seijdell, Vingerhoeds, de Wijk, & van Trijp, 2018).There is, however, limited applications of digital nudge in the mHealth app setting. For that reason, in this dissertation, I simply follow the paradigm to default the healthier food option by system pre-selection of the radio button.

3. EXTENDED USABILITY FRAMEWORK FOR PERSUASIVE TECHNOLOGY

As technologies are getting smarter (by applying machine learning algorithms / artificial intelligence to offer a higher level of cognitive support), computer's role is also changing from a tool to a social actor and make a greater impact to human society. In this context, the mindset of developing persuasive technologies is also evolving from considering persuasive strategies as value-added features to implementing persuasion as the core value of the final product. However, the current persuasive design framework is merely focused on developing persuasive features in design stages but not the entire product development process. There is a need of engineering design approaches to bridge the persuasive design and the software engineering. The current usability engineering framework serves the need to bring together design and engineering but not specific to persuasive technologies, since there's a research gap of human factors evaluation methods for persuasiveness.

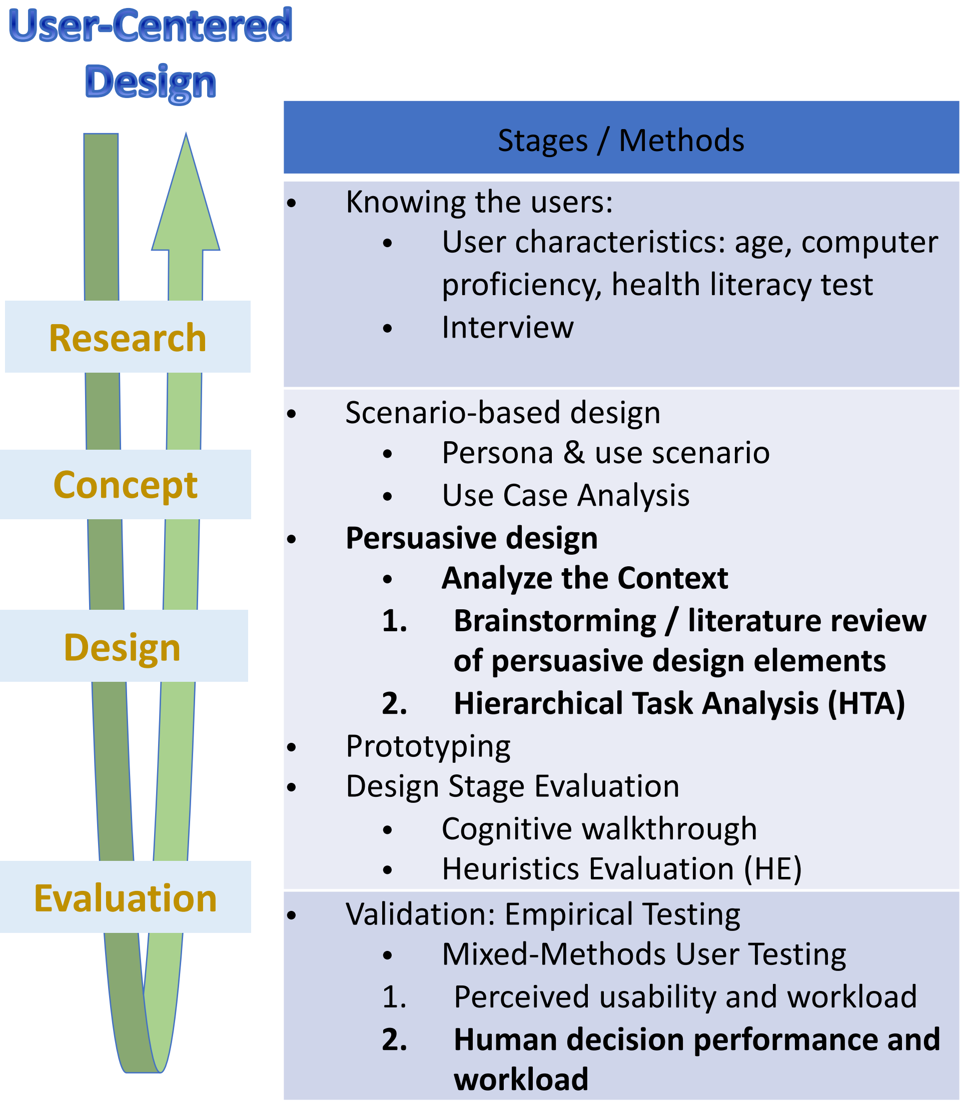
This chapter proposes an extended usability engineering framework for persuasive mHealth apps (Section 3.1) as a holistic approach to develop persuasive technology. The theoretical framework is based on Nielsen's usability engineering framework (Section 1.2.2) and Oinas-Kukkonen & Harjumaa's persuasive design framework (Section 2.3.4).

3.1 Theoretical Framework

In this dissertation, an extended usability engineering framework for persuasive technology is proposed which integrates persuasive design research methods with Nielsen's framework.
Figure 8 shows the proposed framework. The framework inherits the structure of the iterative user-centered design lifecycle with four stages: research, concept, design, and evaluation. In this framework, a design project begins by getting to know the intended users; the related user research methods used in this dissertation for this stage include survey, interview, and literature review. The next part of the design reflects the integration of Oinas-Kukkonen & Harjumaa 's persuasive design framework with the user-centered design approaches (i.e. persona, use scenario) in the concept and design stages. The first step of Oinas-Kukkonen & Harjuma's framework is to analyze

the context of persuasion. Oinas-Kukkonen & Harjuma (2009) have suggested that the analysis should include the recognition and the understanding of the intent, the event, and the strategy. In this dissertation, I propose to draw the user task flow chart based on the use case analysis and the task analysis to facilitate the context analysis. Use case analysis and task analysis could be the practical methods to recognize the intent of persuasion to better understand the persuasive event. The persuasive strategy could be either be formed by brainstorming or identified from literature review. The user task flow chart could be used to identify the timing of when to apply the persuasive strategy. For the purpose of validating design ideas, expert review methods such as cognitive walkthrough and heuristics evaluation could be firstly conducted in the design stage.

In the evaluation stage of this framework, usability of the proposed persuasive design is further verified by empirical user testing study in terms of effectiveness, efficiency, and satisfaction. The frequently used mixed-methods usability testing measures in Neilsen's framework include survey, human task performance measures (task completion time), interview, and think aloud protocol are still adapted to evaluate efficiency and satisfaction. To evaluate effectiveness, objective human decision performance and workload measures are proposed based on Signal Detection Theory (SDT; which has been mentioned in 2.3.5.). The proposed human factors method for persuasiveness is the key to better integrate the persuasive design framework with to Neilsen's usability engineering framework. The detail research framework and the experimental protocol is described in the next section and chapter 5.



Stages / Methods

- Knowing the users:
  - User characteristics: age, computer proficiency, health literacy test
  - Interview

- Scenario-based design
  - Persona & use scenario
  - Use Case Analysis
- **Persuasive design**
  - **Analyze the Context**
  1. **Brainstorming / literature review of persuasive design elements**
  2. **Hierarchical Task Analysis (HTA)**
- Prototyping
- Design Stage Evaluation
  - Cognitive walkthrough
  - Heuristics Evaluation (HE)

- Validation: Empirical Testing
  - Mixed-Methods User Testing
  1. Perceived usability and workload
  2. **Human decision performance and workload**

Figure 8. Theoretical Framework

3.2 Pilot UX Design Project of a Smart Healthy Food Recommender System

Currently, most of the mHealth apps are designed based on the practitioner's viewpoint to extend the healthcare delivery process by utilizing the portability and the mobility of the information technology. For example, the majority of the mHealth tools target to provide the function of monitoring patient's health behavior at home by tracking food intakes or physical activities. However, this kind of mHealth app seldom hooks the consumer as the lower user retention rate has been reported (Jaehee Cho, 2016).

According to the 2013 Pew Study of the Quantified-Self movement, 69% of adults would track with their own health-related measures such as weight, body circumstances, and food intakes. However, half of them just keep rough numbers or attributes in their head, seldom do they use technologies. Kwon et al. (2017) reported the finding that elderly adults are reluctant to use mHealth apps until they were forced to do so, for example, "prescribed" by a doctor (Kwon, Mun, Lee, McLeod, & D'Angelo, 2017). However, older adults are particularly interested in nutrition information (Guo, Sun, Wang, Peng, & Yan, 2013). Sanjari et al. (2017) did the literature review about the customer's attitudes and responses to the front-of-package (FOP) nutrition label from 1990 to 2016. They found that compared to younger adults, older adults would pay more attention to nutrition labels for packaged food when doing the grocery shopping. When making a food choice, more older adults take nutrition information into consideration (Sanjari et al., 2017). This finding is aligned with our general impression obtained from general observation and informal conversations with older adult families and friends that older adults expect that mHealth app technology could help them enhance their interaction with the nutrition information retrieval process.

To better serve the requirements of both the practitioner and the real user, a smart food decision support system and its choice-based user interface of an mHealth app was designed and developed by the User-Centered Design (UCD) process. The design project was kicked off by user research. And then, the persona and use scenario are drawn based on the user research results. In this chapter, a prototype of the system was created and roughly evaluated by the cognitive walkthrough of the designer to make the design decision for the use cases. The iterative design process is continued with making more design decisions of critical UI design element based on the proposed user testing researching project in the next chapter.

3.2.1 User-Centered Design of a Dietary Management App

The needs of older adults were collected from literature, informal conversations and observations in several social service events in public; and were synthesized as the persona and use scenario to inspire the design. Use case analysis was done to determine the technical requirements and task analysis was used to optimize the user task flow.

Persona & Use Scenario

Sanjari et al.'s finding (2017) about older adults pay more attention to nutrition information during grocery shopping was further developed into a persona, which is the profile of a virtual user, Serena. Below is the detail user story we created for Serena:

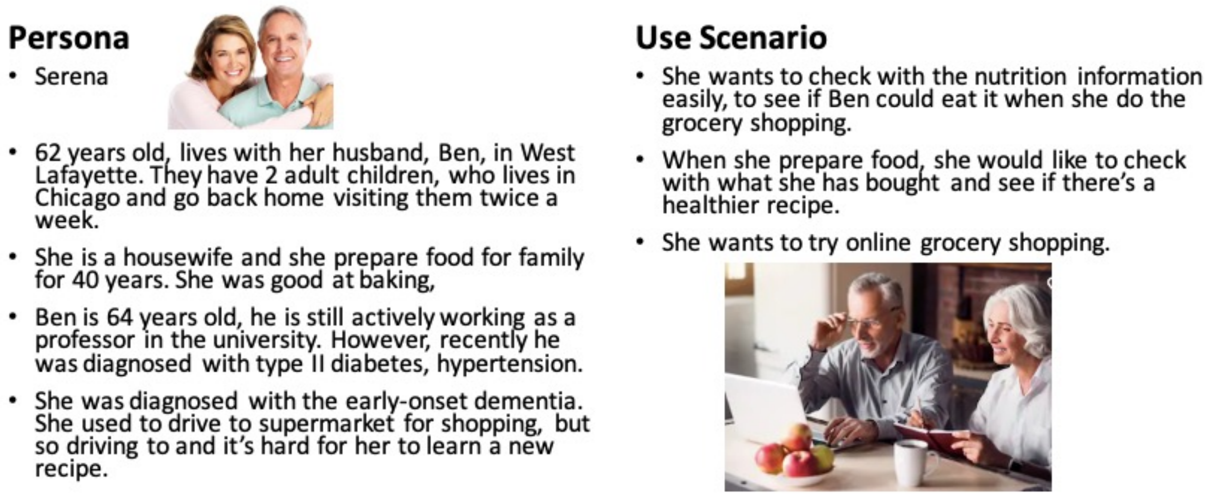
"*Serena is a 62 years old woman, she lives with her husband, Ben, in West Lafayette. They have 2 adult children, who live in Chicago and go back home visiting them twice a week.*

She is a housewife and she has prepared food for family for 40 years. She is especially good at baking. Fresh fruits, fine sugar and butter were the most important food ingredients in her recipes.

Ben is 64 years old. He is still actively working as a professor at the university. However, recently he was diagnosed with type II diabetes and hypertension. The doctor set up food restrictions with daily intake of sugar and sodium.

Serena was diagnosed with early-onset dementia. It's hard for her to learn new recipes. What's more, she used to drive to the supermarket, but she is now having trouble driving independently.

After the persona was created, the use scenarios of the system were drafted based on the persona following the User-Centered Design (UCD) principles. The persona may need to use the system when she does the grocery shopping (checking nutrition information for acceptability against husband's diet) and the meal preparation (search for healthier recipes). Thus, the system is redesigned based on these use scenarios.



Persona

- Serena
- 62 years old, lives with her husband, Ben, in West Lafayette. They have 2 adult children, who lives in Chicago and go back home visiting them twice a week.
- She is a housewife and she prepare food for family for 40 years. She was good at baking,
- Ben is 64 years old, he is still actively working as a professor in the university. However, recently he was diagnosed with type II diabetes, hypertension.
- She was diagnosed with the early-onset dementia. She used to drive to supermarket for shopping, but so driving to and it's hard for her to learn a new recipe.

Use Scenario

- She wants to check with the nutrition information easily, to see if Ben could eat it when she do the grocery shopping.
- When she prepare food, she would like to check with what she has bought and see if there's a healthier recipe.
- She wants to try online grocery shopping.

Figure 9. Persona & Use Scenario

Use Case Analysis

To ensure the design language could be successfully translate to the engineering language, use case analysis was performed. The designer has considered several use cases based on the persona and use scenario and drawn the use case diagram. And then, I further considered the data flow and the information architecture to draw the Unified Modeled Language (UML) type class diagrams for the further system implementation. Based on the feasibility of the system implementation, the use cases of the proposed mHealth app is determined by the diagram as shown below.

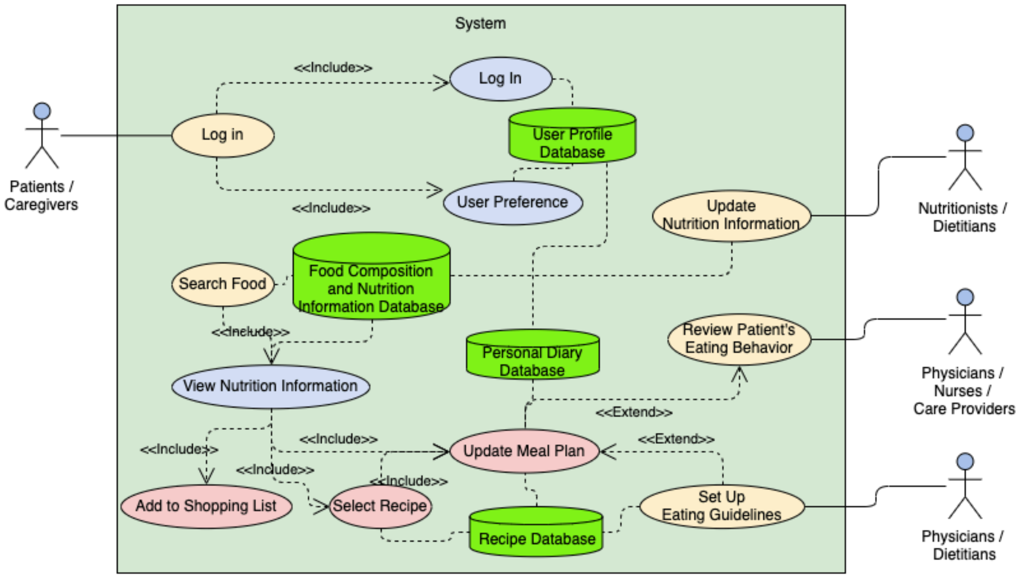


Figure 10. Use Case Diagram of the proposed dietary self-management app.

Hierarchical Task Analysis

In addition to performing the use case analysis to validate the system feasibility, the designer has also performed the cognitive walkthrough to consider the user task flows in the user-centered viewpoint to better determined the ideal use case of recording an entry on food diary.

The hierarchical task analysis was firstly performed to define the user task flows for three different kind of use cases of recording an entry on the food diary. The first one is writing the diary from the scratch, which is a task designed inspired by the daily life activities of writing paper notes and diary for some older adult friends. Although it may be a common scene for some older adults to write a diary which means the process flow could be easily learned by users, however, the tedious workload of data input for writing a digital diary may harm the usability tremendously. The second use case is adopted by most of the competitors who have provided the similar diet tracking functions on mHealth apps. It reduces the data input loading of users by the pre-defined template for dietary assessment such as Food Frequency Questionnaire (FFQ), Dietary Record (DR), or 24-hr dietary recall food record method. Those tools are some existing instruments developed by domain experts to collect detail information of consumed food and beverage in 24 hours for a respondent in the real-world healthcare setting. To serve the dietary assessment purpose, it guides the respondent not only to recall the eaten item and the time of consumption but also to think through the detail of the item including the ingredients and the serving size. However, from the user-centered viewpoint, the respondents actually have to perform the recalling and recording task repetitively to finish the questionnaire. It requires higher mental workload and more human errors would be introduced to the process. Although as it has been mentioned in the chapter two, there are several related technologies were developed to support this use case, which may help with reducing user's workload to perform the task and prevent potential human errors. But it is still difficult to persuade users to regularly do the task since fill in the questionnaire is not what users need and there may be no immediate and visible paybacks to motivate them to do it.

In this design project, the designer has thought beyond the box to redefine the use case of recording the food diary based on users' needs. Based on the user desired scenario of grocery shopping and meal planning, the system would be used as a decision support tool to provide real-time information. In this context, the major task flow would be defined by the searching and selecting task, and the goal of recording the entry on the food diary could be accomplished by recording the users' reactions and decisions when making the meal plan.

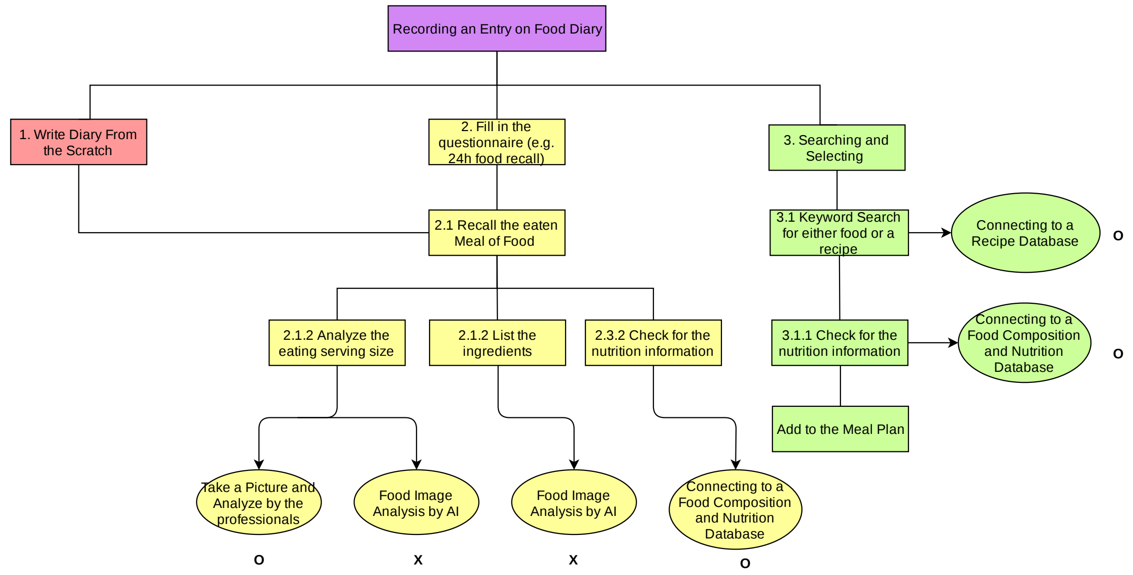

Recording an Entry on Food Diary
1. Write Diary From the Scratch
2. Fill in the questionnaire (e.g. 24h food recall)
3. Searching and Selecting
2.1 Recall the eaten Meal of Food
3.1 Keyword Search for either food or a recipe
Connecting to a Recipe Database
o
2.1.2 Analyze the eating serving size
2.1.2 List the ingredients
2.3.2 Check for the nutrition information
3.1.1 Check for the nutrition information
Connecting to a Food Composition and Nutrition Database
o
Add to the Meal Plan
Take a Picture and Analyze by the professionals
Food Image Analysis by AI
Food Image Analysis by AI
Connecting to a Food Composition and Nutrition Database
o
x
x
o

Figure 11. Hierarchical Task Analysis (HTA) for recording an entry on food diary.

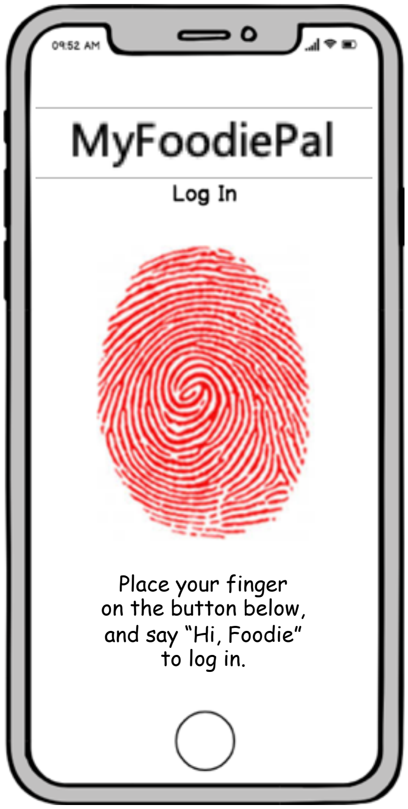
3.2.2 Prototype

In this section, the prototype walkthrough of the smart service system is presented by wireframes and descriptions. The real low-fidelity prototype with clickable buttons and links is attached as a pdf file in the Appendix, to better demonstrate the functionalities.

Log In

Health is a personal issue, and users are likely to have different health conditions. A majority of older adults suffer from multimorbidity (the coexistence of multiple chronic health conditions), which makes their needs of health-related service unique. This indicates the need to set up an account to access personalized services and retrieve personal health data with privacy. To reduce the log in effort, the system would use the biometric characteristics to retrieve the personal account.

When the user launches the application, he/she would see a log-in page with the icon of a fingerprint and the instructions for the login steps. This system allows users to log in via the individual's biometric characteristic of fingerprint and voice control. The 2-way authentication method could grant accessibility of the system to a wider audience while still enhancing the protection of personal data privacy.



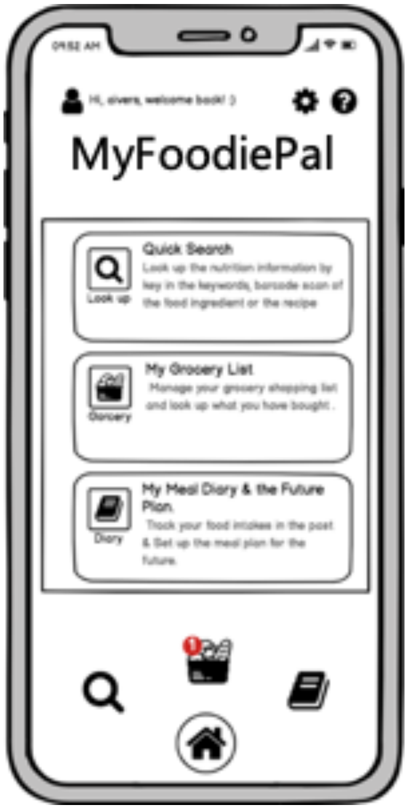
Place your finger
on the button below,
and say "Hi, Foodie"
to log in.

Figure 12. Log-in Page

Home

There are three core functions in this system, they are: "Quick Search", "My Grocery List", and "My Meal Diary & Future Plan". "Quick Search" allows the user to quickly look up the nutrition information for specific food ingredients by keyword search or barcode scanning; "My Grocery List" provides the grocery shopping list when the user finds a healthy food ingredient; "My Meal Diary & Future Plan" helps the user make a healthy meal plan and track food intakes.

On the homepage, there are 3 buttons with icons and the text description of these core functionalities. The homepage could also serve as the help document. The 3 icons are used on the menu bar for the user's navigation of the system. The menu bar is always at the same bottom location. Older adults may be new to the technology and may even suffer from degeneration of cognitive abilities, making it easier for them to get lost in such systems. Should they get lost in the system, they just need to click on the home button to return to the homepage to access the primary functions.



Quick Search
Look up the nutrition information by key in the keywords, barcode scan of the food ingredient or the recipe

My Grocery List
Manage your grocery shopping list and look up what you have bought .

My Meal Diary & the Future Plan.
Track your food intakes in the past & Set up the meal plan for the future.

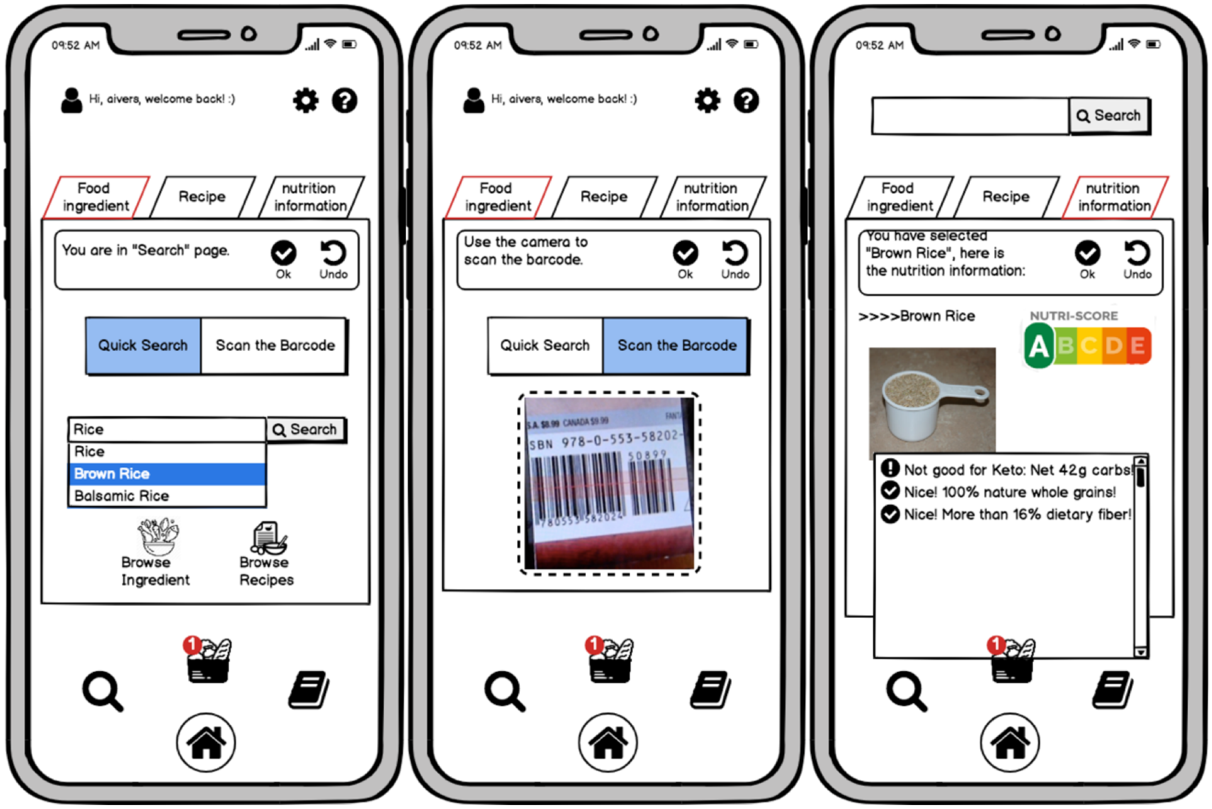
Figure 13. Homepage

Search

The walkthrough of searching for a food ingredient is shown in Figure 14. The user could either select to use the keyword search or the barcode scanning function to search for the nutrition information of the specific food ingredients.

The design of the search function is inclusive to those users with limited health literacy. There's an auto-complete function of the keyword search using the natural language processing algorithm. The nutrition information is displayed in an easy-comprehensive scoring and traffic light system.

There are also personalized health-related warning messages for each ingredient regarding their impact on health conditions.



You are in "Search" page.

Use the camera to scan the barcode.

You have selected "Brown Rice", here is the nutrition information:

Not good for Keto: Net 42g carbs!

Nice! 100% nature whole grains!

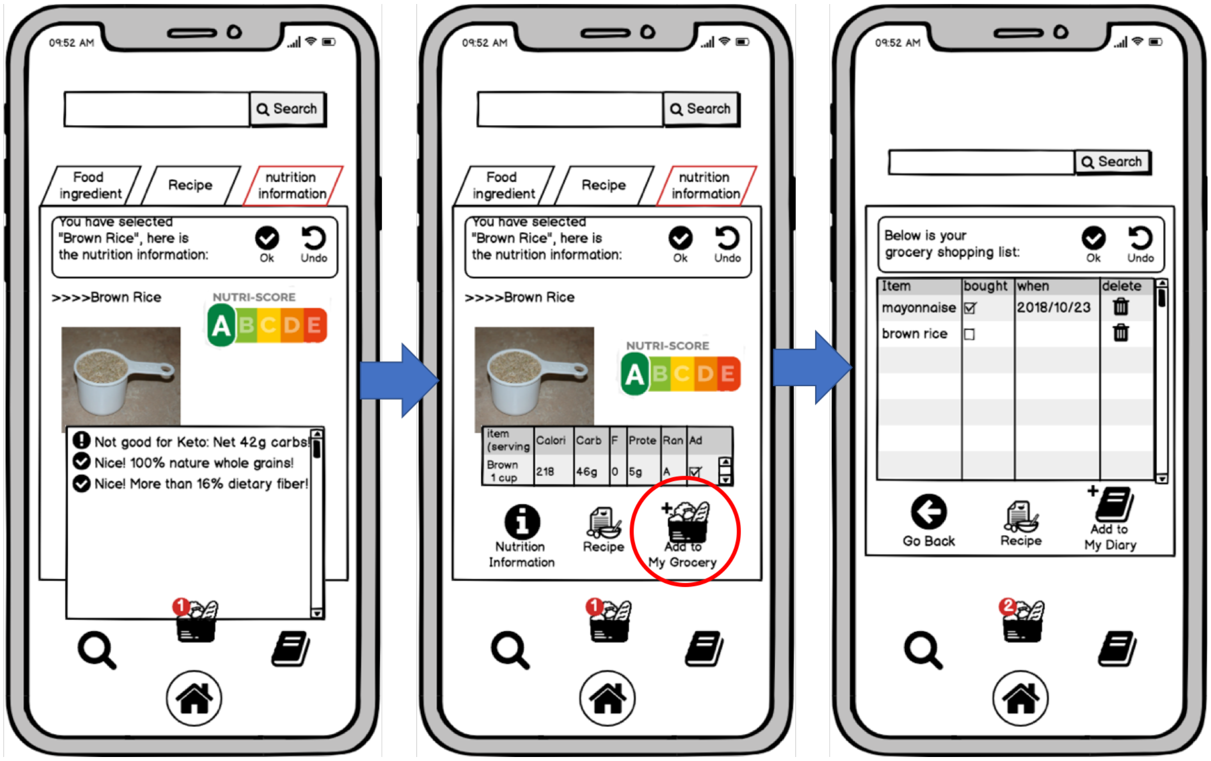
Nice! More than 16% dietary fiber!

Figure 14. The walkthrough of the searching function.

My Grocery List

After the user reads through the nutrition information and decide to buy the food ingredient, they could add the ingredient into their grocery list as shown in Figure 15.

The grocery list could be very useful for users who suffer from degeneration of cognitive functions. Integration of the grocery list with an online grocery shopping website delivery services offers the possibility of eliminating the need of the user of driving to the store. This can be critical in helping older adults maintain independence in the community setting.



You have selected "Brown Rice", here is the nutrition information:

>>>>Brown Rice

Not good for Keto: Net 42g carbs!
Nice! 100% nature whole grains!
Nice! More than 16% dietary fiber!

You have selected "Brown Rice", here is the nutrition information:

>>>>Brown Rice

item (serving	Calori	Carb	F	Prote	Ran	Ad
Brown 1 cup	218	46g	0	5g	A	☑

Below is your grocery shopping list:

Item	bought	when	delete
mayonnaise	☑	2018/10/23	🗑
brown rice	□		🗑

Figure 15. The walkthrough of adding an item to the grocery list.

My Diary

When the user clicks on the button of “Recipe” for the ingredient, they could first select the recipe by clicking on different cooking methods. The system would then recommend 2 easy recipes based on the user’s health condition. (The user should have set up their personal information when they sign up for an account.) The user could directly select from these two options, or he/she could search for alternatives. Once the recipe is selected and the button of “Added to the diary” is hit, the user would be navigated to the meal plan page as shown in Figure 17. On the meal plan page, the user could also select a past date to show the daily report of the summary of the food intakes and the nutrition consumption for a day.

The food diary function could provide a good health record for communication with the primary care provider. Meal plans could also be prescribed by the primary care provider or a nutritionist. And the user could then simply follow the plan by checking with the “My Food Diary” page.

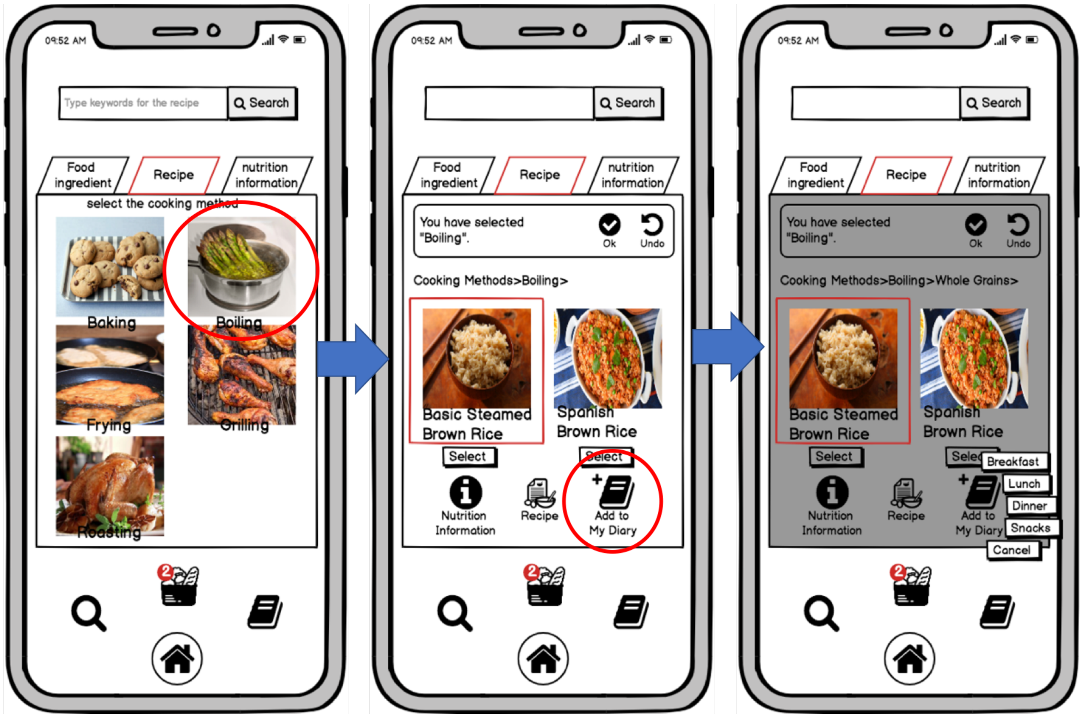


Figure 16. Browsing the recipe for the ingredient and adding the recipe to the grocery list.

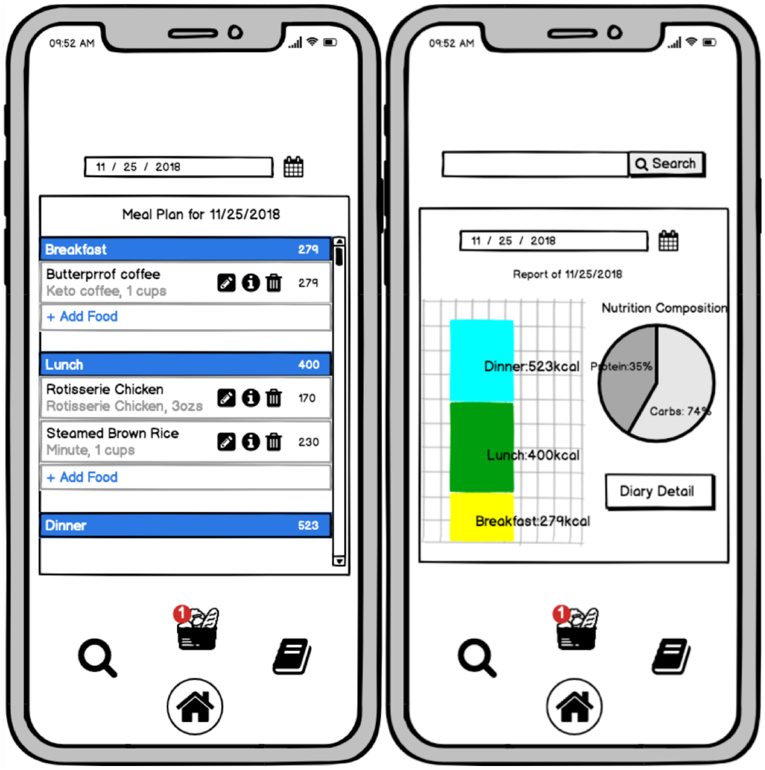


Figure 17. The meal plan and the daily report

3.2.3 Context Analysis

Follow the proposed extended usability engineering framework, the context of implementing persuasion were further analyzed based on task analysis. In order to improve the usability by reducing the data entry workload, I replaced text input task with the searching and selecting task for the use case of recording a food diary entry according to the cognitive walkthrough results. Hierarchical task analysis was performed again to further break down the searching and selecting task and create the user flow chart shown as Figure 18.

There are two design decisions to make, the first one is selecting between 2 proposed UI alternatives, and the second one is selecting between two formats to present nutrition information. Two UI alternatives are browsing-based UI, which is a vertical list of search result entries inspired by Google searching layout UI; and choice-based UI which is a side-by-side presentation of two alternatives recommended by the system based on user's preference and the healthier option of user's preferred alternative. Two nutrition information formats are FDA Nutrition Facts Panel and FSA Nutri-Scores.

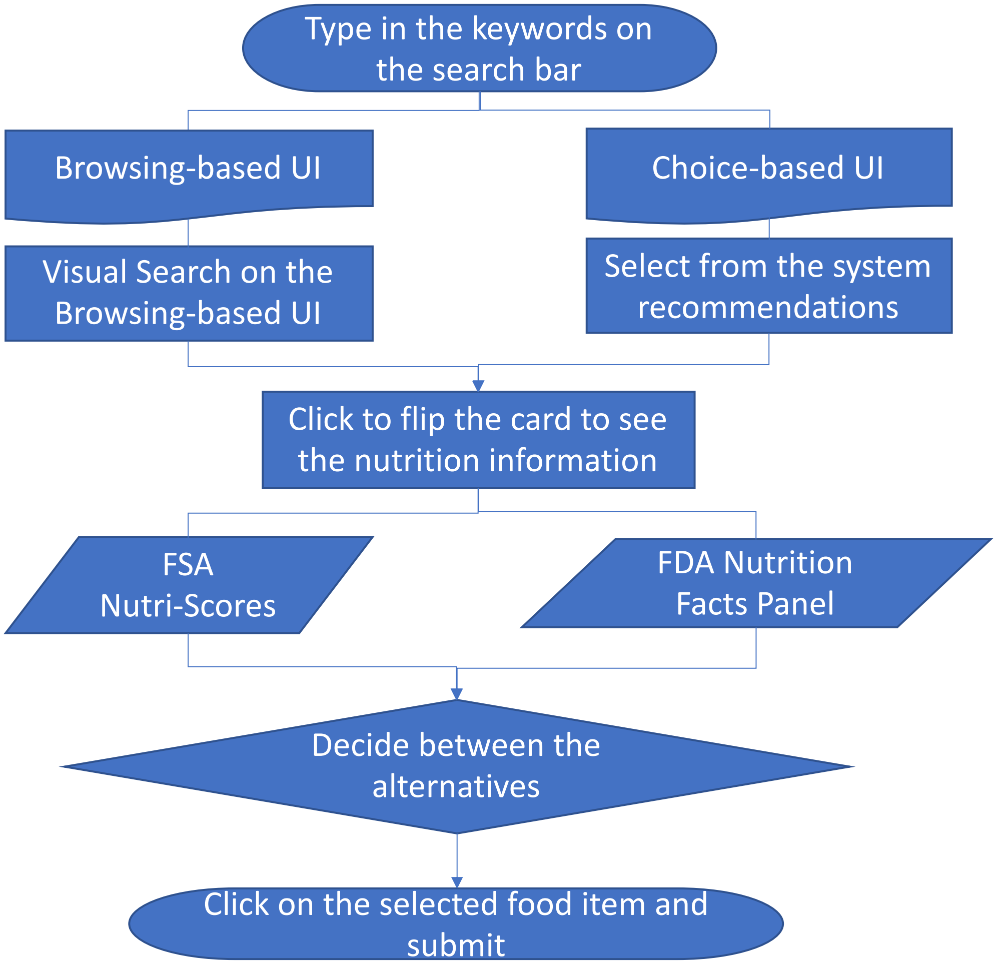

Type in the keywords on the search bar
Browsing-based UI
Choice-based UI
Visual Search on the Browsing-based UI
Select from the system recommendations
Click to flip the card to see the nutrition information
FSA Nutri-Scores
FDA Nutrition Facts Panel
Decide between the alternatives
Click on the selected food item and submit

Figure 18. User task flow chart for recording an entry in food diary by searching and selecting task.

Since the proposed mHealth app is designed to support users' meal planning decisions, it also provides chances to intervene users' thinking process and influence their decisions and behaviors as persuasive technology. According to Weinmann (2016), the selections of critical UI design elements may influence not only users' perceptions and subjective feeling about using the app but also their decisions and behavior. However, there's a research gap of empirical studies of digital nudge design and the best practice of what approaches to use, when to nudge, …etc. But with this task flow chart, the designer could gain a better idea of potential timings to implement nudge interventions and further decide between the design alternatives based on the effectiveness of nudge. As Figure 19 has shown, the designer has decided three timings of implementing the related nudge interventions for the proposed mHealth app:

1. Digital nudge intervention: when the user performs searching and selecting task on different UIs, different UI layout may also make different impacts on nudge, which could be considered as the implementation of digital nudge.
2. Information nudge intervention: providing information is a classic information nudge approach. However, different formats to present information may also induce different impacts of nudge effect.
3. Default nudge intervention: Inspired by the classic default nudge approach, it is also possible to set up the system default pre-selection when the user is making a decision between alternatives.

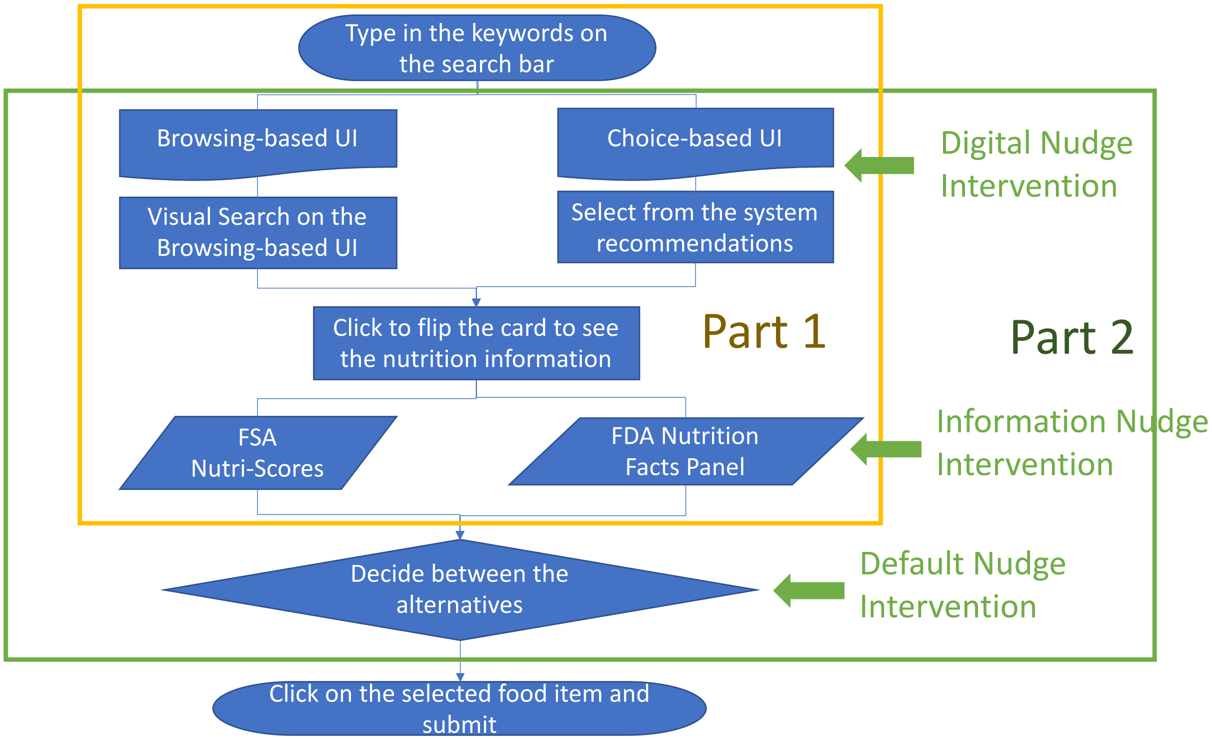

Type in the keywords on the search bar
Browsing-based UI
Choice-based UI
Digital Nudge Intervention
Visual Search on the Browsing-based UI
Select from the system recommendations
Click to flip the card to see the nutrition information
Part 1
Part 2
FSA Nutri-Scores
FDA Nutrition Facts Panel
Information Nudge Intervention
Decide between the alternatives
Default Nudge Intervention
Click on the selected food item and submit

Figure 19. Timing for nudge

4. RESEARCH FRAMEWORK

In previous chapters, the project scope has been narrowed down to persuasive mHealth apps. Chapter three focused the following discussion in this dissertation on the pilot design project of a dietary management app. This chapter proposes a research framework to evaluate the pilot design project and implicitly validate the proposed extended usability framework. The goal of this research project is to make the critical UI design decisions based on the evaluation of the product usability and persuasiveness for the proposed dietary management mHealth app.

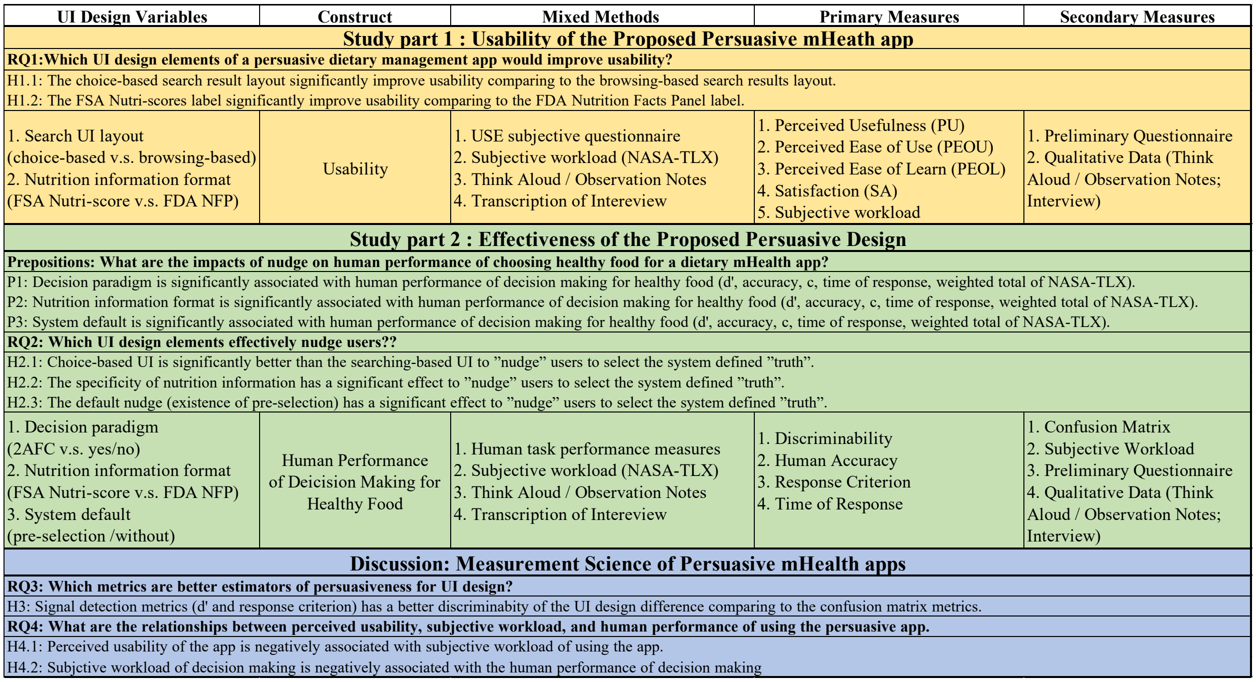
4.1 Research Framework

A two-stages user testing lab study was proposed to answer the research questions regarding to the usability of the proposed mHealth app for older adult users and the effectiveness of proposed nudge design interventions for health behavior change.

Study part one is focused on evaluating usability using mixed methods of subjective measures, including subjective questionnaire, think aloud, and interview. There are two reasons to adapt subjective measures rather than objective measures in this study. At first, it is the first time for participant to encounter with this innovative dietary management apps. They may need a period to learn and adapt themselves. In other words, study part one could be regarded as a training session, so the human performance data was not collected since it is not representative to the regular user behavior. Secondly, the subjective methods collect user's personal statements of preference, which is useful for design purpose. To better control the noise from individual differences in subjective methods, a preliminary questionnaire is used to investigate the participants' demographics and characteristics including age, computer proficiency, and health literacy. A secondary analysis is conducted to examine the mediating effect of individual differences.

Study part two is focused on evaluating persuasiveness, and in this dissertation, I proposed the human factors evaluation method to measure human performance of healthy food choice. The method is based on Signal Detection Theory (SDT), and thus the primary measures including discriminability, d'; response criterion, c; and accuracy. The regular human task performance measures, including the efficiency measure (time of response) and the workload measure (NASA

Task Load Index; NASA-TLX), are also collected. Confusion matrix measures are collected for the secondary analysis to justify the selection of SDT measures. And the secondary analysis is also conducted to evaluate the mediating effect of individual differences.



UI Design Variables	Construct	Mixed Methods	Primary Measures	Secondary Measures
Study part 1 : Usability of the Proposed Persuasive mHeath app				
RQ1:Which UI design elements of a persuasive dietary management app would improve usability?				
H1.1: The choice-based search result layout significantly improve usability comparing to the browsing-based search results layout. H1.2: The FSA Nutri-scores label significantly improve usability comparing to the FDA Nutrition Facts Panel label.				
1. Search UI layout (choice-based v.s. browsing-based) 2. Nutrition information format (FSA Nutri-score v.s. FDA NFP)	Usability	1. USE subjective questionnaire 2. Subjective workload (NASA-TLX) 3. Think Aloud / Observation Notes 4. Transcription of Intereview	1. Perceived Usefulness (PU) 2. Perceived Ease of Use (PEOU) 3. Perceived Ease of Learn (PEOL) 4. Satisfaction (SA) 5. Subjective workload	1. Preliminary Questionnaire 2. Qualitative Data (Think Aloud / Observation Notes; Interview)
Study part 2 : Effectiveness of the Proposed Persuasive Design				
Prepositions: What are the impacts of nudge on human performance of choosing healthy food for a dietary mHealth app?				
P1: Decision paradigm is significantly associated with human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX). P2: Nutrition information format is significantly associated with human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX). P3: System default is significantly associated with human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX).				
RQ2: Which UI design elements effectively nudge users??				
H2.1: Choice-based UI is significantly better than the searching-based UI to ”nudge” users to select the system defined ”truth”. H2.2: The specificity of nutrition information has a significant effect to ”nudge” users to select the system defined ”truth”. H2.3: The default nudge (existence of pre-selection) has a significant effect to ”nudge” users to select the system defined ”truth”.				
1. Decision paradigm (2AFC v.s. yes/no) 2. Nutrition information format (FSA Nutri-score v.s. FDA NFP) 3. System default (pre-selection /without)	Human Performance of Deicision Making for Healthy Food	1. Human task performance measures 2. Subjective workload (NASA-TLX) 3. Think Aloud / Observation Notes 4. Transcription of Intereview	1. Discriminability 2. Human Accuracy 3. Response Criterion 4. Time of Response	1. Confusion Matrix 2. Subjective Workload 3. Preliminary Questionnaire 4. Qualitative Data (Think Aloud / Observation Notes; Interview)
Discussion: Measurement Science of Persuasive mHealth apps				
RQ3: Which metrics are better estimators of persuasiveness for UI design?				
H3: Signal detection metrics (d' and response criterion) has a better discriminabity of the UI design difference comparing to the confusion matrix metrics.				
RQ4: What are the relationships between perceived usability, subjective workload, and human performance of using the persuasive app.				
H4.1: Perceived usability of the app is negatively associated with subjective workload of using the app. H4.2: Subjctive workload of decision making is negatively associated with the human performance of decision making				

Figure 20. Research Framework

After reviewing the literature, secondary hypothesis was further developed to validate the proposed signal detection method and metrics to evaluate the persuasiveness.

RQ3: Which metrics are better estimators of persuasiveness for UI design?
H3: Signal detection metrics (d' and response criterion) has a better discriminability of the UI design difference comparing to the confusion matrix metrics.

RQ4: What are the relationships between perceived usability, subjective workload, and human performance of using the persuasive app.
H4.1: Perceived usability of the app is negatively associated with subjective workload of using the app.

H4.2: Subjective workload of decision making is negatively associated with the human performance of decision making.

In this dissertation, I firstly compared the signal detection metrics with the confusion matrix metrics, which are the frequent-used decision making metrics to evaluate the system performance. And then, to better understand the characteristics of signal detection metrics and their relationships with usability, I conducted structural equation modeling analysis to verify the assumption.

4.2 Research Design

For the first part of the study, a full factorial experiment of 2^2 design alternatives would be used to evaluate the efficiency and the effectiveness of the proposed user interface design. The design variables are: 2 different display formats of food catalog; (image-based grids, text-based list) and 2 different display format of nutrition facts information (the FDA text-list fact or data visualization).

The responses is the perceived usability collected by Lund's USE questionnaire (Lund, 2001). The factors and levels are listed in Table 2 and the Figure 22.

Table 2. Factors and levels for the proposed 2^2 design alternatives.

Factors	Searching UI Layout	Nutrition Information Format
Level 1	Browsing-based UI	FDA Nutrition Facts Panel
Level 2	Choice-based UI	FSA Nutri-Scores

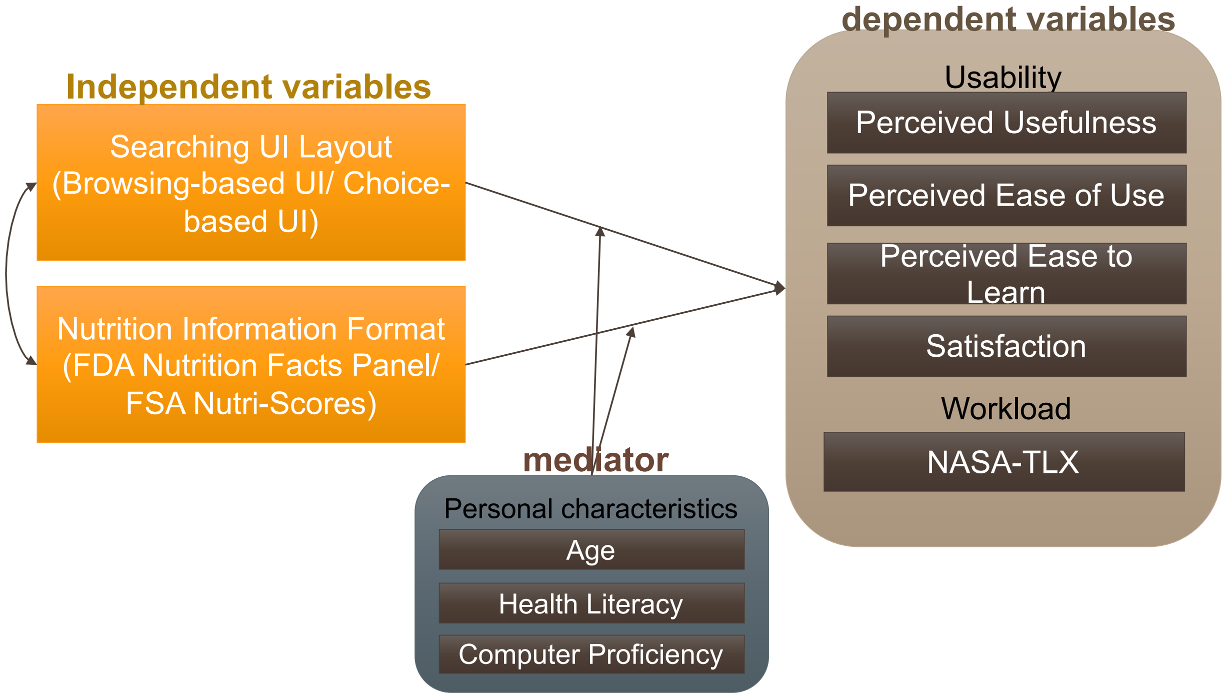


Figure 21. The path diagram of the study part 1.

For the second part of the experiment, the experimental scope is focusing on the decision-making scene to better control the noise to interfere with the major nudge intervention design variables. It's assumed the visual searching task has been done, and the participant is focusing on judging the healthiness of interested items and making a decision to put it on the meal plan. In this context, two simplified testing paradigms for decision-making were used. One is the yes-no test paradigm, the participant judges if the showing stimuli is healthy food or not and answers yes or no; another one is the two-alternatives forced choice (2AFC) test paradigm, two stimuli are shown at a time and the participant is asked to judge which food is healthier. The former testing paradigm could represent the typical use scenario when an individual examines the healthiness for each interested item derived from the traditional searching UI results one by one. And the latter testing paradigm could represent a new use scenario when there's always a recommendation on the side of the interested item.

For information nudge, I tested 2 different nutrition information format design, which were used as the nutrition labels on the package food: one is the back-of-package label in the U.S., Nutrition Facts Panel, derived from U.S. Food and Drug Association (FDA) standard; another one is the front-of-package label used in France and other European countries, Nutri-scores, derived from the United Kingdom Food Standards Agency (FSA) nutrient profiling system.

For default nudge, the preferred, healthy or healthier food option according to the FSA Nutri-scores and the American Diabetics Association eating guidelines will be pre-selected without notifying the participants. However, the participants could still make a change when they realized the existence of the pre-selection.

The above three design variables were deployed to a within-subjects 2^3 full factorial experiment. The three 2-level experimental factors for the design were: decision paradigm (yes-no paradigm vs. 2AFC paradigm); nutrition label format (FDA Nutrition Facts Panel v.s. FSA Nutri-score); default nudge (no pre-selection v.s. pre-selection).

Table 3. Factors and levels for the proposed 2^3design alternatives.

Factors	Decision Paradigm	Nutrition Information Format	Pre-Selection
Level 1	Yes / No Paradigm	FDA Nutrition Facts Panel	No pre-selection
Level 2	2AFC paradigm	FSA-Nutri Scores	With pre-selection

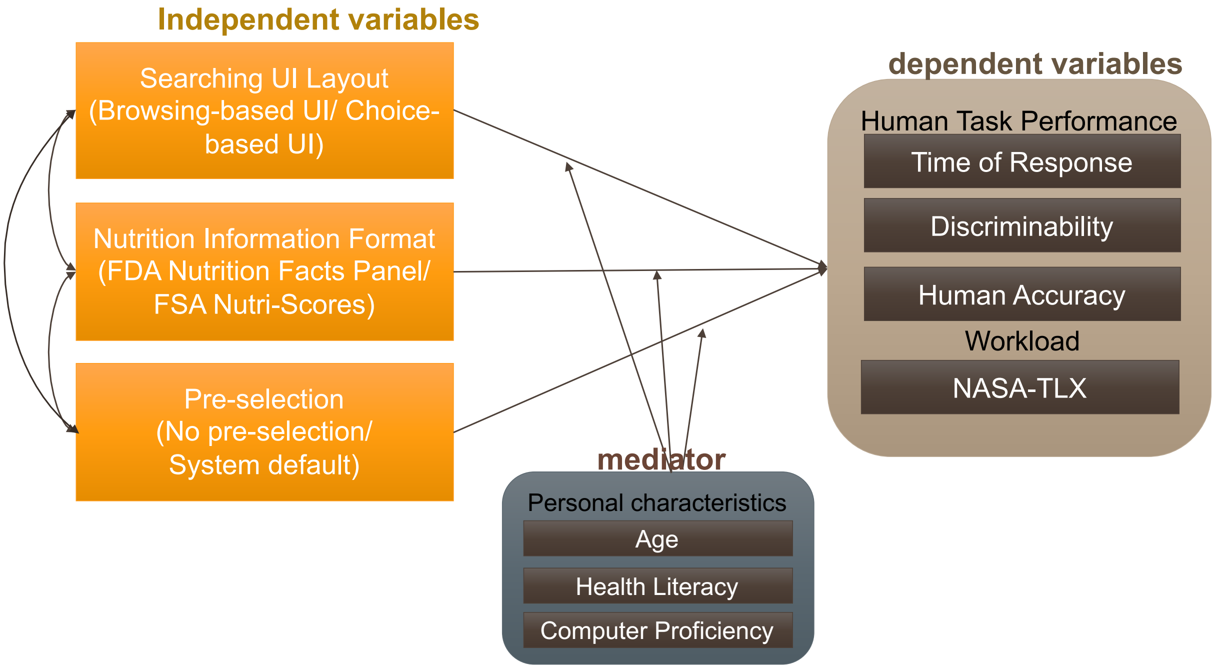


Figure 22. The path diagram of the study part 2.

5. METHODS

5.1 Participants

Participants came from two populations in the greater Lafayette area in Indiana. The experimental group consists of adults 60 years and older, who live independently or with families and consider themselves capable of meal planning. These individuals are likely to represent at least a subset of those who would most benefit from and use the diet change recommendation system under study. Individuals were screened by the preliminary questionnaire (see the appendix) and brief interview with the question "Are you comfortable with planning a meal by yourself?".

The control group was recruited from college and graduate students aged from 20-35 in Purdue University. These individuals are more likely to be early adopters of technology to enhance personal health management in their daily life. This comparison will allow us to draw inference about how ageing impacts the design of the application. Participants from this age group who don't have experience with smartphone apps were excluded from the experiment.

Sample Size Calculation

Based on a power analysis assuming a mean time difference of 100 seconds and a standard deviation of 25 for an F-test of the ANOVA with repeated measures and between factor effects, there's 80% chance of correctly rejecting the null hypothesis of no difference with a total sample size 18 (Cohen, 1977). Given the uncertainty around the power analysis assumptions we will seek to recruit 30 participants. To add in the allowance of data loss, the study will accept a maximum of 60 participants would be recruited from a university in the Midwest of the US.

In the related studies of applying eye-tracking for evaluating the user experience of smartphone apps, Qu et al. (2017) designed a within-subject experiment with 40 participants to quantify the user experience by both subjective and objective data (Qu, Zhang, Chao, & Duffy, 2017).

Recruitment Process

To ensure there is no unwanted pattern of the data due to the noise of uncontrolled demographic variables. Participants were recruited from multiple resources with different methods.

To recruit the older adults, a 3-min advertisement of this experiment were done in the weekly social event at Tippecanoe Senior Center. Tippecanoe Senior Center is a local non-profit organization provide social services such as the community-based senior nutrition program of providing free meal to the needed local seniors in great Lafayette area. Experimenters firstly introduced the purpose and the process of this experiment. And then, the older adult clients there were notified that this is a voluntary experiment. There won't be any penalty for those who don't want to participate in the experiment. Clients can ask any questions regarding the experiment. All of them have got experimenters' contact information, so that they can enroll in the experiment if desired.

In addition to the recruitment in Tippecanoe Centers, flyer advertising the study was also posted on the Purdue Today, at the Nursing Center for Family Health located in Lyles Porter Hall, and in several bulletin boards on the main campus of Purdue University. Additional subjects were recruited via word of mouth. The recruitment process was done under the regulation by Purdue's Institutional Review Board (IRB), with the approval case number IRB2019-214. And all the above-mentioned documents were also reviewed and approved by Purdue's IRB.

Data Collection Interruption

Data was collected during November, 2019 to March, 2020. Due to the outbreak of COVID-19, data collection was terminated earlier based on the IRB guidelines about prohibiting any forms of in-person study since March 24. Before the announcement. was posted, 42 participants had finished the whole study, one participant had only finished the study part one, and 4 participants had only finished study part two. However, there was another interruption in the early stage of data collection due to the programming bug in the formula of data collection for study part two. Ten data collected before November 26 has been discarded from data analysis of study part two. In the end, 43 data were included in study part one and 36 data was included in study part two.

Participant Characteristics

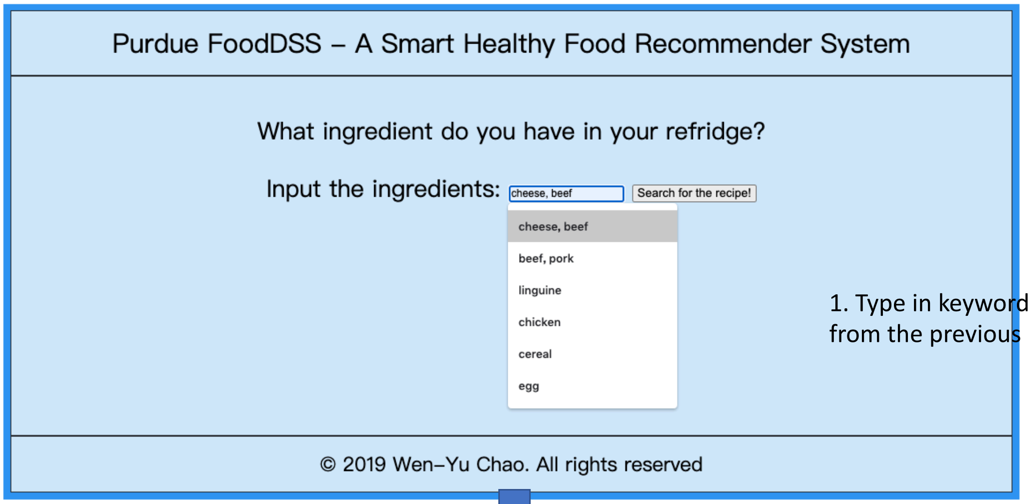
In the end, data of twenty older adults experiment subjects aged 60 years and older (mean=65.31) and the same amounts of students between the ages 18-35 (mean=28.88) were included. All the participants were from the great Lafayette area in Indiana, who consider themselves capable of meal planning, and had no history of diabetes.

Since ageing group is heterogeneous and chronological age is not the only explanation of their behavior (Boot et al., 2015b; Czaja, Lee, Branham, & Remis, 2012), computer proficiency level and health literacy level were also measured by the questionnaire adapted from Boot et al.'s computer proficiency questionnaire (2015) (Boot et al., 2015b) and Weiss' Newest Vital Sign questionnaire (2005) (Weiss et al., 2005) to check the mediation effect of individual difference.

5.2 Apparatus

For the purpose of conducting a user testing study in the computer lab environment setting. The prototype system was created on cloud as a portable web application using the free plan of goormIDE cloud service. The app was implemented by HTML/CSS/JavaScript source codes under the Node.JS environment using Express framework and MongoDB non-SQL database to support the back-end data flow. The backend databases are from two online resource: the recipe database is from a third-party company, EDAMAM, which provides a commercial food and recipe database and related APIs. Food database and the testing data of frequent-seen commercial cereal products in the U.S. and the related nutrition information including FDA Nutrition facts and FSA Nutri-scores were downloaded from an open source organization, Open Food Facts, which provides a free crowdsourcing food database.

High-fidelity prototype UIs were created for the experimental purpose, including the recipe search UI to train the participants to familiarize themselves with the searching and selecting task; the experimental UIs which would be used as the stimuli in each part of the experiment.



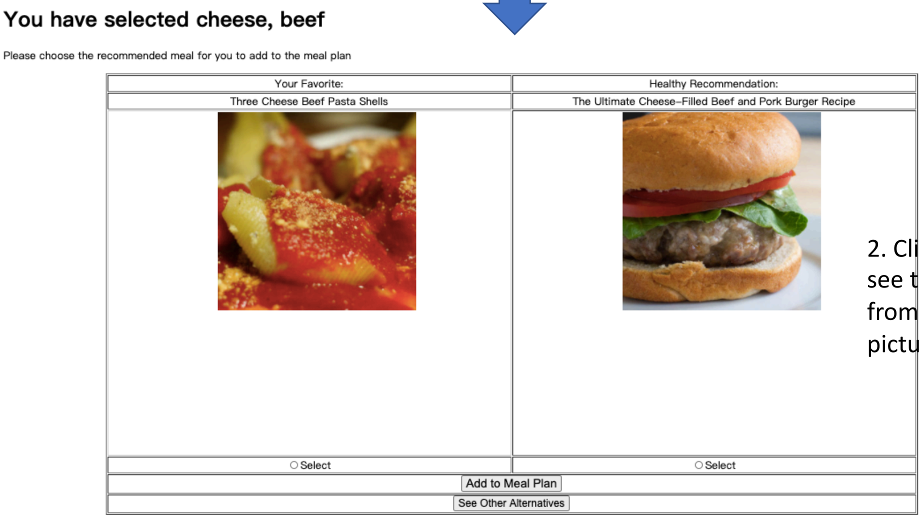
Purdue FoodDSS – A Smart Healthy Food Recommender System

What ingredient do you have in your refridge?

Input the ingredients: cheese, beef

© 2019 Wen–Yu Chao. All rights reserved

1. Type in keywords or selected from the previous record

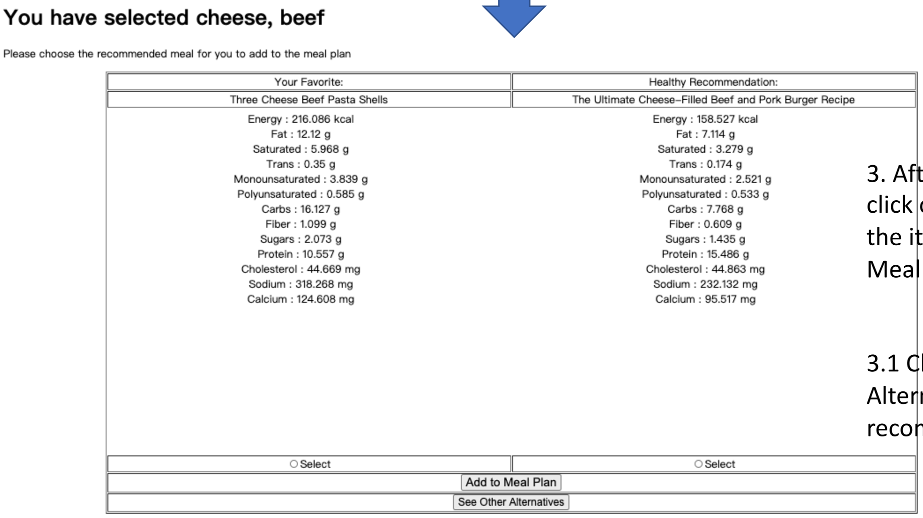


You have selected cheese, beef

Please choose the recommended meal for you to add to the meal plan

Your Favorite:	Healthy Recommendation:
Three Cheese Beef Pasta Shells	The Ultimate Cheese–Filled Beef and Pork Burger Recipe
○ Select	○ Select

2. Click on the picture to flip and see the nutrition information from the back side of the food picture.



You have selected cheese, beef

Please choose the recommended meal for you to add to the meal plan

Your Favorite:	Healthy Recommendation:
Three Cheese Beef Pasta Shells	The Ultimate Cheese–Filled Beef and Pork Burger Recipe
Energy : 216.086 kcal Fat : 12.12 g Saturated : 5.968 g Trans : 0.35 g Monounsaturated : 3.839 g Polyunsaturated : 0.585 g Carbs : 16.127 g Fiber : 1.099 g Sugars : 2.073 g Protein : 10.557 g Cholesterol : 44.669 mg Sodium : 318.268 mg Calcium : 124.608 mg	Energy : 158.527 kcal Fat : 7.114 g Saturated : 3.279 g Trans : 0.174 g Monounsaturated : 2.521 g Polyunsaturated : 0.533 g Carbs : 7.768 g Fiber : 0.609 g Sugars : 1.435 g Protein : 15.486 g Cholesterol : 44.863 mg Sodium : 232.132 mg Calcium : 95.517 mg
○ Select	○ Select

3. After reading the information, click on the radio button to select the item and click on the “Add to Meal Plan” button.

3.1 Click on the “See Other Alternatives” to check with other recommendations.

Figure 23. The Prototype UI and Standard Operation Procedures for the Training Purpose

To reduce the noise in the experiment, most experiments were done in the computer lab in Purdue Discover Park Learning and Resource Center on the Tobii Pro TX 300 Eye Tracker with of a wireless optical mouse controller. Tobii Pro TX 300 is a desktop version of a commercial eye-tracking device to collect the eye movements during the experiment with 300Hz sampling rate. This eye tracker is attached on a 23” personal computer monitor, which is large enough to show the stimuli in a reasonable size. Gazes and the scan path data could be captured and visualized as heat maps and gaze plots in Tobii Studio software. However, to better accommodate some community-dwelling older adult participants, there were some experiments done on a laptop (MacAir 2013 with a 13” monitor) without any attached eye tracker in the local community centers or other public spaces which is accessible for participants. Due to this reason, eye tracking data is not considered as the major response variable in our study. In addition, the same wireless optical mouse controller has been used to reduce the noise.

5.3 Measures

In this study, the mixed-methods measures were used to assess the human performance, perceived usability and subjective workload in both quantitative and qualitative aspects.

Most researchers have utilized mixed methods measures to evaluate human factors impacts on a system. A primary advantage of mixed methods measures is to ensure explainable and generalizable experimental results. Quantitative data could be used to verify the assumptions and generalized the results in a scientific manner while qualitative data be used to explain the findings and help designers better understand users’ thoughts and improvement directions in an iterative design process.

5.3.1 Quantitative Measures

Human Performance

Traditionally, human performance, in terms of speed and accuracy, has been assessed by measuring the task completion time and the error rate when an individual performs the experimental task under the simulated use scenario. In this study, given that the healthy food choice scenario was selected as the experimental task, I measured the time of response for completing the decision-making process on each UI.

Accuracy

In this article, the definition of accuracy was adopted from the confusion matrix, which is frequently used in the classification task of an information retrieval scenario. The definition aligns with the common sense of “proportion of correctness”, and the formula is as shown below:

Accuracy = (True Positive + True Negative) / (Total Opportunity)

The “true conditions” of healthy food is based on the FSA Nutri-Scores and suggested guidelines of American Diabetics Association. One may argue that would there be “truth” regarding to the healthiness. However, when selected the official guidelines as the “truth condition”, the calculation of “accuracy” in a free choice task according to the above formula, actually means “the proportion the user agrees with the truth conditions”, which would be a good metric to evaluate the successful nudge.

Discriminability

Since nutrition information format has been considered as an experiment factor, I proposed to adopt the discriminability from Signal Detection Theory (SDT) to measure information quality and implicitly evaluate persuasiveness according to literature review in section 2.3.5. The discriminability between signal and noise is defined as human sensitivity of a signal detection experiment, which is d’ = Z(Hit Rate) – Z(False Alarm Rate). The higher discriminability means the better signal could be perceived, which implies the information quality is better. Additionally, the formula based on the “proportion of correctness” is also proven as an unbiased estimator of “proportion of correctness” for imbalanced dataset. The detail is discussed in section 8.1.1

When the healthy food is regarded as the signal in a signal detection experiment, information quality could also be measured by discriminability, d’, which gives us an objective measure of UI design quality. The adjustment of McMillan and Creelman was employed for extreme values of 0 and 1 (Macmillan & Creelman, 2005).Traditionally, UI design quality is simply measured by subjective usability questionnaire, such as Lund’s USE questionnaire, with the latent constructs of perceived usefulness, perceived ease of use, perceived ease to learn and satisfaction (Lund, 2001).

Subjective Questionnaires

The self-rated metrics of subjective workload and perceived usability were the quantitative measures collected. The subjective workload was measured by NASA-TLX questionnaire with the weighted total scores as the primary workload metric. The perceived usability was measured by USE questionnaire, which is a survey instrument with 4 constructs on a 7-points Likert scale. The output metric from those four constructs were the total score of Perceived Usefulness (PU), the total score of Perceived Ease of Use (PEOU), the total score of Perceived Ease of Learning (PEOL), and the total score of Satisfaction (SA).

Perceived Usability

Perceived usability is measured by a subjective questionnaire after the experimental task was done on each UI. The main body of this post-experiment questionnaire mainly adopts Lund's USE questionnaire (Lund, 2001), with slight revision to fit the questions in the context of evaluating the proposed dietary self-management apps. The questionnaire includes the self-rating questions presents on a 7-points Likert scales to measure 4 constructs: Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Perceived Ease of Learning (PEOL), and Satisfaction. The USE questionnaire was used, rather than other frequent used questionnaire such as IBM CUSQ as a subjective measurement. Because the definition of the constructs for USE aligned with the definition of Technology Acceptance Model, and the Perceived Ease of Learning could be used to test the learnability dimension of usability.

Subjective workload

Workload is another frequent used human performance metrics. In this study, mental workload is measured by Hart and Staveland's NASA Task Load Index (NASA-TLX) subjective questionnaire (Hart & Staveland, 1988). This questionnaire gives a weighted total score of overall workloads on six dimensions including mental demand, physical demand, temporal demand, performance, effort, and frustration. The questionnaire firstly asks participants to rank the importance of each dimension by pair-wise comparisons, and these subjective ratings are used to calculate the weighted total workload of each task.

5.3.2 Qualitative Measures

Qualitative data were collected in the form of notes taken by the observer during the experiment when the participants practiced thinking aloud, and from the brief interview at the wrap-up session of the experiment. These data were firstly transcribed and then printed out on paper to mark the keywords. The experimenter read each transcript several times to mark keywords and form a concept. After marking on all the transcripts, the frequency for each concept were counted to analyze the tendency of critical themes.

Think aloud and observation notes

For study part 1, concurrent think aloud protocol was used. During the experiment part 1, the participants were encouraged to think aloud. However, for study 2, the retrospective think aloud protocol was used since the human task performance was the primary measure. Since not all the participants are used to thinking aloud concurrently, the experimenter also took notes by observation.

Short interviews

After the experimental task was finished, a semi-structured interview was conducted. the experimenter firstly asked questions about some points she has recorded on the observation notes, such as “I have noticed that you … when …, could you please explain more?”, to probe the participants think aloud retrospectively. And then, she conducted the interview using the following probes:

- For study part 1:

1. Which interface do you like best? Why?
2. To compare between A and B, which one do you like better?

- For study part 2:

1. Did you notice there’s a system guidance? What kind of guidance is it?
2. Do you like that guidance? Why?

Eye Gaze Plot

For those participants who has done the test on the desktop with an eye tracker, eye gaze plot is also used as the probe for retrospective thinking aloud and a supplemental measure to better analyze qualitative data.

5.4 Procedures

5.4.1 Preparation

As Figure 24. Flow Chart of Preparation has shown, the first step of the preparation is asking all the subjects to fill out a preliminary questionnaire. The questionnaire started with the screening questions on the front page, the experimenter asked the participants' age and the question "Are you comfortable with planning a meal by yourself?" (see the preliminary questionnaire in APPENDIX A.) to do the initial screen. The eligible participants then signed the consent form and continued to fill in the questionnaire, which is an instrument adapted from the Computer Proficiency Questionnaire (Boot et al., 2015a), and Newest Vital Sign Questionnaire (Weiss et al., 2005) to measure the subject's level of computer proficiency and health literacy. The Computer Proficiency Questionnaire contains self-rated questions in three sub-domain areas of using computer functions to assess an individual's computer proficiency. The Newest Vital Sign Questionnaire is a comprehension test of the NFP label for evaluating an individual's knowledge level in health and nutrition. The above assessments serve as a proxy to control for the participant's previous experience of using computers and making food choices based on nutrition information (Chao & Hass, 2020).

If the testing location was in the eye tracking lab of the Discovery Learning Resource Center (DLRC) at Purdue University, the experimenter briefly introduced the apparatus and asked if participant's were willing to use the eye tracker. Those participants who agreed to do so then moved in the lab to sit in front of the Tobii TX300 eye tracker, which is a desktop equipped with an eye tracking device, to complete the following experiment. Prior to the experiment, eye tracker calibration has been done for those participants using the eye tracker. The experimenter will ask participants to sit straight in a comfortable position, and then the experimenter would help to adjust the table height and the location of the monitor to make sure the eye-tracker is in an appropriate distance for the participant to read the information on the screen, and meanwhile the sensor can read the participants' eye movements. Participants will then be asked to complete a calibration task. This would include staring at nine points and visually tracking a red dot on the monitor. The experimenter may ask participants to redo the calibration task until the eye-tracker can read their eye movement. Remaining participants sit in front of a laptop without an eye tracker in a designated meeting room in DLRC or public area at the recruitment location. The experimenter sit close to the subject, at hand for assistance

The experimenter then introduced the experimental scenario and demonstrated the operation procedures of the proposed mHealth app. The participants were then allowed to practice the operation procedures and freely interacted with the prototype system until they have familiarized themselves with how to use the system.

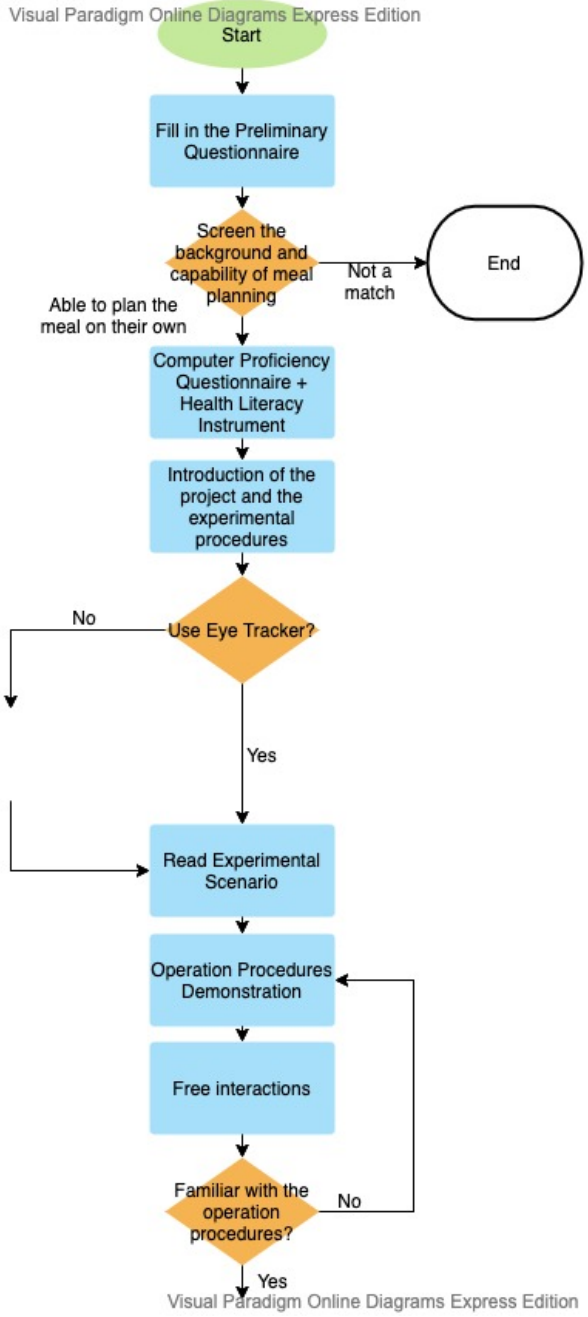

Visual Paradigm Online Diagrams Express Edition
Start
Fill in the Preliminary Questionnaire
Screen the background and capability of meal planning
Not a match
End
Able to plan the meal on their own
Computer Proficiency Questionnaire + Health Literacy Instrument
Introduction of the project and the experimental procedures
Use Eye Tracker?
No
Yes
Read Experimental Scenario
Operation Procedures Demonstration
Free interactions
Familiar with the operation procedures?
No
Yes
Visual Paradigm Online Diagrams Express Edition

Figure 24. Flow Chart of Preparation

Experimental Scenario

Since the pilot project is a scenario-based design, the proposed usability engineering project is also scenario-based to better collect human task performance (Rosson & Carroll, 2002). The experimental scenario is written based on the summary of a variety of scientifically based guidelines and nutritional recommendations for adult people with diabetes from FDA, USDA, American Diabetes Association (ADA), etc. The written script of the experimental scenario is as shown below:

- *John Doe has recently been diagnosed with the pre-diabetes, the doctor set the dietary goals for him to control the weight and the daily food intakes as shown below.*
    - *Control the weight by controlling the daily total calories intake to 1400 kcal and take the light exercise (walking or jogging) 30 minutes per day.*
    - *Control the daily intake of the saturated fat 20g below.*
    - *Control the daily intake of sugars below 36g, daily monitor the blood sugar level.*
    - *Control the daily intake of sodium below 2300 mg. (1 teaspoon of salt)*
    - *Eating 20 grams of dietary fiber per day."*
- *John Doe has consulted with the dietitian, the dietitian has set up an initial meal plan for him, and John Doe was suggested to use the healthful food recommender system to draft a detailed meal plan and send it back to dietitian to check.*
- *John Doe's favorite food is cereal. And he used to eat cereal for breakfast in the past. However, the dietitian has set up the limitation for this right now. According to the meal plan, he is allowed to have 1/2 cup of cereal, which is about 25g, for breakfast. Fortunately, John Doe is allowed to make a choice of the cereal as long as the food restrictions could be followed. Could you make a decision whether the recommended cereal healthful or not based on John Doe's case?*

5.4.2 Experiment part 1: Usability Evaluation for the Proposed mHealth app.

Study part 1 focused on evaluating the impacts of two design variables, search design layout and nutrition information formats, on the user's perceived usability and subjective workload of using the system. As Figure 25. Flow Chart of Study Part 1 has shown, the study part 1 starts with a stimulus, which is one of the four UI treatments including the combinations of browsing-based user interface or choice-based user interface and text-list nutrition information or symbolized nutrition information. The order of the four versions were randomized to avoid confounding results with a learning effect (Chao & Hass, 2020).

The experimental task was to choose a breakfast cereal out of six (or six pairs of) alternatives within the context of the experimental scenario. The participants were encouraged to think aloud when they performed the task following Ericsson and Simon's concurrent think aloud protocol, this procedure helped the experimenter gain further insights of user behavior.

After finishing each UI treatment, participants completed an after scenario questionnaire including the self-rated questions adapted from USE questionnaire on 5-point Likert scales for measuring Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Perceived Ease of Learning (PEOL), and Satisfaction; and a NASA-TLX subjective workload instrument for measuring the mental workload (Hart & Staveland, 1988; Lund, 2001). The experiment concluded with a short interview where participants were asked for their comments and suggestions to improve the system design (Chao & Hass, 2020).

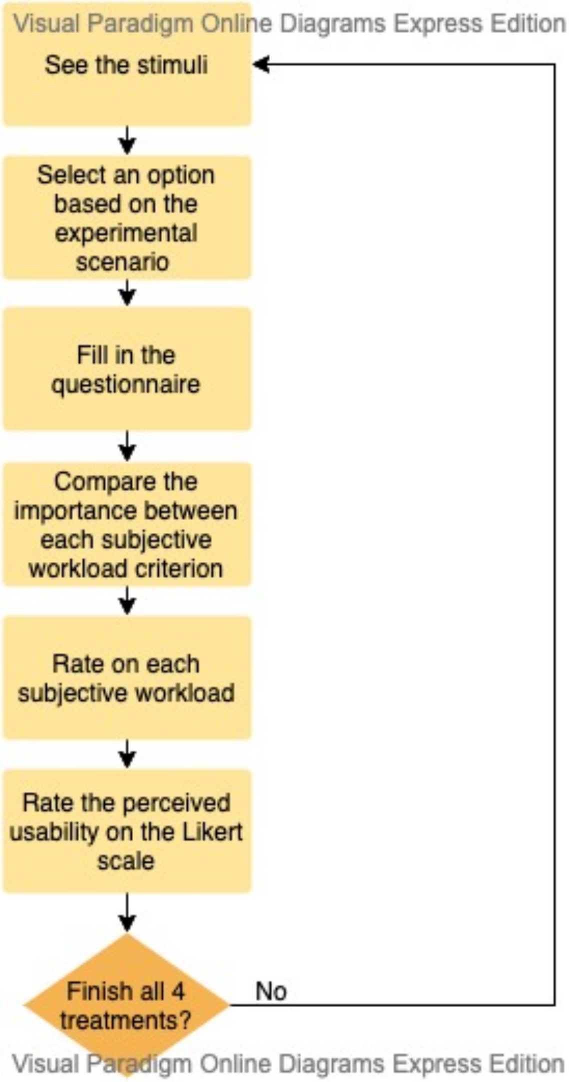
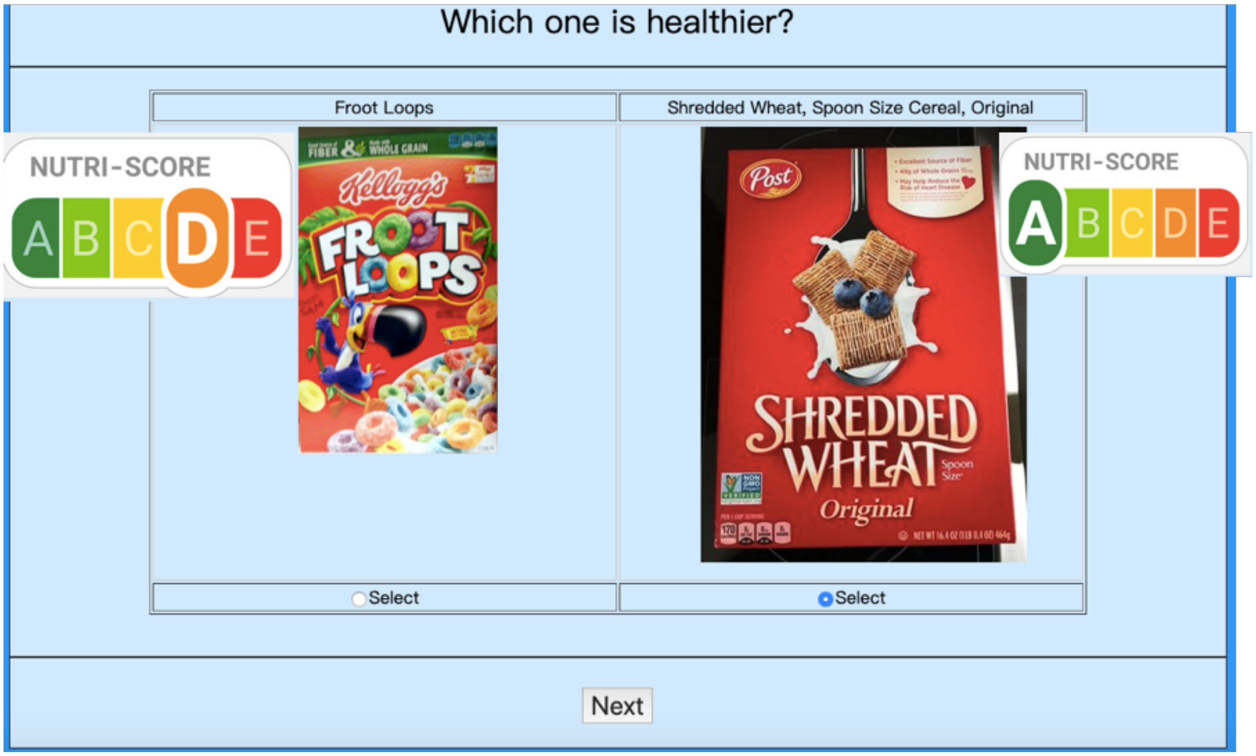

Visual Paradigm Online Diagrams Express Edition
See the stimuli
Select an option based on the experimental scenario
Fill in the questionnaire
Compare the importance between each subjective workload criterion
Rate on each subjective workload
Rate the perceived usability on the Likert scale
Finish all 4 treatments?
No
Visual Paradigm Online Diagrams Express Edition

Figure 25. Flow Chart of Study Part 1

5.4.3 Experiment Part 2: Healthy Food Choice Experiment

Stimuli

In order to better reduce the potential noise, the complexity of the UI and task operations were simplified. Eight user interfaces were created to accommodate 2^3 treatments of the combinations of three 2-level UI design variables, which are decision paradigm (yes/no task vs. 2AFC task) x nutrition information format (FDA NFP vs. FSA Nutri-scores) x system default pre-selection (no system default value vs. with default pre-selection). Figure 26 shows the experimental UI of (2AFC task) x (FSA Nutri-scores) x (with default pre-selection) as an example. To better present the UI characteristics on this image below, the nutrition information was shown next to the food image. But in the real task operation of the experiment procedures, the nutrition information shows only after the participant click to flip the food image to ensure the experimental steps are identical the real app operations.



Which one is healthier?

Froot Loops	Shredded Wheat, Spoon Size Cereal, Original
○Select	●Select

Figure 26. The experimental UI for healthy food choice experiment.

Experimental Procedures

The experiment then started with the experimental scenario of selecting a healthy cereal option for a care recipient who was newly diagnosed with Type II Diabetes. The subject then saw the experimental UI presenting randomly selected stimuli of the pictures and nutrition information of healthy/unhealthy breakfast cereal. Each of 8 UI designs (2^3) were presented in a random order and replicated 6 times to reduce the test difficulty level and learning effect. For each replicate, the subject was asked to respond yes/no or identify the healthy option following different test paradigms of the signal detection experimental procedures. The subjects were encouraged to think aloud while completing the experimental tasks. After each of the 8 designs, subjects completed a NASA-TLX questionnaire (Hart & Staveland, 1988) and commented on their performance and shared their thoughts of the UI. The detail experimental task flow chart is shown as Figure 27. Flow Chart of Study Part 2 has shown.

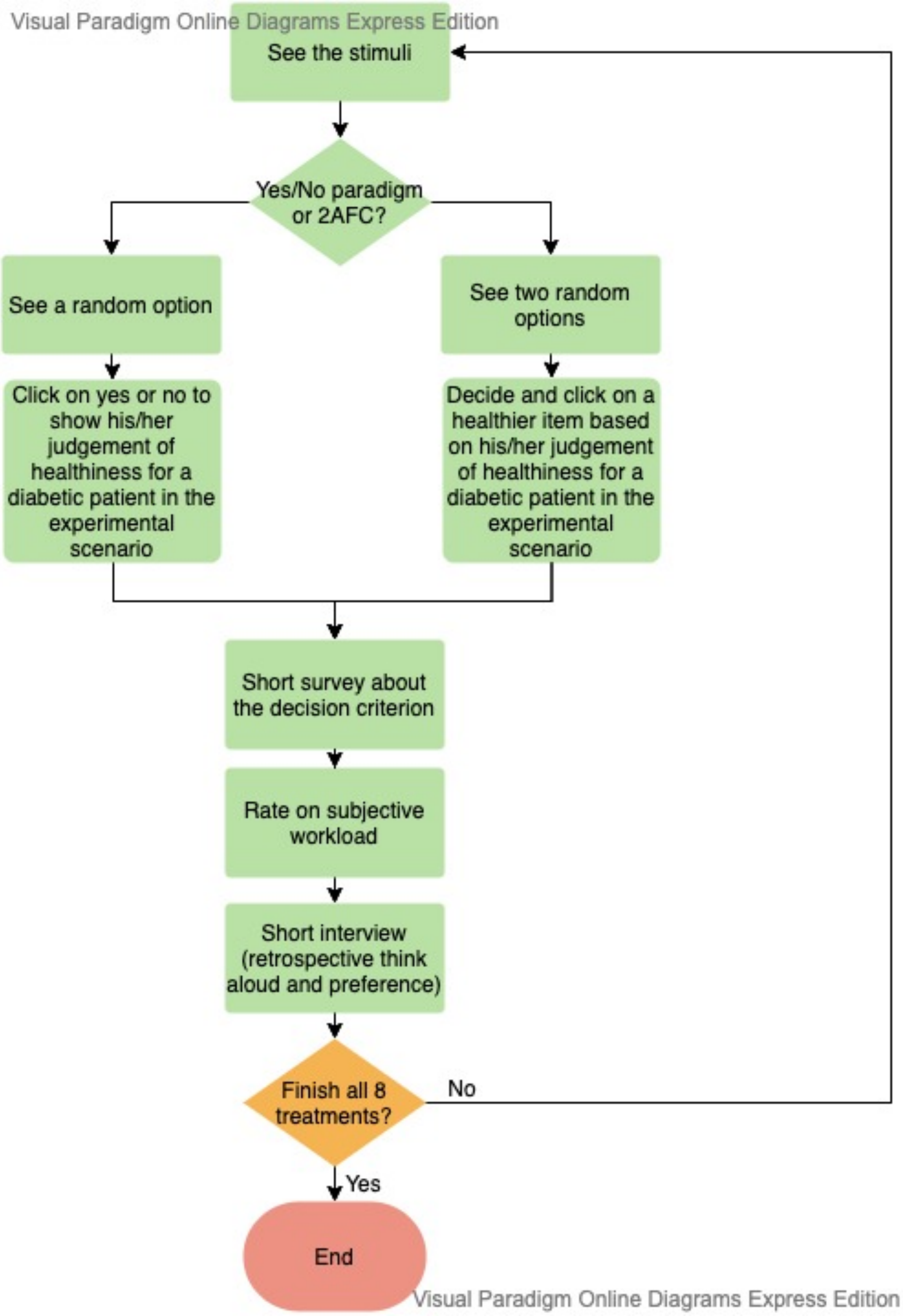

Visual Paradigm Online Diagrams Express Edition
See the stimuli
Yes/No paradigm or 2AFC?
See a random option
See two random options
Click on yes or no to show his/her judgement of healthiness for a diabetic patient in the experimental scenario
Decide and click on a healthier item based on his/her judgement of healthiness for a diabetic patient in the experimental scenario
Short survey about the decision criterion
Rate on subjective workload
Short interview (retrospective think aloud and preference)
Finish all 8 treatments?
No
Yes
End
Visual Paradigm Online Diagrams Express Edition

Figure 27. Flow Chart of Study Part 2

5.4.4 Human Subject Payment and Rewards

All the participants who has completed the entire study (preliminary questionnaire and the experiments) have received $10 cash paid at the time of completion. Participants were required to sign the human subjects log to receive payment. According to the rules of the Internal Revenue

Service (IRS), payments that are made to the participants as a result their participation in a study may be considered taxable income. Healthy refreshment (eg. chips, cookies, water) was provided for the subjects in a needed base during the study.

5.5 Data Analysis

5.5.1 Individual Difference

Since it would be difficult to find the participants if we set up too many conditions, the individual characteristics (computer proficiency and health literacy) are not considered as control factors in our experimental design. Instead, we recruited on the rolling basis, and did the k-means cluster analysis to determine the distinct clusters based on age, computer proficiency, and health literacy to roughly check the individual difference of the response between groups. To determine k, we performed hierarchical cluster analysis using ward's method to draw the dendrogram and the scree plot based on the height.

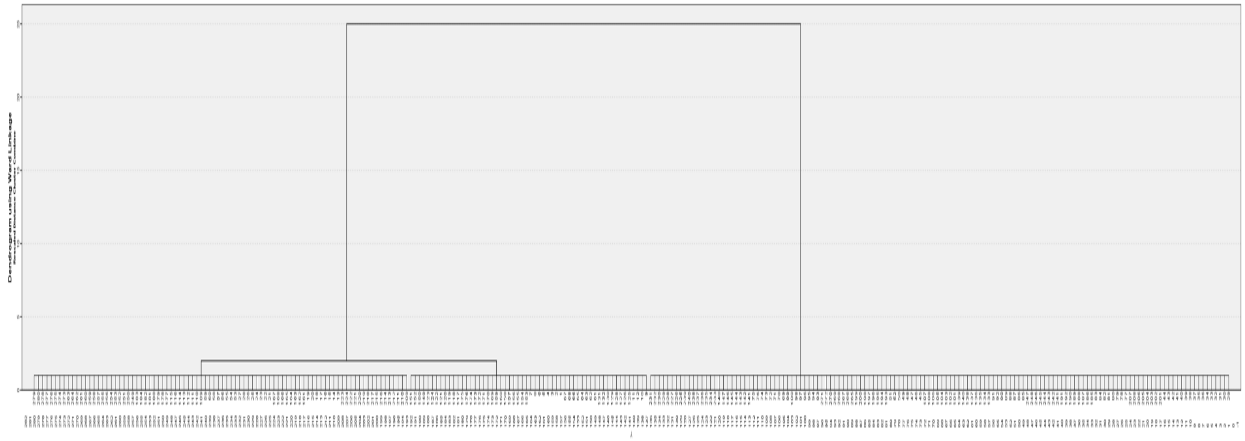
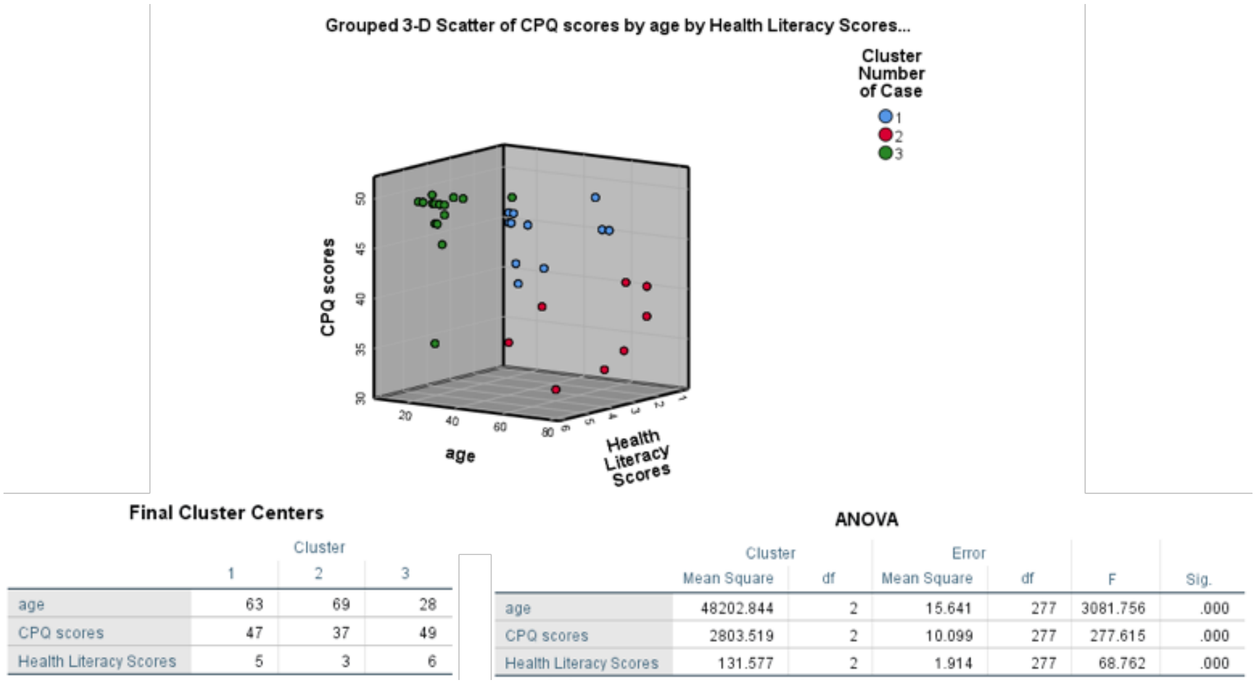


Figure 28. Dendrogram of Hierarchical Cluster Analysis for Determining k.

Based on the computer proficiency scores and the health literacy scores, the data could be further grouped by younger Older Adults(OA) (older adult participants in average aged 63), older OAs (older adult participants in average aged 69), and students. The terms of older OAs and younger OAs used in this dissertation are simply the naming of the participant groups based on cluster analysis results. In previous literature, older OAs may indicate the 85 or older population (Binstock, 1985). Students have the highest computer proficiency and health literacy scores. They perform slightly better than younger OAs in computer proficiency and health literacy; the older OAs have significantly lower scores than the other two groups.



Final Cluster Centers

	Cluster		
	1	2	3
age	63	69	28
CPQ scores	47	37	49
Health Literacy Scores	5	3	6

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
age	48202.844	2	15.641	277	3081.756	.000
CPQ scores	2803.519	2	10.099	277	277.615	.000
Health Literacy Scores	131.577	2	1.914	277	68.762	.000

Figure 29. K-means Cluster Analysis Results (k=3)

One-way analysis of variance (ANOVA) analysis was employed to analyze the differences of the means between groups. Following ANOVA, Tukey's Honest Significant Difference (HSD) test was performed for all pair comparisons to find out specific groups which are significantly different from others. Dunnet's test was also performed using the student group as the control group. Significance levels were set at the 5% level and all the statistical analysis were performed using Minitab® software (version 19).

5.5.2 Primary Analysis

General Linear Model analysis was employed for analyzing quantitative data. The models' main effects were the design variables estimated as fixed effects. A random effect for subject was used to capture the within subject correlation. All possible interactions between fixed effects were estimated. Significance was set at 5% and all data analyses were performed in Minitab® statistical software.

Grounded theory was applied for analyzing qualitative data. The transcription of the qualitative data was read several times to tag the keywords of the concepts for each paragraph. The related tags were collected as a theme and classify the related concepts by themes. And the numbers of concepts for each theme were counted to sort the frequently mentioned themes.

5.5.3 Secondary Analysis

For the purpose of justifying the proposed measures, secondary analysis was conducted. The results and the discussion were described in chapter 8.

To verify the H3 that the signal detection metrics are more sensitive to detect the changes of UI design variables comparing to the confusion matrix metrics, the general linear model analysis was performed with the same dataset again using the confusion matrix metrics as the responses.

Structure Equation Modeling (SEM) analysis was conducted to explore and confirm the relationships between the mixed-methods measures and the relationships between the constructs of usability, workload, and human performance of food choices.

6. USABILITY OF THE PROPOSED APP

Partial results of the presented work in Chapter 6 have been published in page 221-234 of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) by Springer with the title "Choice-Based User Interface Design of a Smart Healthy Food Recommender System for Nudging Eating Behavior of Older Adult Patients with Newly Diagnosed Type II Diabetes" on July 19th, 2020. The article was co-authored with Dr. Zachary Hass.

6.1 Hypotheses

The goal of study part one was to examine two UI design variables for a relationship with participants perceived usability and subjective workload rating. Two key hypotheses were tested:

H1: The search result layout is significantly associated with the self-rated metrics of workload and usability (weighted total of NASA-TLX, PU, PEOU, PEOL, Satisfaction).

H2: The nutrition information format is significantly associated with the self-rated metrics of workload and usability (weighted total of NASA-TLX, PU, PEOU, PEOL, Satisfaction).

6.2 Results

6.2.1 Quantitative Results

The relationship between the outcomes variables of workload and perceived usability (PU, PEOU, PEOL, Satisfaction) and the design variables of search result layout and nutrition information format were assessed by fitting general linear models with a random effect for subjects. Table 4 shows the summary of the model coefficients and the p-values with the significance level of 0.05 being highlighted. For the outcome of subjective workload, the nutrition information format has a significantly negative association with the total score from NASA-TLX (P=0.000). The main effect of search result layout was not found to be statistically significant (P=0.615). Nutrition information format was positively associated with PEOU scores (P=0.000), of PEOL scores

(P=0.007), and also with Satisfaction scores (P=0.049). Neither the search result layout main effect (P=0.820 for PEOU, P=0.615 for PEOL, and P=0.892 for Satisfaction) nor the two-way interaction (P=0.558 for PEOU , P=0.140 for PEOL, and P=0.496 for Satisfaction) were statistically significant for both PEOU score, PEOL score and Satisfaction score. The FSA Nutri-scores nutrition information format has lower mean workload score and higher mean PEOU, mean PEOL, and mean Satisfaction in comparison to FDA NFP label. However, the lack-of-fit test result is significant for the model with PEOL, which suggested the inadequacy of the fitted model. For the model with PU as the outcome variable neither main effects nor the interactions were significant. The residual plots of PU (Figure B.1), PEOU (Figure B.2), PEOL (Figure B.3), Satisfaction (Figure B.4), and NASA-TLX (Figure B.5) show the proposed models follow the normality assumption. (See the APPENDIX B.)

Table 4. Summary of GLMs for subjective workload and perceived usability

Coefficients of GLMs for		**NASA-TLX**		**PU**		**PEOU**		**PEOL**		**Satisfaction**	
Source	**DF**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**
Searching Results Layout (Choice-based=1)	1	-1.240	0.615	-0.730	0.607	0.390	0.820	-0.224	0.615	-0.160	0.892
Nutrition Information Format (FSA-Nutri-scores=1)	1	-12.750	0.000	1.330	0.351	6.210	0.000	1.231	0.007	2.370	0.049
Subjects	42		0.000		0.000		0.000		0.000		0.000
Choice-based*FSA-Nutri-scores	1	6.020	0.086	-1.170	0.558	-3.550	0.140	-0.299	0.633	-1.140	0.496
Error	128										
Lack-of-Fit	124		0.074		0.852		0.066		0.023		0.367
Pure Error	4										
Total	173										

NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study); PU is Perceived Usefulness; PEOU is Perceived Ease of Use; and PEOL is Perceived Ease of Learning.
DF stands for Degree of Freedom; Coef stands for the coefficients of general linear models; p is p-value, the significance level is 0.05. (See the ANOVA tables in APPENDIX B.)

Notably, for each of the general linear models, variability in the outcome was dominated by the between subject variability. The result implies that the significance of experimental variables may be underestimated because individual differences mediate the effects, findings from section 6.3. also support this explanation. Although between subject variability is also a possible warning of lower test reliability, however, since many preventive procedures have been taken during the study (such as within-subjects design, comparison with the baseline, etc.), it is confident for me to say that the results are protected. The detail is discussed in section 6.4

6.2.2 Qualitative Results

Additional insights were gleaned from the qualitative data to supplement the quantitative analysis and better explain the meaning of the results. As chapter 5 has mentioned, grounded theory was applied to develop themes from the transcriptions of the think aloud exercise and interviews. Since the interview questions were semi-structured, major themes were developed around the impacts of design variables. The findings aligned with the quantitative results and provided more insights as expected.

The first theme for discussion helps to explain the quantitative result that there were no significant impacts of the searching layout format on perceived usability. Interviewees did not come to a common agreement of the preferred format (choice-based versus browsing-based). The contradicting preferences appeared to have canceled each other out in the effect estimate and may be explained by the presence of difference search strategies used by different groups. For example, the participants who are proficient at using computers, included three students and two younger OAs (see the classification of participants in this study and the definition of younger OAs in 5.5.1), preferred the vertical list view over the horizontal side-by-side view. They specifically mentioned that they preferred performing an exhaustive search of checking all alternatives in the list view. They are familiar with the browsing-based UI and it requires less steps to achieve their searching goals. Contrastively, one student and three older adults advocated for the side-by-side view layout. They enjoyed performing the pairwise comparisons task and thought it was more intuitive and straightforward to perform the task on the choice-based UI. One of the older adults suggested a hybrid, so that the exhaustive search could be performed on the list-view layout but in the horizontal presentation of the alternatives, for the ease of comparison. This horizontal list-view layout was also mentioned by the other two older adults since the horizontal scanning tallies with their natural reading behavior.

The study has provoked the participant's interest in co-designing the search results layout. Given different searching strategies were expected to be adopted, several participants shared ideas to improve the design by using dynamic and adaptive UIs. For example, list-view layout could be used as the default to present the searching results. When a user shows their interest in an alternative, there would be a buffer for them to temporary store the items of interest, and the side-by-side view layout would be used to present those items for a final decision. Another compelling suggestion was contributed by a student who recommended enabling the faceted search function, which means users could search an item by applying the system pre-defined filters. A novel design idea is the personalized facet search, the system learned from the user inputs and memorized the user's preferred setting for the next time use.

Another theme arising from the qualitative data is that most participants preferred the Nutri-scores labels, which also lines up with the quantitative results, too. However, important concerns about the FSA Nutri-scores format were expressed by some interviewees (four students and two older adults). First of all, the general version of the nutrition labels was used in this study, which may confuse the user in the use scenario for the patient with type II diabetes and the caregivers. Secondly, the absence of critical information to control daily intake values such as serving size and sugar content on the Nutri-score label was pointed out. In affected the perceptions of two older adult participants in the same way, they mentioned their hesitation in deciding between NFP and Nutri-scores because the detailed information of NFP seems to promise the utility in the health context, but they actually desired a simple and easy-to-use label for decision making.

Another surprising theme was about learning the Nutri-scores label. The comprehension issue was reported and guidance about how to interpret the Nutri-score label is needed. One student and one older adult reported their confusion since the FSA Nutri-score label was in the reverse order of numerical scale to show the rating. The label employs letter grades on an A to E scale with the traffic light color coding, "A" represents the best and "E" the worst. However, the scale starts with

"A" on the left-hand side, which contradicts with the customary order of usual mathematical scales (increasing numerical order).

The last prominent theme observed was about the credibility of the new nutrition label. Supplemental information of how to use the label and who created the label seem to be two determinants of trust. Most of the participants encountered the Nutri-scores label for the first time during the experiment and they adopted a tentative strategy to trust the label. They carefully requested more explanations about the context of the label and how to interpret it, and repeatedly confirmed that the label aligned with their own knowledge. Some participants raised their level of trust when they learned that the creators of the label are government authorized experts. Some participants trust the label when they get back the sense of control by become familiar with its interpretation. Only a few participants devotedly trusted the label without going through this process.

6.3 Individual Difference

Table 5 presents the summary of the results obtained from the one-way ANOVA of each of the perceived usability measures and subjective workload by age group (See the ANOVA tables and the Tukey Post-Hoc comparison in APPENDIX C.). The results show the significant differences between group means of PU, PEOU, and Satisfaction. The post-hoc Tukey's HSD test suggests the older OA group is significantly different from other two groups as shown by the superscripts (common letters indicating no difference), with higher subjective ratings. A possible explanation is the age-related social desirability since the questionnaire was done with the experimenter on spot. Social desirability is a tendency that individuals give socially desirable answers for social approval or avoid critics and arguments. Several studies have also confirmed that social desirability is positively correlated with age (Dijkstra et al., 2001; Fastame & Penna, 2012; Ray, 1988).

The residual plots of PU (Figure C.1), PEOU (Figure C.2), PEOL (Figure C.3), Satisfaction (Figure C.4), and NASA-TLX (Figure C.5) show the proposed measures follow the normality assumption. (See the APPENDIX C.)

Table 5. Means of perceived usability and subjective workload between groups.

Measure	Older OA	Younger OA	Students	Significance
Perceived Usefulness (PU)	46.88^{A}	40.70^{B}	40.18^{B}	0.021
Perceived Ease of Use (PEOU)	65.50^{A}	61.33^{B}	57.68^{B}	0.009
Perceived Ease of Learning (PEOL)	24.74	24.88	25.18	0.701
Satisfaction (SA)	34.00^{A}	27.69^{B}	29.46^{B}	0.007
NASA-TLX	36.96	32.33	29.92	0.254

NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study). OA is Older Adult. Significance is from a one-way ANOVA for differences among the row means. ^{AB} give the Tukey-post hoc comparison groups, different letters are significantly different from each other. Significance level is 0.05. (See the ANOVA tables and Tuckey-post doc results in APPENDIX C.)

6.4 Discussion

The study part one was designed to determine the effects of the design variables of search results layout and nutrition information format on subjective workload and perceived usability. The survey method used in this part of the study is the most frequently-seen method in related empirical usability evaluation studies for older adults (Zapata, Fernández-Alemán, Idri, & Toval, 2015). The method is widely used since it is the easiest way to collect a large amount of quantitative data in a short period of time by the self-rating style subjective questionnaire. Nevertheless, inconsistent rating style and different rating strategies from untrained users may induce the survey bias. In this study, several steps have been taken for reducing the impact of the within-subjects variance. For example, to ensure each participant followed a consistent rating rule, participants were encouraged to consider the rating of the first-encountered UI as the baseline to evaluate the subsequent UIs during the experiment. However, between-subject variability was still found as the dominating resource of the variance in this study. In this context, qualitative data collected by the think aloud method was reviewed to supplement the quantitative analysis following Holzinger's suggestion to combine the use of a subjective questionnaire with objective data to reveal human behavior (Holzinger, 2005). Each of these pieces of evidence is supportive to arriving at the final conclusion.

I originally expected that the choice-based UI would be thought to be easier to use especially for older adults. Contrary to my expectations, this study did not find a significant difference between choice-based and browsing-based UIs, but rather there were advocates for each UI. At least two reasons could give explanation to this contrary finding. First, it may be simply due to an underpowered study, and it is still likely the choice-based UI would be preferred by continually collecting opinions from a sufficient sample size of representative older adults. Although this study may have a relatively small sample size (twenty older adults), the second reason is considerably compelling. Findings from qualitative data suggested that at least two search strategies were being employed among participants. Some individuals tend to prefer exhaustively searching all the options to make an optimal choice. Other users adapted the satisficing rule to search and focus on picking a quality option. In the former cases, the list-based UI was preferred while in the later cases, the choice-based UI with limited alternatives was a natural preference. It is possible to hypothesize that an individual, especially a first-time user of the app, will decide which search strategy to use and which UI is most natural based on their previous experience on the most common application designs. For example, in our study, the computer-proficient participants preferred the list-view UI which is similar to Google's search engine UI. In addition, the conditionally adaption of search strategies suggest that an adaptive UI dynamically presents different type of layout when interacting with users could possibly be more useful than presenting either UI by itself.

Mixed-methods findings about nutrition information design suggested the Nutri-score label was generally preferred for ease of decision making as hypothesized. There is room for improvement for the current FSA Nutri-scores label to improve usability for U.S. customers who are more familiar with the government authorized Nutrition Facts Panel (NFP). First, the rating scale style should follow the general custom to meet U.S. customer's expectation (e.g. order of quality going from worst on the left to best on the right). Secondly, some additional information from the NFP would still be required to support diet management in a special health condition context (such as calorie count).

In terms of reducing biases for subjective questionnaire, conjoint measurement is suggested based on my observation from the study, which also reflects to the work of Sattler & Hensel-Börner (2001). It is easier for the novel user to respond to comparing questions rather than making unbiased judgement on an absolute scale. Further, this questionnaire style avoids the bias of social desirability which is a commonly found characteristic among older adult raters (Sattler & Hensel-Börner, 2001).

This part of the study employed subjective measures, but it is also possible to have objective data in this same setting. Previous studies measure human task performance or eye tracking data to support the assumptions of efficient UI design variables for older adults (Sharit, Hern, Czaja, HernándezHern, & Czaja, 2008; Wilson, 2011). However, the objective of this part of the study is to investigate the user preference of the innovative design for assuming the technology acceptance of older adults, so the subjective methods were selected. The study also could be viewed as a participatory design process since user statements are found useful to improve the current design. Study part two which take emphasis on verifying the effectiveness of persuasion is focused on human task performance measures.

7. PERSUASIVENESS OF THE PROPOSED APP

Partial results of the presented work in Chapter 7 have been published in page 342-348 of Advances in Intelligent Systems and Computing volume 1201 by Springer with the title "Evaluation of the Effectiveness of an Interpretive Nutrition Label Format in Improving Healthy Food Discrimination Using Signal Detection Theory" on July 16th, 2020. The article was co-authored with Dr. Mark Lehto, Dr. Brandon Pitts, and Dr. Zachary Hass.

7.1 Hypothesis

The goal of study two was to test impact of three nudge variables on human performance of choosing healthy food for a dietary mHealth app. Three prepositions and their corresponding hypotheses that were tested are presented below.

P1: Decision paradigm is a significant predictor of human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX).
P2: Nutrition information format is a significant predictor of human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX).
P3: System default is a significant predictor of human performance of decision making for healthy food (d', accuracy, c, time of response, weighted total of NASA-TLX).

RQ2: Which UI design elements effectively nudge users?
H2.1: Choice-based UI is significantly better than the searching-based UI to "nudge" users to select the system defined "truth".
H2.2: The specificity of nutrition information has a significant effect to "nudge" users to select the system defined "truth".
H2.3: The default nudge (existence of pre-selection) has a significant effect to "nudge" users to select the system defined "truth".

7.2 Results

Table 6 shows the summary of ANOVA analysis for each general linear model with outcomes of discriminability (d'), accuracy, response criterion (c), subjective workload (NASA-TLX) and the time of response. (See the ANOVA tables in APPENDIX D.) The independent variables are the three binary design variables and their two- and three-way interactions and a random intercept for subject. The results indicate that the 2AFC paradigm, absence of pre-selection, and the 2-way interactions between the 2AFC decision paradigm and the pre-selection were positively associated with human accuracy and the discriminability of healthy food.

Furthermore, pre-selection significantly reduces the time of response for the healthy food discrimination task. There's also a significant reduction of response time and subjective workload when using the interpretative nutrition label, FSA Nutri-Scores, compared to the current FDA Nutrition Facts Panel. However, there's an interaction between the nutrition information format and the pre-selection towards time of response. When the FSA Nutri-Scores was presented with the system default pre-selection, it significantly increases the time of response.

There is significant between-subjects difference on the subjective workload and the time of response. From the secondary analysis of individual difference (additional detail in the next section), overall there was a significant difference between the older and younger adults in the subjective workload and the time of response, but not in the human accuracy and the healthy food discriminability.

Another critical point that stands out from Table 6 is the time-accuracy trade-off. The factors associated with d' and accuracy are different to those associated with the time of response. Then again, the factors associated with subjective workload aligned with the time of response. The secondary analysis of Pearson correlations which showed positive correlation between the subjective workload and the time of response, as well as between discriminability and accuracy.

The residual plots of d' (Figure D.1), Accuracy (Figure D.2), c (Figure D.3), Time of Response (Figure D.4), and NASA-TLX (Figure D.5) show the proposed measures follow the normality assumption. (See the APPENDIX D.)

Table 6. Summary of GLMs for human decision performance

Coefficients of GLMs for		d’		Accuracy		c		NASA-TLX		Time of Response	
Source	**DF**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**	**Coef**	**p**
Decision Paradigm (2AFC = 1)	1	1.036	0.002	0.312	0.000	-0.289	0.000	1.510	0.575	1.470	0.865
Nutrition Information Format (Nutri-scores =1)	1	-0.092	0.776	-0.009	0.839	-0.066	0.063	-11.02	0.000	-31.98	0.000
System Default (Pre-selection=1)	1	-0.928	0.005	-0.133	0.003	-0.069	0.052	-2.430	0.371	26.410	0.003
Subjects	35		0.823		0.468		0.000		0.000		0.000
2AFC* Nutri-scores	1	0.789	0.085	0.100	0.104	0.079	0.110	-1.320	0.728	2.100	0.866
2AFC* Pre selection	1	0.944	0.041	0.150	0.015	0.081	0.104	-4.750	0.213	5.000	0.689
Nutri-scores* Pre selection	1	0.736	0.109	0.111	0.073	0.089	0.074	4.920	0.200	36.000	0.005
2AFC* Nutri-scores* Pre-selection	1	-1.100	0.089	-0.140	0.108	-0.137	0.050	3.720	0.489	-14.10	0.431
Error	237										
Lack-of-Fit	236		0.405		0.570		0.145		0.208		*
Pure Error	1										
Total	279										

d’ stands for d-prime (the sensitivity measures in this study); c is response criterion; NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study). 2AFC stands for Two-Alternative Forced Choice paradigm.

DF stands for Degree of Freedom; Coef stands for the coefficients of general linear models; p is p-value, the significance level is 0.05. (See the ANOVA tables in APPENDIX D. ANOVA TABLES FOR GLMS OF STUDY PART 2.)

7.3 Individual Difference

Individual difference such as age, computer proficiency, and health literacy may mediate human decision performance (Czaja et al., 2013). However, this study recruited on the rolling basis. To verify the mediating effect of individual difference, cluster analysis was conducted to identify the individual difference between groups for our sample (see the detail of the method in section 5.5.1). Based on age, computer proficiency scores and health literacy scores, the data could be further grouped by younger Older Adults (OA), older OAs, and students. Students have the highest computer proficiency and health literacy scores. They perform slightly better than younger OAs in computer proficiency and health literacy; the older OAs have significantly lower scores than the other two groups.

There is a significant difference between the three groups in response criterion (c), time of response, and subjective workload (NASA-TLX scores). The older OAs with limited health literacy and lower computer proficiency take a significantly longer time to discriminate the healthier food option; whereas the "younger" older adults with adequate health literacy and higher computer proficiency have nearly the same performance as the students. However, there were no significant differences between groups in human performance of accuracy and discriminability when deciding healthy food. But the response criterion of older adults (either older OAs or younger OAs) is significantly higher than students, which means they tend to adopt a liberal response criterion to suppose all the given options are healthy and accepted.

Table 7. Means of task performance and subjective workload between groups

Measure	Older OA	Younger OA	Students	Significance
d-prime; (d')	0.9180	0.9440	1.106	0.658
Accuracy	0.7107	0.7532	0.7368	0.675
Response criterion; c	0.2815^{A}	0.2571^{A}	0.1813^{B}	0.008
Time of Response	71.79^{A}	44.1^{B}	46.04^{B}	0.000
NASA-TLX	33.85	26.96^{B}	25.65^{B}	0.0013

NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study). OA is Older Adult. Significance is from a one-way ANOVA for differences among the row means. ^{AB} give the Tukey-post hoc comparison groups, different letters are significantly different from each other. (See the ANOVA tables and Tuckey-post doc results in APPENDIX E. ANOVA TABLES FOR INDIVIDUAL DIFFERENCE IN STUDY PART 2.)

The residual plots of d' (Figure E.1), Accuracy (Figure E.2), c (Figure E.3), Time of Response (Figure E.4), and NASA-TLX (Figure E.5) show the proposed measures follow the normality assumption. (See the APPENDIX E.)

7.4 Discussion

7.4.1 Impacts of Interpretative Nutrition Information

The results of the study suggest that the interpretative nutrition label (FSA Nutri-scores) significantly influences subjective workload and efficiency but not the effectiveness of persuasion. The label increased usability of the app in terms of reducing subjective workload and time of response. The findings aligned with the previous study results which suggested using interpretive nutrition information for adding significant value to a usable mHealth app (Burton-Jones, Grange, & Student, 2011; Zapata et al., 2015). Though caution must be applied for generalizability with limited sample size. Notably, the co-existence of pre-selection and the interpretative nutrition label significantly harm the efficiency by increasing the time of response. This finding could possibly be explained by the reluctance of accepting the pre-selection guidance, which is further discussed in section 7.4.2.

The exact impact of interpretative nutrition information on human decision performance is likely to vary across individuals. The interpretative nutrition information significantly improved the efficiency of older adults, especially older OAs, while it made no significant effect to students. Nevertheless, it helps to reduce workload significantly for students. This finding contributes to fill in the research gap of undetermined relationship between literacy, numeracy, health literacy and the use of nutrition label based on Malloy-Weir and Copper's (2017) systematic review (Malloy-Weir & Cooper, 2017).

7.4.2 Effectiveness of the Proposed Persuasive Nudge Design Elements

Although accuracy and d' were able to reflect the effects of the UI design variables in the same way, d' is a preferred metric for evaluating persuasiveness given its interpretation of discriminability and the unbiased property for imbalanced dataset. The measurement science of persuasiveness is further discussed in chapter 8.

Two alternative forced choice (2AFC) decision paradigm had the largest nudge effect on the user's decision-making behavior in terms of d' and accuracy, which supports hypothesis 2.1. Surprisingly, pre-selection was found to have a negative effect on accuracy and discriminability in contrast to the hypothesis 2.3. A theme of rebelling against the pre-selection came up from notes taken during observation supporting the unexpected quantitative results. Some subjects tried to avoid being nudged, especially when they regarded pre-selection as "choice manipulation" According to the observation notes, nearly half of the participants tried to avoid the pre-selected alternative once they have realized the existence of it. A few subjects mentioned in the interview that they didn't like to be nudged. For example, one interviewee mentioned, "*I would take more time to try to make my own decision since I don't want to be guided*.". This concern also partially explained the finding from quantitative results that the co-existence of pre-selection and Nutri-scores label increased the time of response.

This finding seems to be consistent with some criticisms of nudge, including: 1. the techniques used in nudge "works best in the dark" (Bovens, 2009)(Burgess, 2012) and that the effect of a nudge disappears if it is recognized by people.(Selinger & Whyte, 2010); 2. nudging harms user's autonomy and ability to make choices for themselves. However, most of the participants still agreed that the pre-selection is helpful from a time-saving viewpoint. Approximately a third of the participants followed the pre-selection after they have established the credibility of the pre-selection and began to trust the suggestions. Moreover, the success of 2AFC testing paradigm design in this empirical study, demonstrated the potential of digital nudge. It also implies that one of the essential features of nudge is the re-design of the choice architecture rather than the application of a collection of the "psychological manipulation" techniques. Nudge design is not just finding a hammer from the toolbox to fix all the problems, but more like fitting a screw into the system with the right screwdriver. The study reflects the importance of context analysis and verified the use of human factors methods in the proposed theoretical framework. Before jumping into adapting the nudge approach to a persuasive technology, human factors methods of cognitive work analysis and human error analysis should be applied to examine the critical breakthrough point to nudge case by case. In chapter eight, the theoretical framework is further discussed, the detail would be modified in order to be proposed as a human-centered digital nudge design framework.

8. GENERAL DISCUSSION

8.1 Measurement Science of Persuasive Technology

8.1.1 Evaluate Persuasiveness via Signal Detection Analysis versus Confusion Matrix

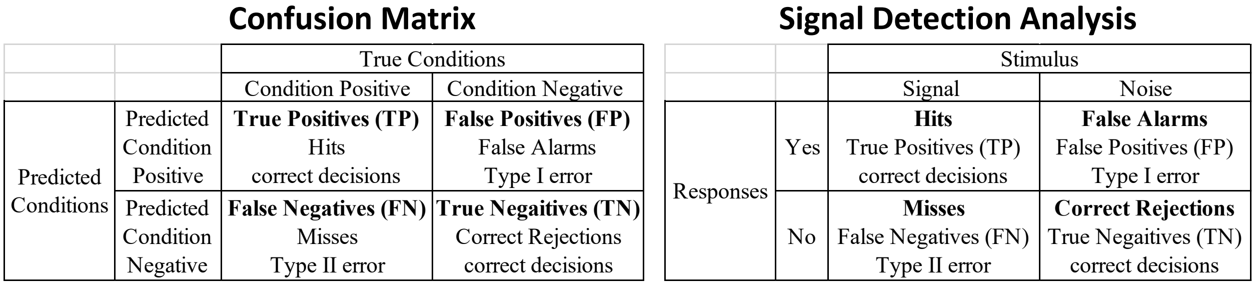
Measuring human performance of decision making for system evaluation could be tricky since human bias is involved, and thus, the variances from the measures are confounded with the variances from the subjects. For example, proportion of correct is an intuitive but poor measure of sensitivity for the yes/no experiment of discrimination if the responder is biased (Green & Swets, 1988; Macmillan & Creelman, 2005). In this dissertation, discrimination if the responder is biased (Green & Swets, 1988; Macmillan & Creelman, 2005). In this dissertation, discriminability and response criterion adapted from the signal detection theory is suggested as the primary measure of human performance of decision making. And in section 7.4.2, I have briefly mentioned that discriminability is preferred, rather than accuracy which is adapted from the confusion matrix, for evaluating the information quality based on my experimental data.

In comparison to the proposed signal detection measures, many Artificial Intelligence (AI) researchers may think of recall and precision measures on confusion matrix for cross validation of binary classification algorithm. Since there are some similarities between the signal detection analysis and the confusion matrix: Both of them are the metrics of decision science, used to evaluate AI / human performance of labeling / predicting the class. So they both apply the contingency table of statistical decision theory to calculate the proportion of correct decisions, type I error, and type II error in order to estimate the performance of decision making based on Deming's concept of precision and accuracy (here I denoted them by Deming's precision and accuracy to distinguish from the terminologies of the measures, "accuracy" and "precision" on the confusion matrix.). However, due to the differences between the dataset characteristics and testing themes and data collection procedures, some metrics which are frequently used in machine learning discipline for algorithm performance assessment actually cannot be adapted to the human evaluation context. For example, the sensitivity on confusion table, which is the proportion of correct and defined as the True Positive (TP) rate, is not taking errors into account and thus insensitive to biased human raters.

In previous sections, I had explained why discriminability was proposed to evaluate the information quality and why accuracy was used for assessing the effectiveness of nudge approaches based on the idea of Deming's precision and accuracy. The same concept could also be applied to the general decision-making theme in human-computer interactions scenario. In the following paragraphs, few more metrics on confusion matrix would be introduced and discussed in comparison with the proposed metrics in my study. The empirical data has justified the use of signal detection theory in this study and similar human factors evaluation theme. An implication of this is the possibility that some arguments about the biased confusion matrix measures in some specific cases probably could be resolved and reconsolidated by adapting the metrics developed in signal detection theory, which reflects Powers' (2003, 2011) suggestion about using of Informedness, Markedness, and ROC curve analysis as the system decision performance metrics (Powers, 2003, 2011).

Confusion Matrix

In the Machine Learning discipline, especially for supervised learning, the performance of classification algorithms is evaluated by the confusion matrix of the cross-validation results. The confusion matrix adapts the statistical decision theory with the contingency table to denote four cases of type I, type II errors and correct decisions by True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN). TP denotes those cases of correctly predicting the positive condition, which is the same thing to hits in signal detection theory; FP denotes those cases of wrongly predicting the positive condition, which is the same thing to False Alarms (FAs) in signal detection theory and also the Type I error; FN denotes those cases of wrongly predicting the negative condition, which is the same thing to misses in signal detection theory and also the Type II error; TN denotes those cases of correctly predicting the negative condition, which is the same thing to correct rejections in signal detection theory. Below is the cheat sheet to readers for quickly comparing between the terms to avoid the confusion.



Confusion Matrix

		True Conditions	
		Condition Positive	Condition Negative
Predicted Conditions	Predicted Condition Positive	**True Positives (TP)** Hits correct decisions	**False Positives (FP)** False Alarms Type I error
	Predicted Condition Negative	**False Negatives (FN)** Misses Type II error	**True Negaitives (TN)** Correct Rejections correct decisions

Signal Detection Analysis

		Stimulus	
		Signal	Noise
Responses	Yes	**Hits** True Positives (TP) correct decisions	**False Alarms** False Positives (FP) Type I error
	No	**Misses** False Negatives (FN) Type II error	**Correct Rejections** True Negaitives (TN) correct decisions

Figure 30. Comparison between the terms of the confusion matrix and the signal detection analysis

Frequent-Used Metrics for Classification Assessment

It's similar to Signal Detection Theory, some empirical metrics derived from TP, FP, TN, FN are then used to assess the algorithm performance for classification problems. The frequent-used ones including:

1. Sensitivity and Specificity:

Sensitivity and specificity are two terms that indicate the true positive rate (hit rate in SDT) and the true negative rate (correct rejection rate in SDT) respectively, which are directly adapted from the statistical decision theory. Sensitivity is mostly referred to as Recall and specificity as selectivity when they are mentioned in the assessment of machine learning algorithms. Below are the formulas of sensitivity and specificity:

- Sensitivity (Recall; True Positive Rate):

TP / (TP+FN) = Hits / (Hits + Misses) = Hit Rate

- Specificity (Selectivity; True Negative Rate):

TN / (TN+FP) = CRs / (CRs + FAs) = Correct Rejection Rate

2. Accuracy and Precision

When a new research instrument is built, it's always important to examine the reliability and the validity of the tool. Reliability means the instrument could measure the outcomes consistently while validity means the instrument could measure the outcomes accurately. The similar idea is applicable to other scientific methods, when comparing measured values with the standard, the results could be further analyzed by two indicators: precision and accuracy. Precision measures the closeness between measurements whereas accuracy measures the closeness between the measurements and the target. A classic example to explain precision and accuracy is the dartboard, accuracy is how close does the player shoot near the bullseye while precision is how consistent does the player shoot on the same target.

The formulas of accuracy and precision are listed below:

- Accuracy:

  (TP+TN) / (TP+FP+TN+FN) = (Hits + CRs) / (Hits + Misses + CRs + FAs)

- Precision (Positive Predictive Rate):

  TP/ (TP+FP) = Hits / (Hits + False Alarms)

3. F1-scores (harmonic mean of recall and precision):

   When assessing the machine learning algorithm, recall is frequently used rather than accuracy. However, to better consider recall and precision at once, the harmonic mean of recall and precision was proposed and named as F1-scores. The formula of F1-scores is listed as below:

2*(precision * recall) / (precision + recall) = 2TP/(2TP+FP+FN) = 2*Hits / (2*Hits +FAs + Misses)

From the formulas of accuracy and F1-scores, one could easily learn that accuracy take the “True” system performance into account while the “False” system performance would also be considered in F1-scores. In the real-world practice, many researchers prefer F1-scores than accuracy since in many scenarios the FP and FN are also crucial.

Assessment of Classification Algorithms with the Imbalanced Dataset

However, some concerns have been raised about these frequent-used metrics, including recall precision, and accuracy, as they are biased measures to evaluate the performance of the biased classifiers on an imbalanced dataset. Imbalanced data means the observations of each class is not evenly distributed. For example, in binary classification, there is a huge amount of data from the so-called majority class, and the rest in the minority class. In supervised machine learning, imbalanced training dataset would result in a biased classifier which would favor the frequent-seen class in the binary classification.

Brodesen et al. (2010) warned that in the fold-wise cross validations, the average accuracy of all folds is a non-parametric point estimator so the confidence interval can’t be derived, and it misleads to an optimistic estimation of the biased classifier on an imbalanced dataset. They suggest using the posterior distribution of balanced accuracy instead (Brodersen, Ong, Stephan, & Buhmann, 2010). Powers (2011) has also pointed out recall, precision, F-measures, and accuracy are biased variants, especially for imbalanced data. He further suggested to adopt Informedness (probability that a prediction is informed in relation to the condition) rather than recall and to use Markedness (the probability that a condition is marked by the predictor) rather than precision. (Powers, 2011).

1. Restore the data balance

   In the real-world practice of developing machine learning algorithms, many researchers would try either oversampling the minority class or undersampling the majority by bootstrapping (random sampling with replacement), SMOTE, or emsemble learning,…etc. to restore the balance of the training set (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Japkowicz & Stephen, 2002; M. Zhu, Xu, & Wu, 2013).

2. Balanced Accuracy

   However, as Brodersen et al. (2010) have mentioned, oversampling and undersampling worked to prevent the biased classifier under specific conditions but not a generic method to avoid an optimistic estimation of accuracy. They worked on the posterior distribution of accuracy and find the balanced accuracy is an unbiased estimator of accuracy in terms of the generalizability. The balanced accuracy is defined as:

   $$(TP / (TP + FN) + TN / (FP + TN))/2 = (\text{Hit Rate} + \text{Correct Rejection Rate}) /2$$

3. Bookmaker Informedness and Markedness

   Despite the fact that most studies in machine learning discipline use recall and precision or their synthetic metric, F1-scores, as the primary measures to assess the algorithm. One of the major concerns is that the recall and precision measures ignore the cost of error, similar to the baseline of guessing (Powers, 2003). Powers (2003) suggested the bookmaker evaluation technique which estimate accuracy in a betting scenario assuming guessing (random choice) would be given 0 gain, perfectly correct performance would be giving maximum gain and perfectly incorrect performance would get a maximum loss; and making a perfect correct decision G% of the time and guessing otherwise would gains G% of the maximum gain. In other words, reconsidered the contingency table as the combination of the guesswork matrix and the perfect decision matrix. And the bookmaker Informedness, which is percentage of the time we are using definitive information to make a correct decision rather than just guessing, G% can be calculated.

For the binary case, Informedness could be simplified as TPR–FPR from the formula:

$$\text{Recall} + \text{Inverse Recall} - 1 = \text{TPR}-\text{FPR} = 1-\text{FNR}-\text{FPR} = \text{Hit Rate} - \text{FA Rate};$$

Although deducted from the different formula in different viewpoints, there is a high similarity between the Bookmaker Informedness and the discriminability as both measures are the difference between the hit rate (TPR) and the false alarm rate (FPR) to measure the probability of an informed decision. A slight difference between the two, is the formula of discriminability is Z(Hit rate)-Z(FA rate) since it's assumed that the responses followed the normal distribution.

Other than Bookmaker Informedness Powers (2011) has also suggested to replace precision with Markedness, which is the chance of the conditions being marked by predictors. For the binary case, Markedness is defined as:

$$\text{Precision} + \text{Inverse Precision} - 1 = \text{TP}/(\text{TP}+\text{FP}) - \text{FN}/(\text{FN}+\text{TN})$$
$$= 1-\text{FP}/(\text{TP}+\text{FP}) - \text{FN}/(\text{FN}+\text{TN})$$

Why Signal Detection Measures

Recall and precision, or F1-scores measures are currently the most popular methods for cross-validation to assess the AI/ machine learning algorithm performance of the classification problem. However, the cross-validation sampling of the dataset from big data, which is usually collected by automated measures, and thus the imbalanced dataset could be easily dealt with re-sampling since dataset is sufficient; whereas human factors evaluation is based on the scarce human task performance data which was collected from the human subjects in the lab with a higher cost of data collection. First of all, human bias should be expected in an experiment with human subjects involved. The same rater could make different responses every time and thus the proportion of correct responses is confounded with the response criterion making the hit rate a biased metric of sensitivity, but the discriminability, which is the distance between the mean of the signal distribution and the mean of the noise distribution, is an unbiased metric of sensitivity in this context (Stanislaw, 1999). Secondly, limited by the human capacity of performing a task repetitively, the experimental data must be relatively insufficient compared to the big data used for machine learning. In this context, every data point is important and can't be easily discarded. Additionally, human factors evaluation looking for more insights about human behavior from the metrics rather than a go/no-go criterion to judge the performance. So, it would prefer discriminability which not only consider "true" data (which are correct decisions) but also keep those "false" data (which are type I and type II errors). Based on the above reasons, signal detection analysis is ideal since it provides explainable metrics derived from both true and false data.

Moreover, precision is usually not a main concern in a human decision-making experiment over the binary classes since the within-subject variance would always exist. Instead, response criterion is considered and Receiver Operating Characteristic (ROC) curve, which is a plot of the sensitivity (hit rate) and specificity (false alarm rate) and the graphical presentation of discriminability over different response criterion, is thought more informative in psychology and diagnosis test (Hanley & McNeil, 1982; Swets, 2014).

The above discussion has justified the use of signal detection measures instead of the recall and precision measures from the confusion matrix. In the next section, an empirical comparison between signal detection measures and confusion matrix measures was conducted using the data from the study.

Empirical Comparison between Signal Detection Metrics and Confusion Matrix Metrics

In order to justify the use of the proposed measures in this study, an empirical comparison between signal detection metrics and confusion matrix metrics based on study result was conducted.
The table below shows sample data from a single subject across all UIs to demonstrate the calculation of SDT measures and Confusion Matrix measures. Treatment 1 is the combination of the design variables, yes/no paradigm, FDA Nutrition Facts Panel, and no pre-selection; Treatment 2 is the combination of the design variables, yes/no paradigm, FDA Nutrition Facts Panel, and with the pre-selection; Treatment 3 is the combination of the design variables, yes/no paradigm, FSA Nutri-Scores Label, and no pre-selection; Treatment 4 is the combination of the design variables, yes/no paradigm, FSA Nutri-Scores Label, and with pre-selection; Treatment 5 is the combination of the design variables, Two-Alternatives Forced Choice (2AFC) paradigm, FDA Nutrition Facts Panel, and no pre-selection; Treatment 6 is the combination of the design variables, 2AFC paradigm, FDA Nutrition Facts Panel, and with the pre-selection; Treatment 7 is the combination of the design variables, 2AFC paradigm, FSA Nutri-Scores Label, and no pre-selection; Treatment 8 is the combination of the design variables, 2AFC paradigm, FSA Nutri-Scores Label, and with pre-selection. The table illustrates that SDT measures are more discriminable between treatments in this study, even for a single observation. One reason is that there are limited trials for each treatment in this study, which makes it more likely that responses would result in an extreme number of 0 or 1 for many Confusion Matrix measures (which is not informative for small sample sizes). For example, recall, F1-scores, and Informedness are nearly 0 for treatment 2 to 4, because TP cases were rare for this subject. However, SDT considers all cases and projected the numbers to the normal distribution smoothing out this effect. So, the metrics are not affected than much by uneven distributed responses.

SDT measures also give more explainable numbers. For example, the negative number of c represents the subject's decision criterion is on the left-hand side of the neutral criterion, 0, which indicates this subject adopted a liberal decision strategy for a particular treatment; and the positive number of c indicates the subject adopted a conservative decision strategy for a particular treatment.

Table 8. Calculations of SDT Measures and Confusion Matrix Measures

	SDT Measures		Confusion Matrix Measures				
Sample Data	**d-prime**	**c**	**accuracy**	**recall**	**precision**	**F1-scores**	**Informedness**
Formula	Z(Hit rate) - Z(FA rate)	- (Z(Hit rate)+Z(FA rate))/2	(TP+TN) / (TP+FP+TN+FN)	TP / (TP+FN) =TP rate	TP/ (TP+FP)	2TP/(2TP+FP+FN)	TP rate - FP rate
Treatment 1	0.602	-0.468	2.400	0.667	1.000	0.800	2.000
Treatment 2	0.674	1.168	0.200	0.000	0.500	0.000	0.000
Treatment 3	0.659	1.148	0.400	0.000	0.500	0.000	0.000
Treatment 4	0.659	1.148	0.400	0.000	0.500	0.000	0.000
Treatment 5	0.602	-0.468	2.200	0.667	0.667	0.667	1.000
Treatment 6	0.602	-0.468	2.400	0.667	1.000	0.800	2.000
Treatment 7	1.582	-1.724	1.800	1.000	1.000	1.000	1.000
Treatment 8	1.621	-1.774	2.600	1.000	1.000	1.000	2.000

The table below shows the summary of ANOVA analysis for each general linear model of signal detection metrics, discriminability (d'), accuracy, response criterion (c); and confusion matrix metrics, recall, balanced accuracy, precision, and F1-scores. It is apparent from this table that recall and balanced accuracy may not be appropriate metrics to discover UI design effects on human performance as attested by the significance of lack of fit.

As the result section has mentioned, discriminability and accuracy were sensitive to the changes of decision paradigm, system default selection, and their two ways interactions, and so too precision. It's also worthwhile to notice that the metrics such as balanced accuracy and F1-scores which were synthesized from recall and precision to estimate accuracy and precision for imbalanced dataset, in this context lost its discriminability of different design elements.

Table 9. Signal detection metrics measures.

Signal Detection Metrics							
Coefficients of GLMs for		**d'**		**Accuracy**		**c**	
Source	**DF**	**Coef**	**P-Value**	**Coef**	**P-Value**	**Coef**	**P-Value**
Decision Paradigm (2AFC =1)	1	1.036	0.002	0.312	0.000	-0.289	0.000
Nutrition Information Format (Nutri-scores = 1)	1	-0.092	0.776	-0.009	0.839	-0.066	0.063
System Default (Pre-selection = 1)	1	-0.928	0.005	-0.133	0.003	-0.069	0.052
Subjects	35		0.823		0.468		0.000
2AFC * Nutri-scores	1	0.789	0.085	0.100	0.104	0.079	0.110
2AFC * Pre-selection	1	0.944	0.041	0.150	0.015	0.081	0.104
Nutri-scores * Pre-selection	1	0.736	0.109	0.111	0.073	0.089	0.074
2AFC * Nutri-Scores * Pre-selection	1	-1.100	0.089	-0.140	0.108	-0.137	0.050
Error	237						
Lack-of-Fit	236		0.405		0.570		0.145
Pure Error	1						
Total	279						

d’ stands for d-prime (the sensitivity measures in this study); c is response criterion; NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study). 2AFC stands for Two-Alternative Forced Choice paradigm. DF stands for Degree of Freedom; Coef stands for the coefficients of general linear models; p is p-value, the significance level is 0.05.

Table 10. Confusion matrix metrics.

Confusion Matrix Metrics									
Coefficients of GLMs for		Precision		Recall		F1		Balanced Accuracy	
Source	**DF**	**Coef**	**P-Value**	**Coef**	**P-Value**	**Coef**	**P-Value**	**Coef**	**P-Value**
Decision Paradigm (2AFC = 1)	1	0.241	0.000	-0.146	0.001	0.090	0.105	0.322	0.000
Nutrition Information Format (Nutri-scores = 1)	1	0.008	0.883	-0.016	0.727	0.005	0.928	0.011	0.739
System Default (Pre-selection = 1)	1	-0.137	0.012	-0.028	0.532	-0.132	0.018	-0.009	0.771
subjects	35		0.848		0.395		0.848		0.457
2AFC * Nutri-scores	1	0.126	0.097	0.120	0.057	0.147	0.059	0.102	0.027
2AFC * Pre-selection	1	0.173	0.024	0.092	0.144	0.195	0.013	0.036	0.430
Nutri-scores * Pre-selection	1	0.101	0.185	0.045	0.471	0.108	0.167	0.002	0.969
2AFC * Nutri-scores * Pre-selection	1	-0.190	0.076	-0.187	0.035	-0.247	0.02	-0.094	0.146
Error	237								
Lack-of-Fit	236		0.473		0.005		0.418		0.007
Pure Error	1								
Total	279								

d' stands for d-prime (the sensitivity measures in this study); c is response criterion; NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study). 2AFC stands for Two-Alternative Forced Choice paradigm. DF stands for Degree of Freedom; Coef stands for the coefficients of general linear models; p is p-value, the significance level is 0.05. (See the ANOVA tables in APPENDIX F. ANOVA TABLES FOR CONFUSION MATRIX METRICS.)

Table 11 summarize the differences between SDT measures and confusion matrix measures. For reader's better understanding, I presented both SDT measures and confusion matrix measures by SDT terms: Hits, False Alarms (FA), Correct Rejections (CR), and Misses; instead of using terms True Postive (TP), False Positive (FP), True Negative (TN), False Negative (FN). Figure 30 summarizes the translation of the terms on confusion matrix to SDT terms.

The first difference between SDT measures and confusion matrix measures is the application. SDT metrics measure human decision performance while confusion matrix metrics measure system decision performance. The former metrics take human bias and the chances of guessing into account and used estimators independent of human bias. The assumption of SDT measures is human responses follows Gaussian distributions and the equal variance of responses to signal and responses to noise distributions. Confusion matrix metrics are usually calculated based on big data, and the metrics have been criticized as being biased estimators for imbalanced dataset.

In this study, d-prime, accuracy, precision yield the same result of being better able to discriminate UI difference, whereas recall and F1-scores are not that sensitive to UI differences and the results are relatively unexplainable.

Table 11. Summary of the Comparison Between SDT Measures and Confusion Measures

	SDT Measures		Confusion Matrix Measures				
	d-prime	**c**	**accuracy**	**recall**	**precision**	**F1-scores**	**Informedness**
Formula (presented by Hit rate / FA rate)	Z(Hit rate) - Z(FA rate)	- (Z(Hit rate)+Z(FA rate))/2	(Hits + CRs) / (Hits + Misses + CRs + FAs)	Hits / (Hits + Misses) = Hit Rate	Hits / (Hits + False Alarms)	2*Hits / (2*Hits +FAs + Misses)	Hit rate - FA rate
Application	Human decision performance	Human decision performance	System decision performance	System decision performance	System decision performance	System decision performance	Human / System decision performance

Table 11 continued

Assumption	1. Human responses follow Gaussian distributions 2. Equal variances between signal and noise distributions.	1. Human responses follow Gaussian distributions 2. Equal variances between signal and noise distributions.	big data	big data	big data	big data	1. Payoff function of X:Y for Expected Value Analysis 2. Hit rate -FA rate is only valid for the binary classification case.
Human bias	Independent of human bias	Independent of human bias (reference of human bias)	?	?	?	?	Consider guessing
Imbalanced Data			Bias estimator	Unexplainable with imbalanced data	Unexplainable with imbalanced data	Unexplainable with imbalanced data	
Detect UI differences in this study	Better	Not good, but a nice reference to individual's decision strategy	Better and align with SDT measures	Not good, and unexplainable	Better and align with SDT measures	Not good, and unexplainable	Better and align with SDT measures

8.1.2 Measurement Science of the Usability of Persuasive mHealth apps

Usability is a multidimensional construct of effectiveness, efficiency, and satisfaction for the specific users to achieve the specific goals in a distinct environment (International Organization for Standardization [ISO], 2018). In this dissertation, multiple human factors measures are used to evaluate the usability of a proposed persuasive mHealth app, including perceived usability, subjective workload, and human task performance. In study part 1, perceived usability and subjective workload were the primary measures to evaluate usability. Lund's USE questionnaire was used since different dimensions of usability was considered as distinct constructs which would give the designer more insights. NASA-TLX is also a multidimensional self-assessment tool to measure the perceived workload and researchers also agreed on its usage to assess the task, system, and team effectiveness and performance. For example, Hornbæk (2006) found nine usability studies during 1999-2002 use NASA-TLX as an efficiency measure (Hornbæk, 2006); Paas and van Merriënboer (1993) suggested mixed-methods of objective task performance measures and subjective mental effort measures to assess the relative task efficiency. However, the existing subjective questionnaire constructs are still not enough to explain the effectiveness of persuasive technology in the aspect of influencing end users' behavior. And thus, study part 2 was proposed to evaluate the usability in a sense of the effectiveness of persuasion based on the human factors evaluation: the human task performance and workload measures of the human decision-making theme.

In this section, the link between those measures and the latent human factors constructs including usability, (mental) workload, and human performance of decision making is furthered explored. However, since the interruption of data collection due to the COVID-19 pandemic and the fact that some participants have only finished part 1 or part 2 of the study, with a smaller sample size, these results should be interpreted with caution around generalization. Further, since the stimuli and the task scenario in study part 1 and study part 2 are slightly different, data from two parts of study can't be linked together. The perceived usability measures in study part 1 has no direct relationship with the task performance measures in study part 2. So, the relationship between perceived usability and human performance of decision making is excluded from this discussion.

The Relationship between Perceived Usability and Subjective workload

To examine the relationship between perceived usability and subjective workload, a correlation analysis between each perceived usability construct (PU, PEOU, PEOL, SA) and subjective workload (NASA-TLX) was firstly conducted. The correlation matrix of Pearson's coefficients is shown as the table below: All the constructs are significantly correlated, NASA-TLX scores are moderately and negatively correlated with the perceived usability constructs, PU, PEOU, PEOL, and Satisfaction with the scores between -0.3 and -0.5.

Table 12. the correlation matrix between perceived usability and subjective workload

Correlations	USE_PU	USE_PEOU	USE_PEOL	USE_Satisfication
USE_PEOU	0.781***	-		
USE_PEOL	0.402***	0.619***	-	
USE_Satisfication	0.883***	0.757***	0.406***	-
NASA_TLX	-0.374***	-0.46***	-0.374***	-0.323***
* p <0.05 ; ** p<0.01; *** p<0.000				

USE_PU stands for Perceived Usefulness adapted from Lund's USE questionnaire; USE_PEOU stands for Perceived Ease of Use adapted from Lund's USE questionnaire; USE_PEOL stands for Perceived Ease of Learnding adapted from Lund's USE questionnaire; USE_Satisfaction stands for statisfaction adapted from Lund's USE questionnaire. All of them are perceived usability measures in this study. NASA-TLX stands for NASA Task Load Index (the subjective workload measure in this study).

The above result disagrees with Longo's finding that subjective workload and perceived usability are uncorrelated for the information seeking task. As the previous section has discussed, there are some possible explanations for the contradiction: first of all, the difference between the measures of perceived task; secondly, the difference between the task types and the complexity; besides, the sample size of this study is relatively smaller, half of them are older adults, and the subjective workload is just moderately correlated to the perceived usability in this study.

To better estimate the structural coefficients between the measures and the latent construct, perceived usability, a reflective measurement model was established as Figure 31. The exploratory factors analysis was performed, and the standardized estimates and the R-squares are shown on each path.

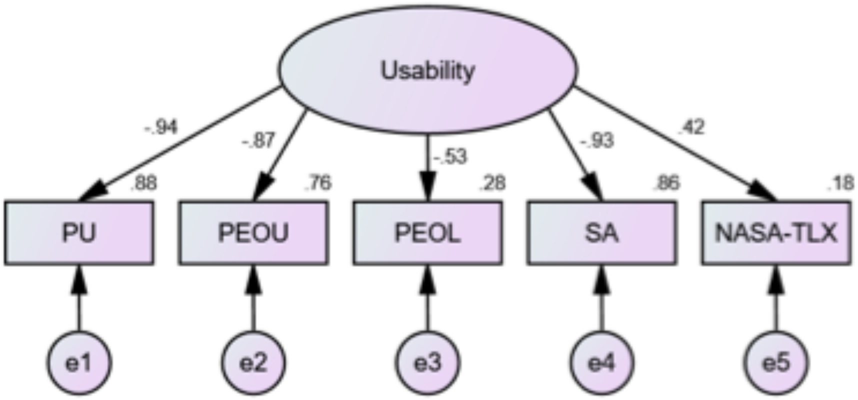


Figure 31. Reflective measurement model

However, the probability value of chi-square test is less than 0.05 which suggests the rejection of the null hypothesis that the model fits the data. According to the model fits summary below, the RMSEA for this model is 0.302 and the Tucker-Lewis Index (TLI) value is 0.742 and the Comparative Fit Index (CFI) value is 0.871, which doesn't follow Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9. So, the model was empirically modified by adding the constraints of correlated residuals as Figure 32 shown.

Table 13. Notes of Model for the reflective measurement model

Notes for Model (Default model)	
Computation of degrees of freedom (Default model)	
Number of distinct sample moments:	20
Number of distinct parameters to be estimated:	15
Degrees of freedom (20 - 15):	5
Result (Default model)	
Minimum was achieved	
Chi-square = 82.179	
Degrees of freedom = 5	
Probability level = .000	

Table 14. Model fits summary of the reflective measurement model

Model Fit Summary					
CMIN					
Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	15	82.179	5	0	16.436
Saturated model	20	0	0		
Independence model	10	608.952	10	0	60.895
Baseline Comparisons					
Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	0.865	0.73	0.872	0.742	0.871
Saturated model	1		1		1
Independence model	0	0	0	0	0
RMSEA					
Model	RMSEA	LO 90	HI 90	PCLOSE	
Default model	0.302	0.247	0.361	0	
Independence model	0.595	0.556	0.636	0	

The thresholds adapted Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

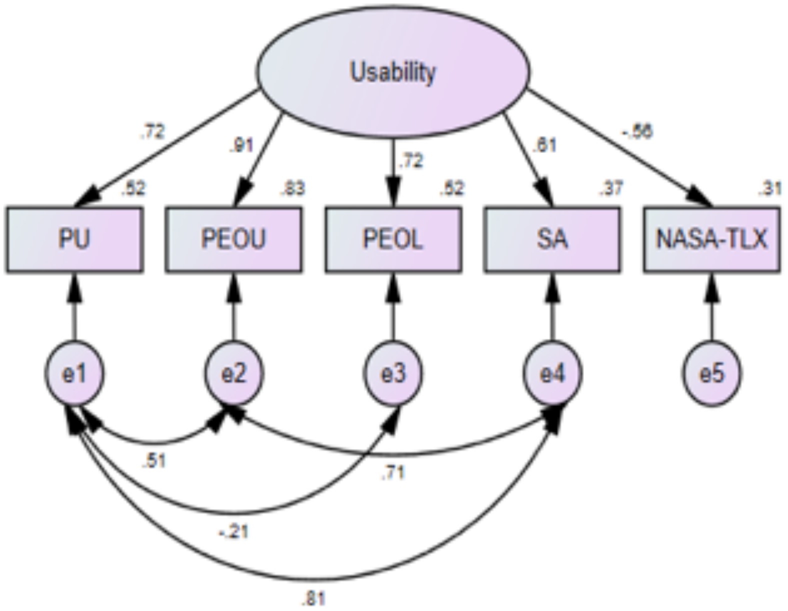


Figure 32. modified reflective measurement model

Table 15 below shows, the probability value of chi-square test is 0.485 which suggests the model fits the data. And according to the model fits summary, the RMSEA for this model is 0.000 and the Tucker-Lewis Index (TLI) value is 1.009 and the Comparative Fit Index (CFI) value is 1.000, which follow Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

Table 15. Notes of Model for the reflective measurement model

Notes for Model (Default model)	
Computation of degrees of freedom (Default model)	
Number of distinct sample moments:	20
Number of distinct parameters to be estimated:	19
Degrees of freedom (20 - 19):	1
Result (Default model)	
Minimum was achieved	
Chi-square = .487	
Degrees of freedom = 1	
Probability level = .485	

Table 16. Model fits summary for the modified reflective measurement model

Model Fit Summary					
CMIN					
Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	19	0.487	1	0.485	0.487
Saturated model	20	0	0		
Independence model	10	608.952	10	0	60.895
Baseline Comparisons					
Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	0.999	0.992	1.001	1.009	1
Saturated model	1		1		1
Independence model	0	0	0	0	0
RMSEA					
Model	RMSEA	LO 90	HI 90	PCLOSE	
Default model	0	0	0.179	0.57	
Independence model	0.595	0.556	0.636	0	

Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

Table 17. Regression Weights

Regression Weights: (Group number 1 - Default model)	Estimate	Standard Estimates	S.E.	C.R.	P	Label
NASA_TLX <--- Usability	-0.919	-0.556	0.153	-6.007	***	par_1
USE_Satisfication <--- Usability	0.508	0.61	0.052	9.866	***	par_2
USE_PEOL <--- Usability	0.211	0.723	0.031	6.902	***	par_3
USE_PEOU <--- Usability	1	0.909				
USE_PU <--- Usability	0.682	0.72	0.069	9.878	***	par_4

Besides correlations, it's very likely that there are causal relationships between the measures. To explore the possible causal relationships, a sequence of stepwise regression analysis trials was performed by selecting each measure as the response and the other measures as the predictors. The hypothetical causal network was drawn as Figure 33. The hypothetical network seems partially credible since the causal relationship between PEOU and PU has already been verified by Technology Acceptance Model (TAM) (Davis, 1985), and the remainder basically follows the TAM framework that the external variables (learnability and workload) would affect the PEOU and thus PU; PU would affect the attitude (satisfaction) and thus the intention of use.

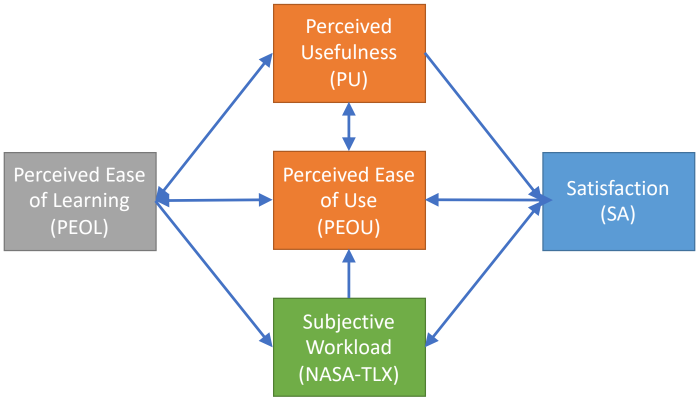

Perceived Usefulness (PU)
Perceived Ease of Learning (PEOL)
Perceived Ease of Use (PEOU)
Satisfaction (SA)
Subjective Workload (NASA-TLX)

Figure 33. Hypothetical causal relationships between subjective workload and each construct of perceived usability

To verify the hypothetical framework, structural equation modeling analysis was utilized. However, due to the limitation of degree of freedom, the reciprocal effects were firstly ignored and a recursive structural equation model with unidirectional causations was assumed as shown in Figure 34. The narrative of hypothetical framework could be written as:
PEOL-->NASA_TLX-->PEOU-->PU-->Satisfaction

H1: An increasing Perceived Ease of Learning (PEOL) would increase Perceived Usefulness (PU).
H2: An increasing PEOL would increase Perceived Ease of Use (PEOU).
H3: An increasing PEOL would reduce NASA-TLX scores (NASA-TLX).
H4: A reducing NASA-TLX scores would increase Perceived Ease of Use (PEOU).
H5: A reducing NASA-TLX scores would increase Satisfaction (SA).
H6: An increasing Perceived Ease of Use (PEOU) would increase Perceived Usefulness (PU).

H7: An increasing Perceived Ease of Use (PEOU) would increase Satisfaction (SA).

H8: An increasing Perceived Usefulness (PU) would increase Satisfaction (SA).

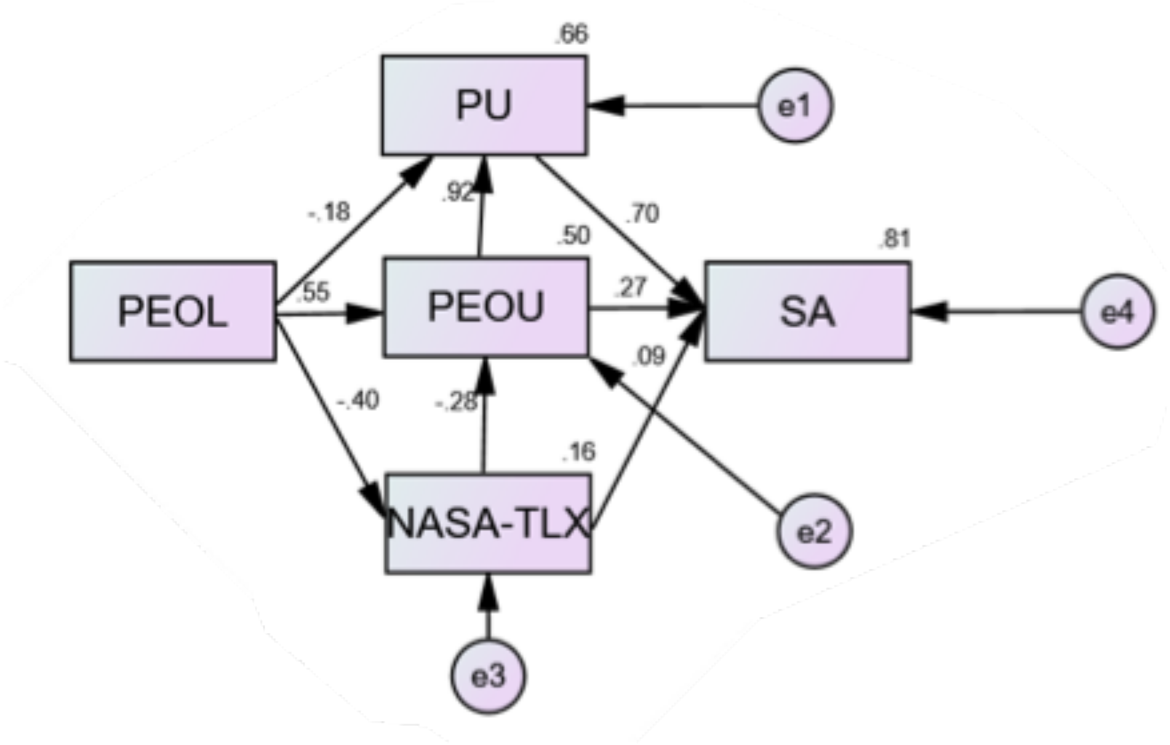


Figure 34. Recursive structural model of usability measures

The below table of model fits summary failed to reject the hypothetical framework since the probability value of chi-square test is 0.875, which is much higher than 5% significance level and suggests non-rejection of the null hypothesis that the proposed structural equation model fits the data; the RMSEA for this model is 0.000, which is way smaller than 0.06; the Tucker-Lewis Index (TLI) value is 1.014, which is higher than 0.95; and the Comparative Fit Index (CFI) value is 1.000, which is higher than 0.9 based on the Hu and Bentler’s guidelines.

Table 18. Notes of the recursive structural model

Notes for Model (Default model)	
Computation of degrees of freedom (Default model)	
Number of distinct sample moments:	20
Number of distinct parameters to be estimated:	18
Degrees of freedom (20 - 18):	2
Result (Default model)	
Minimum was achieved	
Chi-square = .268	
Degrees of freedom = 2	
Probability level = .875	

Table 19. Model fits summary for the recursive structural model

<table>
<tr><th colspan="5">Model Fit Summary</th></tr>
<tr><th colspan="5">CMIN</th></tr>
<tr><td>Model</td><td>NPAR</td><td>CMIN</td><td>DF</td><td>P</td></tr>
<tr><td>Default model</td><td>18</td><td>0.268</td><td>2</td><td>0.875</td></tr>
<tr><td>Saturated model</td><td>20</td><td>0</td><td>0</td><td></td></tr>
<tr><td>Independence model</td><td>10</td><td>608.952</td><td>10</td><td>0</td></tr>
<tr><th colspan="5">Baseline Comparisons</th></tr>
<tr><td rowspan="2">Model</td><td>NFI</td><td>RFI</td><td>IFI</td><td>TLI</td></tr>
<tr><td>Delta1</td><td>rho1</td><td>Delta2</td><td>rho2</td></tr>
<tr><td>Default model</td><td>1</td><td>0.998</td><td>1.003</td><td>1.014</td></tr>
<tr><td>Saturated model</td><td>1</td><td></td><td>1</td><td></td></tr>
<tr><td>Independence model</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><th colspan="5">RMSEA</th></tr>
<tr><td>Model</td><td>RMSEA</td><td>LO 90</td><td>HI 90</td><td>PCLOSE</td></tr>
<tr><td>Default model</td><td>0</td><td>0</td><td>0.076</td><td>0.915</td></tr>
<tr><td>Independence model</td><td>0.595</td><td>0.556</td><td>0.636</td><td>0</td></tr>
</table>

Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

Table 20 further shows the significance of all the estimated regression weights, which confirmed the hypothesis H2-H4 and H6-H8. However, contrary to expectation, a reducing NASA-TLX scores would slightly reduce satisfaction (with the standard estimate 0.09) and an increasing PEOL would reduce PU slightly (with the standardized estimate -0.178), which are contrary to H1 and H5. This rather contradictory result may be due to the bias of the subjective rating. As the previous section has mentioned, older adults are found to be biased raters that they tend to give good rating and sometimes inconsistent responses.

Table 20. Regression weights of the recursive structural model

Regression Weights: (Group number 1 - Default model)	Estimate	Standardized Estimates	S.E.	C.R.	P
NASA_TLX <--- USE_PEOL	-2.27	-0.402	0.398	-5.704	***
USE_PEOU <--- USE_PEOL	2.066	0.55	0.223	9.267	***
USE_PEOU <--- NASA_TLX	-0.185	-0.277	0.039	-4.679	***
USE_PU <--- USE_PEOL	-0.576	-0.178	0.193	-2.989	0.003
USE_PU <--- USE_PEOU	0.792	0.92	0.051	15.442	***
USE_Satisfication <--- USE_PEOU	0.204	0.27	0.046	4.475	***
USE_Satisfication <--- USE_PU	0.619	0.703	0.05	12.403	***
USE_Satisfication <--- NASA_TLX	0.046	0.09	0.02	2.318	0.02

The scatter plot below shows linear relationships between perceived usability scores and NASA-TLX scores classified by user group. The above assumption of the older adult subjects driving the unexpected result was supported since a negatively correlative relationship between the perceived usability scores and the NASA-TLX scores only existed in older OAs' data. This could be further explained by the observation notes that some older OAs did not really understand the meaning of workload and they thought the higher the better, so their response was contradictory. The plot has also shown PEOL is a biased measure in this study. The learnability of each UI was considered equally good for many older OAs, so the discriminability of the index is low.

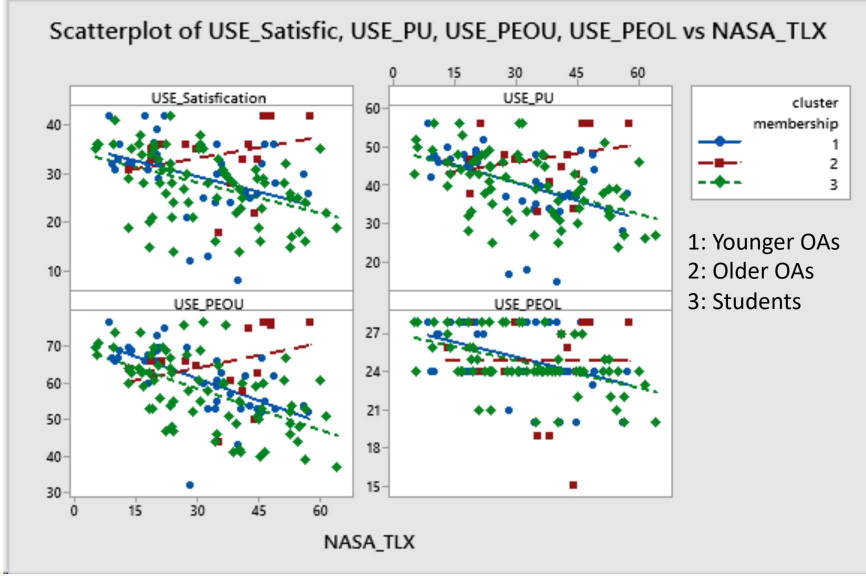


Figure 35. Compare the linear relationships between perceived usability and subjective workload by user group

The Relationship between Human Task Performance and Subjective workload

Human performance and workload are two key human factors constructs. The ultimate goal for them is to estimate cognitive load, which is defined as the use amount of the working memory resource in cognitive psychology. Usability engineering could be based on human factors evaluation because usability could be improved by reducing cognitive loads in theory (Nielsen, 2013).

In this dissertation, human performance and subjective workload measures were used to evaluate the usability of persuasive technology. However, the relationships between these two constructs remains undetermined since the evaluation was conducted for the product persuasiveness, which appears to be a new context of usability. Previous works also provides inconclusive evidence about the relationship between human performance and workload in overall since the related measures are task-specific and scenario-based. For example, Yeh and Wicken (1988) have mentioned that human performance is dissociated with subjective workload measures for most of the multi-tasking driving scenarios. For example, subjective workload measures is insensitive the performance region in overloading region; and performance is dissociated with subjective workload measures when comparing the difficult single task configuration with easy dual task configuration (Yei-Yu Yeh & Wickens, 1988). Whereas Longo (2018) has found subjective workload measure is a predictor of human performance for the information seeking task in a web-based interactive system (Longo, 2018).

In this section, the similar method of determining relationship between perceived usability and subjective workload measure was employed to discover the relationship between human task performance and subjective workload measures. The Pearson correlation analysis was firstly done to roughly discover the relationship between human task performance and subjective workload measures. And then a reflective measurement model of usability of persuasive technology was assumed and verified by confirmatory factors analysis in the end.

In Table 21 is the correlation matrix of human task performance and subjective workload measures. It follows the expectation that d-prime, accuracy, and the response criterion, c, are significantly correlated; d-prime is highly positively correlated with accuracy; accuracy is moderately and negatively correlated to c; and c is weakly negatively correlated to d-prime. The relationship between NASA-TLX and human task performance measures of time of response, d-prime, accuracy are significantly weak correlated. NASA-TLX and time of response are weakly positively correlated, and there are no practically significant correlations between NASA-TLX with d-prime, accuracy, and response criterion.

Table 21. Correlation Matrix of Human Performance and Workload Measure

Correlations	Time of Response	NASA_TLX	d-prime	Accuracy
NASA_TLX	0.172 **	-		
d-prime	-0.039	-0.126*	-	
Accuracy	-0.029	-0.134*	0.868***	-
c	-0.031	-0.051	-0.194**	-0.504***
* p <0.05 ; ** p<0.01; *** p<0.000				

NASA_TLX stands for NASA Task Load Index (The subjective workload measure in this study.); c stands for response criterion; d-prime and c are proposed Signal Detection theory measures for evaluating persuasiveness.

And then, a reflective measurement model was assumed as in Figure 36, I assumed the constructs of effectiveness and efficiency which are caused by the latent factors, usability, could be measured by proposed human task performance and subjective workload measures. The structural equation modeling analysis was performed to estimate structural coefficients between the measures and the constructs. Table 23 shows the model fits summary, which did not reject that model fit the data since the probability value of chi-square test is 0.248. The p-value is higher than 5% significance level and suggests non-rejection of the null hypothesis. Furthermore, the RMSEA for this model is 0.038, which is below the recommended threshold of 0.06; the Tucker-Lewis Index (TLI) value is 0.994, which is higher than 0.95; and the Comparative Fit Index (CFI) value is 0.997, which is higher than 0.9 based on the Hu and Bentler’s guidelines.

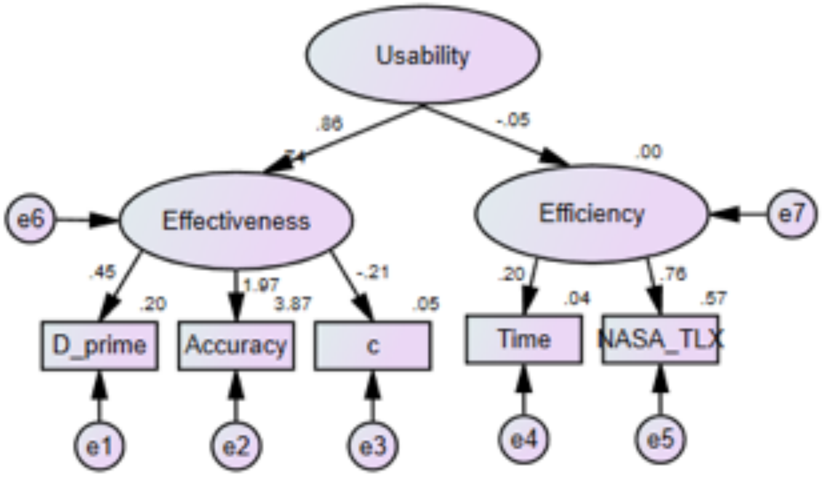


Figure 36. The reflective measurement model of usability based on human performance and workload measures

Table 22. Notes of the reflective measurement model

Notes for Model (Default model)	
Computation of degrees of freedom (Default model)	
Number of distinct sample moments:	15
Number of distinct parameters to be estimated:	11
Degrees of freedom (15 - 11):	4
Result (Default model)	
Minimum was achieved	
Chi-square = 5.408	
Degrees of freedom = 4	
Probability level = .248	

Table 23. Model fits summary of the reflective measurement model

Model Fit Summary				
CMIN				
Model	NPAR	CMIN	DF	P
Default model	11	5.408	4	0.248
Saturated model	15	0	0	
Independence model	5	######	10	0
Baseline Comparisons				
Model	NFI	RFI	IFI	TLI
	Delta1	rho1	Delta2	rho2
Default model	0.99	0.976	0.997	0.994
Saturated model	1		1	
Independence model	0	0	0	0
RMSEA				
Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	0.038	0	0.11	0.519
Independence model	0.471	0.438	0.505	0

Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

Table 24. Regression weights of the reflective measurement model

Regression Weights: (Group number 1 - Default model)	Estimate	Standardized Estimate	S.E.	C.R.	P	Label
Effectiveness <--- Usability	1	0.862				
Efficiency <--- Usability	-1	-0.045				
c <--- Effectiveness	-0.054	-0.214	0.009	-6.208	***	par_1
Accuracy <--- Effectiveness	0.749	1.967	0.399	1.877	0.061	par_2
D_prime <--- Effectiveness	1	0.447				
NASA_TLX <--- Efficiency	1	0.758				
Time <--- Efficiency	0.627	0.203	1.107	0.566	0.571	par_3

Confirmatory factors analysis was performed to confirm the relationship between two constructs as shown in Figure 37. The model fits summary (Table 26) suggests adequate model (p-value 0.242). The relative fits index also supports the model fitness with the RMSEA value as 0.039, Tucker-Lewis Index (TLI) value is 0.993, and Comparative Fit Index (CFI) value is 0.997, all conforming to Hu and Bentler’s guidelines. However, the table of regression weights (Table 27) shows weak correlations between constructs, and the coefficients are insignificant. A possible reason is that the correlation between NASA-TLX and time of response is very weak while d-prime, accuracy, and c are derived from the same source, so they are confounded rather than correlated.

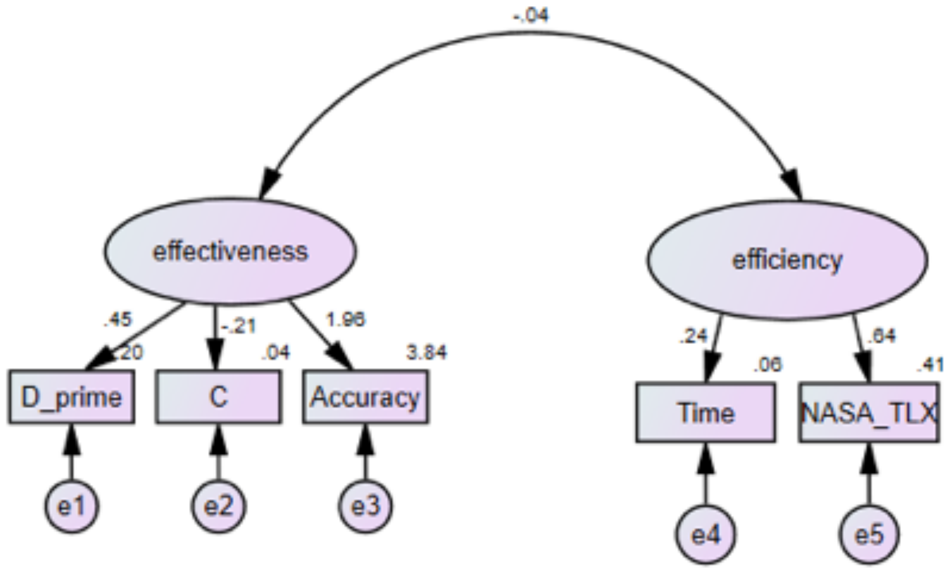


Figure 37. Confirmatory Factor Analysis (CFA)

Table 25. Notes of the CFA

Notes for Model (Default model)	
Computation of degrees of freedom (Default model)	
Number of distinct sample moments:	20
Number of distinct parameters to be estimated:	16
Degrees of freedom (20 - 16):	4
Result (Default model)	
Minimum was achieved	
Chi-square = 5.476	
Degrees of freedom = 4	
Probability level = .242	

Table 26. Model fits summary for the CFA

Model Fit Summary					
CMIN					
Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	16	5.476	4	0.242	1.369
Saturated model	20	0	0		
Independence model	10	528.2	10	0	52.82
Baseline Comparisons					
Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	0.99	0.974	0.997	0.993	0.997
Saturated model	1		1		1
Independence model	0	0	0	0	0
RMSEA					
Model	RMSEA	LO 90	HI 90	PCLOSE	
Default model	0.039	0	0.11	0.512	
Independence model	0.46	0.427	0.494	0	

Hu and Bentler (1999) recommended guidelines of RMSEA values below 0.06, TLI values higher than 0.95, and CFI values higher than 0.9.

Table 27. Regression weights of the CFA

Regression Weights: (Group number 1 - Default model)	Estimate	Standardized Estimate	S.E.	C.R.	P	Label
NASA_TLX <--- efficiency	1	0.642				
Time <--- efficiency	0.874	0.239	1.266	0.691	0.49	par_1
D_prime <--- effectiveness	1.335	0.446	0.728	1.834	0.067	par_2
c <--- effectiveness	-0.071	-0.212	0.046	-1.546	0.122	par_3
Accuracy_adj <--- effectiveness	1	1.961				
Covariances: (Group number 1 - Default model)	Covariance Estimate	Corrrelatioin Estimate	S.E.	C.R.	P	Label
effectiveness <--> efficiency	-0.273	-0.044	0.157	-1.74	0.082	par_4

8.2 Design Considerations of Persuasive mHealth apps for Older Adults

8.2.1 Ageing-Centered Design Guidelines of Persuasive mHealth apps

In this section, the empirical learnings from conducting the ageing-centered design for persuasive mHealth app are summarized. The first part describes the considerations when conducting user research with older adult participants. The second part summarizes from the study findings to list the important design recommendations for the similar kind of persuasive mHealth apps. However, these guidelines must be adapted with caution for the future works as they are based primarily on qualitative data. Few pieces of supported quantitative evidence were found in this study since the older adults (especially older OAs) represent a relatively small sample.

Considerations for User Research with Older Adult Participants

The findings in this study point to the following considerations when conducting an ageing-centered design project.

It is important to consider OA's needs via user-centered design framework.

This work found the proposed extended usability engineering framework helped determine the critical design elements to increase the perceived usability, which implies the potential technology acceptance of the dietary management app for older adults. The mixed-methods results have also confirmed Mitzner et al.'s (210) and Wang et al.'s (2019) finding that older adults are actually willing to use a new technology as long as they find it useful (Mitzner et al., 2010; S. Wang et al., 2019). Most older adults have a relatively lower threshold of giving higher perceived usability scores compared to younger users. A possible explanation might be that, as the technology solves their problems or adds values to their daily life scenarios, they are open to learn and willing to give more tolerance to the "minor but seems adjustable" usability problems.

Older OAs are "good participants", who tend to give "good" but biased subjective ratings due to social desirability.

The study results have shown, most older adults are biased raters who tend to give optimistic subjective rating. The results could be explained by age-related social desirability (see the discussion in section 6.3). This fact could become a threat of external validity to deliver the accurate user testing results. In the context of overly generous raters, mixed-methods user testing would be recommended to moderate the biased subjective user testing data with objective measures. The frequently-used human factors evaluation methods including human task performance measures, biometrics (heart rate, eye movements, EEG,…etc.) are examples of potentially useful objective measures (Hornbæk, 2006; Lin & Imamiya, 2006; Qu, Zhang, Chao, & Duffy, 2017; Wenzel et al., 2015). In this work, I utilize signal detection to accommodate the biased raters and imbalanced dataset in the decision-making scenario. (The in-depth discussion about the use of signal detection theory is continued in section 8.2.1).

Another way to cope with Social Desirability Bias (SDB) is either measuring the correlation between social desirability scale with the self-report measures (Paulhus, 1991) or considering mediating effects of the social desirability in the causal relationship (Nolte, Elsworth, & Osborne, 2013; Soubelet & Salthouse, 2011). Currently, Marlowe-Crowne Social Desirability Scale (MC-SDS, or MC scale) is the most widely used instrument to assess the social desirability of an individual and determine contamination of the SDB for self-report measures (Paulhus, 1991). It is a 33-item instrument in a true-false response format to measure personal traits of seeking for social approval. The reliability and validity of the instrument are good, Crowne and Marlowe (1960) reported a pretty high test-retest correlation (0.88), internal consistency coefficient (0.88), and significant correlation with the 1953 developed Edwards Social Desirability Scale (Crowne & Marlowe, 1960). MC-SDS was translated into different languages in 1970s and many short versions were developed such as Strahan and Gerbasi's 10-item version and Reynolds' 13-items version (Reynolds, 1982; Strahan & Gerbasi, 1972).

However, some researchers still question the representativeness of the social desirability scale to response bias since the personal traits may be indirect indicators of the response style (McCrae & Costa, 1983; Uziel, 2010). In recent years, some objective measures based on biometrics including eye tracking or EEG were proposed (Baumgartl, Sauter, Roessler, & Buettner, 2020; Kaminska & Foulsham, 2013). The related discussion points out a direction of future research to incorporate SDB measures to the research studies of ageing-centered design.

Younger OAs may have a similar mental model to young adults, and they have nearly the same performance as young people.

The study confirms that the older adult is a special demographic of technology users since the individual difference within this chronological ageing group is huge. Two healthy older adults in the same age may still have heterogeneous attitude, behavior, and mental model toward using technology (Czaja, Boot, Charness, & Rogers, 2017; Hänninen, Taipale, & Luostari, 2020). A critical finding from the study is that the healthy ageing group of technology users could be further classified by age and knowledge level. The behavior and performance of younger OAs is more like young adults, such as the student group. The finding could be explained by the demographic trend that the first generation of technology users is ageing and so the technology gap is gradually diminishing.

An implication of the finding is the possibility that the necessity of considering the technology gap will be reduced in future ageing-centered design projects. Another implication is that the student group may represent a sufficient proxy to the younger OA group in user testing studies. This is beneficial to researchers since the student group is relatively homogenous, easily accessible, and fairer on the subjective rating. However, with a small sample size, caution must be applied, as the findings might not be generalizable to other situations.

Reducing the workload of using the apps could improve the perceived usability by increasing the perceived ease of use and satisfaction.

Correlation analysis has shown that NASA-TLX scores are moderately and negatively correlated with PU, PEOU, PEOL, and Satisfaction with the Pearson coefficients between -0.3 to -0.5. Stepwise regression has also confirmed NASA-TLX is a significant predictor of PEOU, Satisfaction, and USE total scores. It is reasonable to assume a causal relationship is present between workload and perceptions of PEOU, Satisfaction, and USE total scores (although this study was not designed to validate this). This relationship would lead the designer to reasonably expect to improve perceived usability by making an effort to reduce task loading, such as simplifying the task flow, avoiding tedious and repetitive task, ...etc.

However, this finding seems contrary to that of Longo (2018) who found subjective workload and perceived usability are significantly uncorrelated, which should be treated as independent constructs of UX (Longo, 2018). This inconsistency may be due to two reasons: First, different measures of perceived usability were used in Longo's and my studies. In Longo's study, Subjective Usability Scale (SUS) was used instead of Lund's USE questionnaire. The former is shorter, and it's a mix of questions from the different constructs on the USE questionnaire. Although Gao et al. (2018) has verified both questionnaires are reliable and highly correlated (M. Gao, Kortum, & Oswald, 2018), the confounded aspects on SUS may make the relationship between workload and usability unexplainable. Functionally this could occur if workload is simply related to a specific construct of usability which is diluted by inclusion of other irrelated questions; Secondly, Longo's conclusion was based on 9 different types of information seeking tasks with different levels of task complexity, difficulty, and interestingness. However, he didn't balance the numbers of each of the task genres, which would be a potential mediator of the relationship between workload and usability. The section 0 gave the in-depth discussion about the relationship between perceived usability, subjective workload, and objective human performance.

Listing all the information in detail may make some OAs feel at ease and be more confident.

According to the observation notes and the OA participants' feedback, most older adults were eager to learn the technology from all perspectives, and many of them felt at ease when they could see the full detail of information presented. During the experiment, instructional material (including task flow chart, standard operation procedures, user manual) were equally helpful to the practice and the guidance of the participants. Older adults tended to pay more attention to the written documents with patience compared to the young people. These findings broadly support the work of other studies in this area that instruction material is needed in designing mobile user interfaces for older adults. However, since the instructional material design is not a main concern in this study, several questions remain unanswered at present. A further study with more focus on the instruction effect on older adult is therefore suggested.

Many older adults hold the same attitude of thorough treatment of information for nutrition information as well (i.e. preferred detail information presented on the UI). Some of them has expressed their preference for the FDA nutrition facts panel during the interview, while also agreeing with the better usability of the FSA nutri-scores label; and some of them changed their mind and found the FSA Nutri-scores label as trustable once they learned it is from a certified authority. Another small group of older adults suggested to have both FDA and FSA labels in the final design.

Empirical Recommendations of Persuasive Design for Older Adults

Several themes have been developed through the study, which are further summarized as the following recommendations of persuasive design for older adults.

1. Comparison task would be easily done by displaying the alternatives horizontally.
   A theme has come up from interview data that older adults may prefer the horizontal presentation for comparison since it's similar to the natural reading behavior. However, the significance is not found in quantitative data due to the individual differences of adopting different searching strategies in the searching context.

2. Nudge could make a greater impact on decision making for older OAs compared to younger adults. The human behavior data (observation notes, and post-scenario interview questions) in this study suggests that older OAs were more likely to rely on nudges in any form. Statistical significance was not found in the quantitative data to verify this observation, but every tested nudge approach slightly increased the discriminability and accuracy for older OAs.

3. Transparency of nudge approaches should be considered for younger OAs.
   For younger OAs and students, a significant negative effect of default nudge has been shown on persuasiveness metrics. The results show that transparent nudge approaches may not be effective to influence younger OAs (and students) who have higher computer proficiency and health literacy. In addition, when a non-transparent approach is found by users, it would cause negative effects on persuasiveness.

4. Both younger OAs or older OAs tend to use a liberal criterion when deciding healthy food to eat, but younger OAs could be more stubborn to change until a sophisticated nudge

approach is used. This finding suggests older adults may have larger bias to decide healthy food, they tend to suppose all the given options are accepted healthy food. Combining with the above two findings, it is reasonable to further assume a sophisticated nudge approach, such as digital nudge, is needed for younger OAs since they would be more rejective of or simply ignore more transparent forms of nudge.

8.2.2 Implication of Digital Nudge Design

The essential feature of digital nudge is to design the choice architecture by manipulating the UI design element, and thus it has a strong potential to influence the user's behavior. However, when it comes to the idea "design the digital nudge in the system". The designers and developers should keep in mind that, 1. The designer is responsible to the appropriate usage of nudge approach to avoid the adverse effect. They should also understand the intention of system design and deliver the appropriate means based on human factors evaluations of the system. 2. Not only "what approach" matters, but also "when" to nudge, "who" is the nudger, and "to whom" matters. In other words, a successful digital nudge design is strongly based on the context as Thaler and Sunstein defined and supported by our study results.

In real world practice, since nudge theory roots from the cognitive psychology, which is based on the understanding of human nature, it requires an iterative design thinking process and systematic human factors approaches to evaluate the design. Human factors methods, such as cognitive work analysis, could be used in this iterative design process to identify the critical human information processing stages and decide the best timing of the intervention to maximize the nudge effect.
In this article, we proposed a human-centered digital nudge design framework to identify the key human factors principles and methods could be used in the regular user-centered design process:

Human-centered digital nude design framework

Digital nudge design could be embedded in the regular User-Centered Design (UCD) process with putting more emphasis on the application of Human Factors Engineering (HF/E) methods to design digital nudge based on human's nature.

In the research and analysis stage, traditional UCD research methods focused on presenting the persona and the use scenario to inspire designers thinking about the user's need and solutions. However, digital nudge design put more emphasis on task analysis methods and human error analysis methods, for a thorough understanding of human natures in a specific context for better defining the ways human interacts with the computer.

In the design stage, traditional UCD methods rely on UI designer's aesthetics, knowledge and previous experience about system usability to create the prototype. However, when considering digital nudge design, the UI designer should keep in mind Hansen and Jespersen's framework if adapting from other nudge approaches. Hansen and Jespersen's framework is based on dual process theory and considering the transparency of the approach for the nudge ethics.
If digital nudge usage is justified in the system, for example, for the individual's health behavior change. Designers could rely on the task analysis and human error analysis results, human factors design principles, and theories roots from cognitive psychology to come up a better idea for digital nudge design. For example, in our study, the design idea of adapting 2AFC testing paradigm as an UI element actually roots from experimental psychology. It has been found human would have a better performance on signal/noise discrimination task in a 2AFC testing paradigm (Macmillan & Creelman, 2005).

Another theoretical basis to inspire digital nudge design could be game theory. It would be extremely useful if designer is able to perform the equilibrium analysis for considering agents' equilibrium reaction to design the incentive architecture (the payoff function) and guide the user's choice vice versa (Spiegler, 2015). The game theoretical analysis, which is based on the observed stimulus-response patterns and the theoretical Nash equilibrium, which would ensure the consequence of digital nudge design has been analyzed in an economical viewpoint. For example, Spiegler (2015) has done equilibrium analysis of the firm price competition game assuming with the default nudge of the auto-renewal of the product subscription in the consumer market. He found banning the auto-renewal would maximize consumer's welfare with forcing the firms' competition game reaches the Nash equilibrium with the lowest price (Spiegler, 2015). Nash equilibrium is a proposed solution of the non-cooperative zero-sum game with two or multiple players in game theory. It is a strategy profile when there's no player can do better by unilaterally changing his or her strategy (von Neumann & Morgenstern, 2007). Of course, when we simply look at human-computer interactions, it is never a zero-sum game. However, in a complex system with multiple human agents and smart agents involved, game theoretical analysis could be used to analyze the digital nudge effect and suggest the UI design of a smart agent. For example, Liang and Yan (2019) have analyzed the crowdsourcing contest system with game theoretical analysis and found the existing issues which suggest the algorithm design in the system (Liang & Yan, 2019).

In the evaluation and implementation stage, traditional UCD focused on collecting user's feedbacks about usability and use experience for the iterative design. Customer satisfaction, intention to use, and user retention rate are the key performance index for kicking up the next design cycle. But the evaluation of digital nudge design would focus on human performance measures and continually collect those metrics as the web analytics to track the long-term user behavior and evaluate the effectiveness of the design. And for a complex system, as previous paragraph has mentioned, the game theoretical analysis would be another useful evaluation method to help designer foresee the theoretical consequence of digital nudge intervention in the economics viewpoint.

Table 28 summarized the suggested HF/E methods used in the UCD process for digital nudge design.

Table 28. Human-Centered Digital Nudge Design Framework

UCD Process stage	UCD methods	HF/E emphasis for digital nudge design

Research	Interview;
	Ethnographic Research Methods;
		Contextual Inquiry;
		Observation

Analysis	Affinity Diagram;
	Persona;
	Use Case Analysis;
		Task analysis / Cognitive Work Analysis;
		Human error analysis

Design	Sketch; Low/High Fidelity Prototypes	Dual Process Theory / Hansen and Jespersen's typology; Game Theory

Evaluation	Interview;
	Usability / UX / Customer Satisfaction questionnaire;
		Primary and secondary task performance measures;
		Decision Analysis

Implementation	Customer feedbacks / Satisfaction questionnaire	Web analytics of human performance

Appropriate use of digital nudge.

To avoid the misuse of nudge techniques and approaches in public policy making, Hansen and Jespersen (2013) have proposed the framework for responsible nudge approaches usage in public policy. Based on Hansen and Jespersen's (2013) framework, there are two types of nudge approaches designed with utilizing different thinking modes involved different cognitive processing stages in dual processing theory, automatic mind v.s. reflective mind. Type 1 nudge approaches utilized the human's automatic mind, in which people made unconscious decision to make a quick response to the environment; type 2 nudge approaches utilized the reflective mind in which people made conscious deliberation. And to make sure the ethics of nudge in response to those critics who is against to "psychological manipulation" and concerns nudge as simply "underhanded deceptions". Hansen and Jespersen proposed to evaluate the manipulation by the "epistemic transparency", which distinguishes the transparency of the nudge approach by the visibility of the intention and means to the agent being nudge. In this context, the thinking modes and the transparency define four categories. Hansen and Jespersen claim only those non-transparent methods were the so-called "psychological manipulation". They labeled the type 1 non-transparent approaches as "behavior manipulation" and the type 2 transparent approaches as “choice manipulation” , and warn the possible controversy of using non-transparent methods in public policy making. They suggest the government (the nudger) and the choice architect (designer) should take full responsibility of the non-transparent methods, including the disclosure of intentions and the effects and side-effect for the “behavior manipulation “methods and limit the usage of “choice manipulation” methods unless the design keep some extents of free choice for the agents being nudge. They encouraged the use of transparent approaches in public policy making.

Nudge approaches usage in a digital world could also inherit the same framework, however, one should notice that Hansen and Jespersen’s framework is proposed in the context of policy making to ensure the “publicity principle” is obeyed. The principle regulates the government to adopt a defendable policy in public for avoiding the government overpowered in policy making. And thus, the intents and the means should be transparent to the agents being nudged and the “choice manipulation” is strictly prohibited since it may derive the citizen’s autonomy of choice. It still requires some revisions of the responsible nudge usage guidelines to adapt the framework to other human computer interactions context. For example, in the context of persuasive technology usage for individual’s health behavior change, the intention of the digital nudge is clear to the consumer, and the free market always allow the consumer to choose not to use. It’s unnecessary to avoid all the possibilities of “psychological manipulation” which would also limit the designer’s thinking. Instead, we encourage the designer consider digital nudge design which is successful at effectively and implicitly guide user’s behavior without instigating any negative effects. Although in some viewpoints, digital nudge may be considered as “choice manipulation” but at least, in our case, it could be justified with the clear and beneficial-to-users intention and the careful examined means based on our proposed human-centered digital nudge design framework.

However, it could still be controversy since the absence of the specific legitimate guidelines to ensure the ethics of digital nudge in the human-computer interactions. We highlighted the research gap and the keep the room for further debate since the proposal of legitimate guidelines is beyond this article’s discussion.

8.3 Conclusion

In this dissertation, an extended usability engineering framework for persuasive mHealth apps, which integrates Nielsen's framework, Oinas-Kukkonen & Harjumaa's persuasive design framework and human factors evaluation method for persuasiveness based on Signal Detection Theory (SDT) was proposed. A dietary management app which aims to influence the dietary decisions of older adult patients for health behavior change was developed to validate the proposed framework. A mixed-methods user testing study was conducted with 40 subjects including twenty older adults and twenty students to investigate the nudge impacts of the proposed UI design elements on perceived usability, subjective workload, and the human performance of selecting healthy food.

There were two parts to the study; the first part tested usability of four different versions of a user interface. These made up the treatments of a 2 x 2 full factorial design (search results layout x nutrition information format). The second part of the study was a human factors evaluation of a 2 x 2 x 2 digital nudge design (decision paradigm x nutrition information format x system default) based on the human performance in a healthy food decision-making experiment.

The study results have shown that the choice-based (Two Alternative Forced Choice; 2AFC) layout significantly increase the d-prime and accuracy which implies the persuasiveness; while the system default pre-selection decreased the persuasiveness; and the interpretative FSA Nutri-scores label saved time of response, reduced workload, and increased perceived ease of use, perceived ease of learning, and satisfaction. In this study, the Older Adult (OA) participants could be further classified as older OAs and younger OAs by age, computer proficiency, and health literacy. There is no significant difference between older adults and students to effectively make healthy food choices. But there are individual differences of perceived usability, subjective workload, and efficiency of making decisions on different UIs. The younger OAs (aged 63 in average) with higher computer proficiency scores and health literacy scores, perform and behave nearly the same as students. The older OAs (aged 69 in average) with lower computer proficiency scores and health literacy scores, perform significantly worse and are biased raters. The ageing-centered design guidelines for persuasive mHealth app were further discussed and a generalized human-centric digital nudge design framework was proposed. In the end of the article, the measuring science of persuasiveness was discussed; the measurement models and a structural model of usability and were proposed and verified by structural equation modeling analysis for the similar research and design of persuasive technology in the future.

8.4 Final Remarks

8.4.1 Study Limitations

There are several study limitations due to the constraints on the research design and the method. In order to better accommodate the needs and the special requirements of older adult participants, several constraints have been set and they could be threats of internal validity. For example, data collection was partially done outside the lab on the laptop without an eye tracker for the community-dwelling older adults who may not be able to travel to the eye tracking lab on campus. The noise from the natural environment could harm the internal validity. And due to the same reason, limited eye movement data could be collected for the further quantitative analysis and thus the eye tracking measures are not selected as the response.

There are also several threats of the external validity of the study. Due to the constraint of the project scope, limited design elements focused on nudge effect have been considered in this research project. However, there are more unselected design elements such as button and font size, display contrast, background color may cause potential usability problems of mobile user interface for older adults (Kurniawan & Zaphiris, 2005). Nurgalieva et al. (2019) did systematic literature review between 2005 to 2017 on 52 research articles and found 434 design guidelines has been proposed for the design elements of display, navigation, context, forms, …etc.(Nurgalieva, Jara Laconich, Baez, Casati, & Marchese, 2019). To avoid those design elements becoming the noise of nudge design elements, the simplified experimental user interfaces of the prototype web application were used; and the experiment was done on the computer platform rather than on the mobile phone platform. Other reason of conducting the experiment on the computer is to shorten the learning curve of experimental task operations on the smart phone. And it's also partially due to the limitation of the author's programming ability to develop a similar high-fidelity prototype web application on the smart phone.

The short time span of the user testing for finishing this project in product development perspective could also be a major concern of the external validity. Health behavior change usually requires longer-term observations and field studies for the evaluation of the health outcomes improvement.

Another threat for both internal validity and the external validity is the smaller sample size due to the Covid-19 pandemic in 2020. Data collection was terminated earlier since the prohibition of in-person study was posted by IRB.

8.4.2 Directions for Future Research

Verification of Human-Centered Digital Nudge Design Framework

The current design project is still in an earlier stage of product development, the longer time span user testing study to evaluate the impacts on the health behavior change and improvement of health outcomes haven't been considered in the project scope. In the future, once the proposed mHealth app is developed as an online commercial product, a longer-term observation study of assessing health behavior change and health outcomes improvements would be needed to better determine the effectiveness of adapting a persuasive mHealth app.

With the continuous usage of the commercial app in an individual's daily life, every decisions s/he has made in app would be recorded. So the daily eating behavior is monitored by continuous data collection of daily food intakes.

The long-term effects of health behavior change could be assessed by health outcomes and complications evaluation for type II diabetes. The direct measures from the clinical research study of including glucose levels (glycemic control of HbA_{1C}), micro-albumin test (for proteinuria), blood pressure monitoring, eye and lower extremities examination (for retinopathy and foot ulcers), and lipid profile (for dyslipidaemia) (Gavin, Stolar, Freeman, & Spellman, 2010; Harris, 2000; Reddy, 2000). Other indirect measures including the healthcare resources engagement, medication involvement, quality of life years (QALYs) and healthy year equivalents (HYEs) (Reddy, 2000). The aggregated big data of food tracking and regular examination of health outcomes could further be used to train the smart system with assessing and predicting health outcomes from the tendency of patient’s behavior and lifestyle change. The predictions could be used to adjust the food recommendations and sent to the users as a reminder to adjust their behavior.

In addition, an empirical human-centered digital nudge design framework has been proposed in the end of this article. A suggestion of the future research direction is to implement other persuasive technology design project based on the same framework to ensure the reliability and the validity of the proposed methodologies.

Social Assistive Robotics (SAR)

Social Assistive Robotics (SAR) is another promising persuasive technology to support elderly care (Abdi, Al-Hindawi, Ng, & Vizcaychipi, 2018). Providing assistance via social interactions on behalf of physical support is the focus of SAR applications, for example, exercise coach (Fasola & Mataric, 2013). Rossi et al. (2018) did two user testing studies to compare the acceptance rate of the recommendations from the SAR app and the mHealth apps. They found users preferred SAR but there’s no significant difference between the acceptance rate of the recommendations (Rossi, Staffa, & Tamburro, 2018).

Social Assistive Robotics (SAR) is a pretty new research area. The earliest SAR may be the Nursebot (also called Pearl), which was developed by Carneige Mellon University in 2003. It integrated the auto reminder system on a humanoid mobile robot with a touch screen, speech recognition & synthesis, and indoor navigation function (Pineau, Montemerlo, Pollack, Roy, & Thrun, 2003).

The Care-O-bot is a mobile robot with manipulator arms, which simulate the human’s ball and socket joints and thus it could perform more dedicate arms and hands movements such as pick and place simple objects in home environment (Graf, Hans, & Schraft, 2004). The 4th generation of the Care-O-bot debuted in 2015 with the modular system design feature, to accommodate the wider ranges of applications including home environment, healthcare institutes,....etc. (Ackerman, 2015).

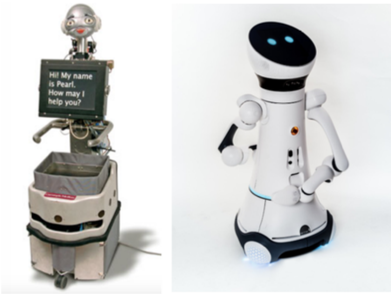


Figure 38. Left: Neurobot Pearl; Right: Care-O-Bot 4
(Source: Left- Carnegie Mellon University; Right-Fraunhofer IPA)

The PARO is a seal-like companion robot, which could move its head and legs, making a sound imitating baby harp seal, and respond to touch, voice, sight, temperature. It is designed for elderly people with dementia, and originated in Japan and then further developed in the Europe beginning in 2003. The commercial version is now in the 8th generation. It may be the most successful social robot. More and more Human Robots Interactions (HRI) studies have found that interacting with PARO boosts the moods for elderly with dementia, and it helps release the heavy burdens of the caregivers of people with dementia, and it helps release some of the heavy burden on the caregivers of people with dementia (Shibata, 2004; Wada, Shibata, Saito, & Tanie, 2002).

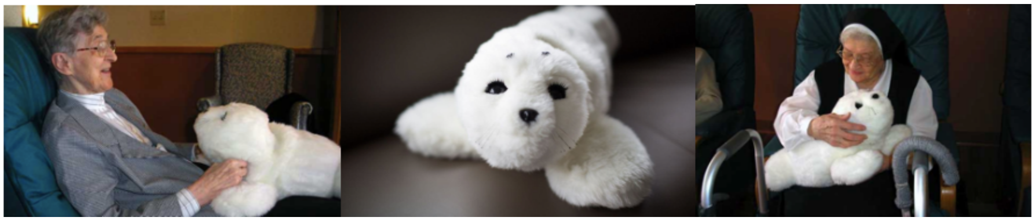


Figure 39. The Commercial Seal-Like Companion Robot PARO;
(Source: PARO US Inc.)

Usability Engineering for a Complex Smart System

As the healthcare service system is getting smarter to provide more complex functionalities, there are more interaction design elements should be taken into consideration at once and data-driven design decisions should be made systematically.

Rossi et al. (2018) did two user testing studies to compare the perceived usability and the acceptance rate of the recommendations from the SAR app and the mHealth apps. They found users preferred SAR but there’s no significant difference between the acceptance rate of the recommendations (Rossi et al., 2018). This study implies that SAR may be a better platform to implement the proposed digital nudge design framework. However, in the context human-SAR interactions, more human aspects should be considered beyond the human information processing, for example, human emotions.

In this context, research protocols of usability engineering for a complex SAR apps should be proposed and practiced. For example, automated usability measures, such as web analytics, eye tracking,..etc., would be needed to collect larger sample size data effectively and efficiently; and quantitative methods should be employed for data analysis in a system engineering viewpoint. In the future, once the current design project gets into the later product development stage, the research project scope should be broadened, and higher external validity of the studies will be needed. The further research work should be done to facilitate usability engineering for a complex human-centered smart service system.

Food Decision Support System

Human decision making of food intakes is actually a complex multi-objectives problem with many constraints of the cost, availability, and psychological factors such as culture, personal belief, health conditions…etc. Most people solve the problem by practicing the satisficing rule to find a suboptimal solution and thus there's a chance to nudge an individual's health behavior.

This behavioral decision-making aspect impacts policy making of providing nutrition information on packaged food. Currently, the major public health systems around the world including US Food and Drug Administration (FDA) and European Union (EU) regulate the use of a text-list based Back-of-Package (BOP) nutrition label. However, this kind of label actually assume the users to be rational decision makers with adequate levels of health literacy (measured by Weiss' Newest Vital Sign)(Vincent Delhomme, n.d.). As the French government developed and regulated the use of the interpretative label, FSA Nutri-Scores, more and more European countries have followed. And thus, the most recent EU official document "Farm to Fork Strategy" has also suggested the direction of making a new policy in 2022 to regulate the use of interpretative FOP label (European Commission, 2020).

This thesis contributes the evidence that providing FSA-Nutri-Scores label would significantly improve human decision efficiency. Although no significant effect was found directly of nudging the one-shot human decisions in the lab, however it's still possible to influence human's long-term behavior since human may tend to adopt an easy option under the time pressure in the real-life scenario. What's more, the qualitative data in this thesis further suggests the research direction of cultural differences in interpreting FSA-Nutri-Scores data, which could be a major concern in the legitimate framework of EU policy and the adoption of the similar policy in the US.

However, currently, there's still a gap of developing such a smart system to deal with so many latent factors as a service engineering nature and the uncertainty in the real world. There's a need of the cooperation between service process design and also the fundamental Artificial Intelligence (AI) research to develop the algorithm and continually improve the system performance vice versa. In the future, once the smart system development is getting mature, a more complete service design of healthy food decision support could be discussed and provided.

APPENDIX A. PRELIMINARY QUESTIONNAIRE

Section 1: Demographics & Dietary Behavior

1.) Gender: □Male □Female

2.) Age:

3.) Are you comfortable with planning a meal by yourself? □Yes □No

4.) Are you currently following a special diet (i.e., vegan, vegetarian, diabetic, low fat, lactose free, gluten free, kosher, halhal)?

□Yes □No If yes, what kind?

Has this diet been prescribed by your health care provider? □Yes □No

5.) Have you ever had a bad reaction or aversion to any foods?

□Yes □No If yes, which foods?

6.) Which criterion is more important to you, when selecting food?

□ Tastiness. □ Healthiness.

How much more important for your selection compare to the other?

1: equally important	3: moderately important	5: strongly important	7: dominantly important	9: extremely important
□	□	□	□	□

7.) Which type of information do you rely on the most to make the selection?

□Brand Name □Image □Nutrition Information □Ingredients □My Previous Experience

8.) If you have selected "nutrition information", which nutrient do you rely on the most to make the selection?

□rating □color coding □energy per 100g □fat per 100g

□saturated fat per 100g □sugar per 100g □salt per 100g □Other

Section 2: Computer Proficiency Questionnaire (CPQ)

1. Computer Basics

I can:	1: Never Tried	2: Not at All	3: Not Very Easily	4: Somewhat Easily	5: Very Easily
a. Use a mobile phone keyboard to type.	□	□	□	□	□
b. Adjust the volume of my phone.	□	□	□	□	□

2. Communication

I can:	1: Never Tried	2: Not at All	3: Not Very Easily	4: Somewhat Easily	5: Very Easily
a. Open and read an email on my phone.	□	□	□	□	□
b. Send an instant message (by Facebook messenger, iMessage, Skype, Line messenger…etc.) by my phone.	□	□	□	□	□

3. Internet

I can:	1: Never Tried	2: Not at All	3: Not Very Easily	4: Somewhat Easily	5: Very Easily
a. Use search engines (e.g. Google, Bing, Yahoo…etc.) to find information about local community resources on the Internet.	□	□	□	□	□
b. Find information about my hobbies and interests on the Internet.	□	□	□	□	□

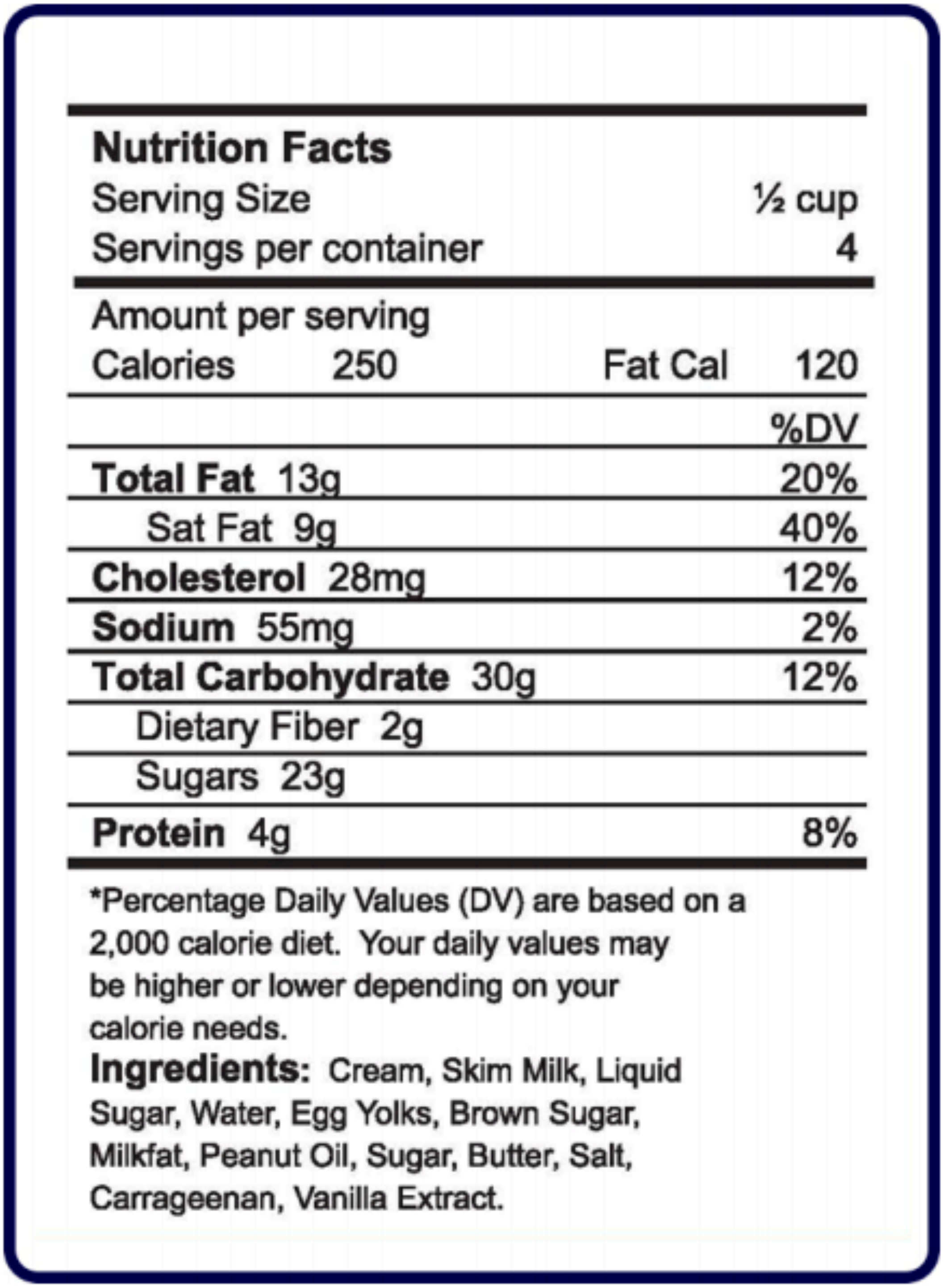
4. Calendar

I can:	1: Never Tried	2: Not at All	3: Not Very Easily	4: Somewhat Easily	5: Very Easily
a. Use my phone to enter events and appointments into a calendar.	□	□	□	□	□
b. Check the date and time of upcoming and prior appointments.	□	□	□	□	□

5. Entertainment

I can:	1: Never Tried	2: Not at All	3: Not Very Easily	4: Somewhat Easily	5: Very Easily
a. Use my phone to watch movies and videos.	□	□	□	□	□
b. Use my phone to listen to music.	□	□	□	□	□

Section 3 Health Literacy



Nutrition Facts

Serving Size			½ cup
Servings per container			4
Amount per serving			
Calories	250	Fat Cal	120
			%DV
Total Fat 13g			20%
Sat Fat 9g			40%
Cholesterol 28mg			12%
Sodium 55mg			2%
Total Carbohydrate 30g			12%
Dietary Fiber 2g			
Sugars 23g			
Protein 4g			8%

*Percentage Daily Values (DV) are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

Ingredients: Cream, Skim Milk, Liquid Sugar, Water, Egg Yolks, Brown Sugar, Milkfat, Peanut Oil, Sugar, Butter, Salt, Carrageenan, Vanilla Extract.

***** This information is on the back of the ice cream container.**

Please answer the questions below based on the nutrition facts label on Page 4:

1. If you eat the entire container, how many calories will you eat?

Answer:

2. If you are allowed to eat 60 grams of carbohydrates as a snack, how much ice cream could you have?

Answer:

3. Your doctor advises you to reduce the amount of saturated fat in your diet. You usually have 42g of saturated fat each day, which includes one serving of ice cream. If you stop eating ice cream, how many grams of saturated fat would you be consuming each day?

Answer:

4. If you usually eat 2500 calories in a day, what percentage of your daily value of calories will you be eating if you eat one serving?

Answer:

***** Pretend that you are allergic to the following substances: Penicillin, peanuts, latex gloves, and bee stings.**

5. Is it safe for you to eat this ice cream?

Answer:

6. If you answer “No” to question 5, why not?

Answer:

APPENDIX B. ANOVA TABLES FOR GLMS OF STUDY PART 1

Output Corresponding to ANOVA Tables for GLMs of Study Part 1 (6.2.1)

Factor Information

Table B.1 Factors Information

Factor	Type	Levels	Values
Searching Results Layout (Choice-based=1)	Fixed	2	0, 1
Nutrition Information Format (FSA-Nutri-scores=1)	Fixed	2	0, 1
Name	Random	43	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43

General Linear Model: USE_PU versus Search UI Layout, Nutrition Information format, Name

Analysis of Variance

Table B.2 ANOVA table for GLM of PU

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Searching Results Layout (Choice-based=1)	1	11.5	11.50	0.27	0.607
Nutrition Information Format (FSA-Nutri-scores=1)	1	37.9	37.93	0.87	0.351
Name	42	9169.1	218.31	5.04	0.000
Choice-based*FSA-Nutri-scores	1	14.9	14.94	0.34	0.558
Error	128	5549.9	43.36		
Lack-of-Fit	124	5256.9	42.39	0.58	0.852
Pure Error	4	293.0	73.25		
Total	173	14835.7			

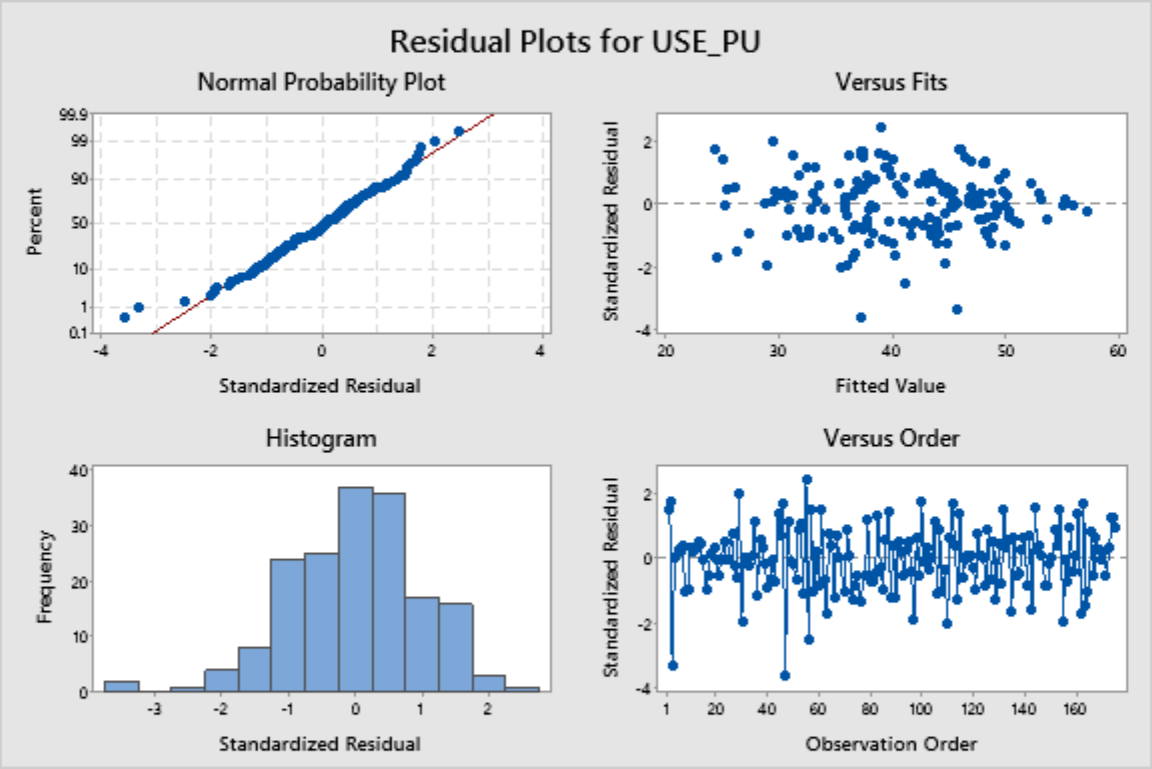


Figure B.1 Standardized Residual Plots for PU

General Linear Model: USE_PEOU versus Search UI Layout, Nutrition Information format, Name

Analysis of Variance

Table B.3 ANOVA table for GLM of PEOU

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Searching Results Layout (Choice-based=1)	1	3.2	3.219	0.05	0.820
Nutrition Information Format (FSA-Nutri-scores=1)	1	821.4	821.434	13.31	0.000
Name	42	10943.6	260.561	4.22	0.000
Choice-based*FSA-Nutri-scores	1	136.3	136.278	2.21	0.140
Error	128	7898.5	61.707		
Lack-of-Fit	124	7846.0	63.274	4.82	0.066
Pure Error	4	52.5	13.125		
Total	173	19921.4			

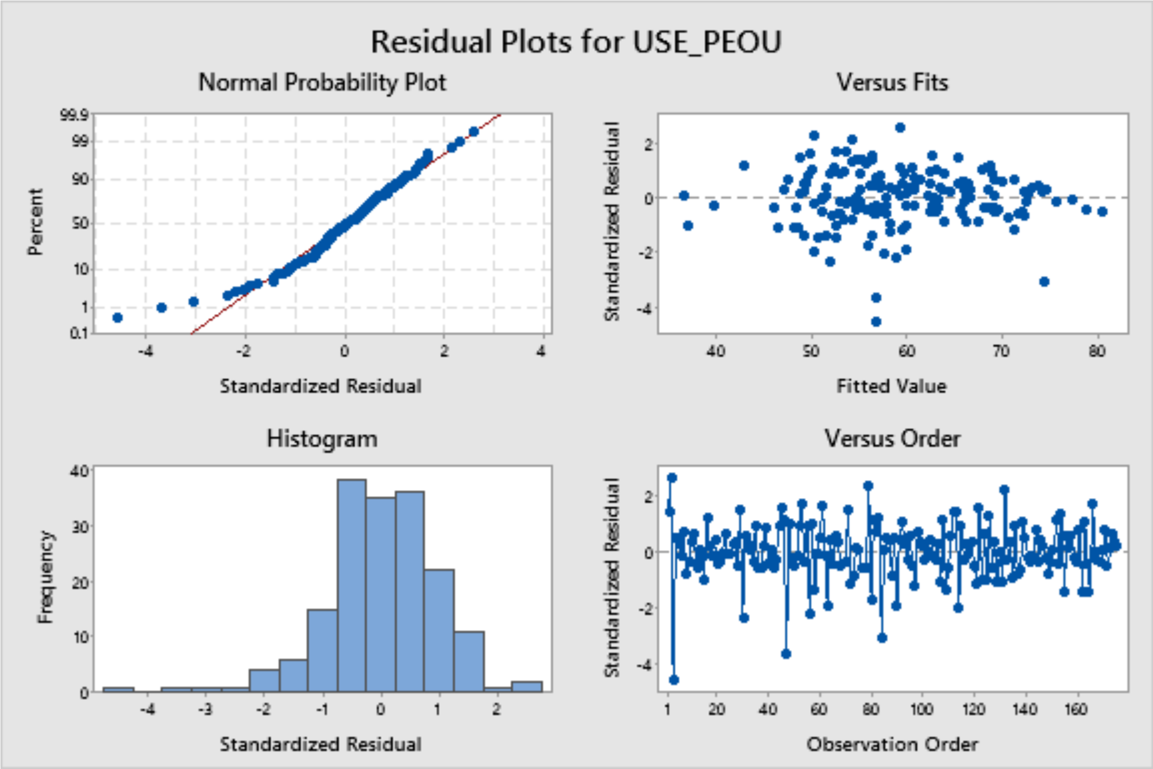


Figure B.2 Standardized Residual Plots for PEOU

General Linear Model: USE_PEOL versus Search UI Layout, Nutrition Information format, Name

Analysis of Variance

Table B.4 ANOVA table for GLM of PEOL

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Searching Results Layout (Choice-based=1)	1	1.07	1.0702	0.25	0.615
Nutrition Information Format (FSA-Nutri-scores=1)	1	32.29	32.2869	7.65	0.007
Name	42	870.75	20.7322	4.91	0.000
Choice-based*FSA-Nutri-scores	1	0.97	0.9661	0.23	0.633
Error	128	540.27	4.2209		
Lack-of-Fit	124	538.27	4.3409	8.68	0.023
Pure Error	4	2.00	0.5000		
Total	173	1467.91			

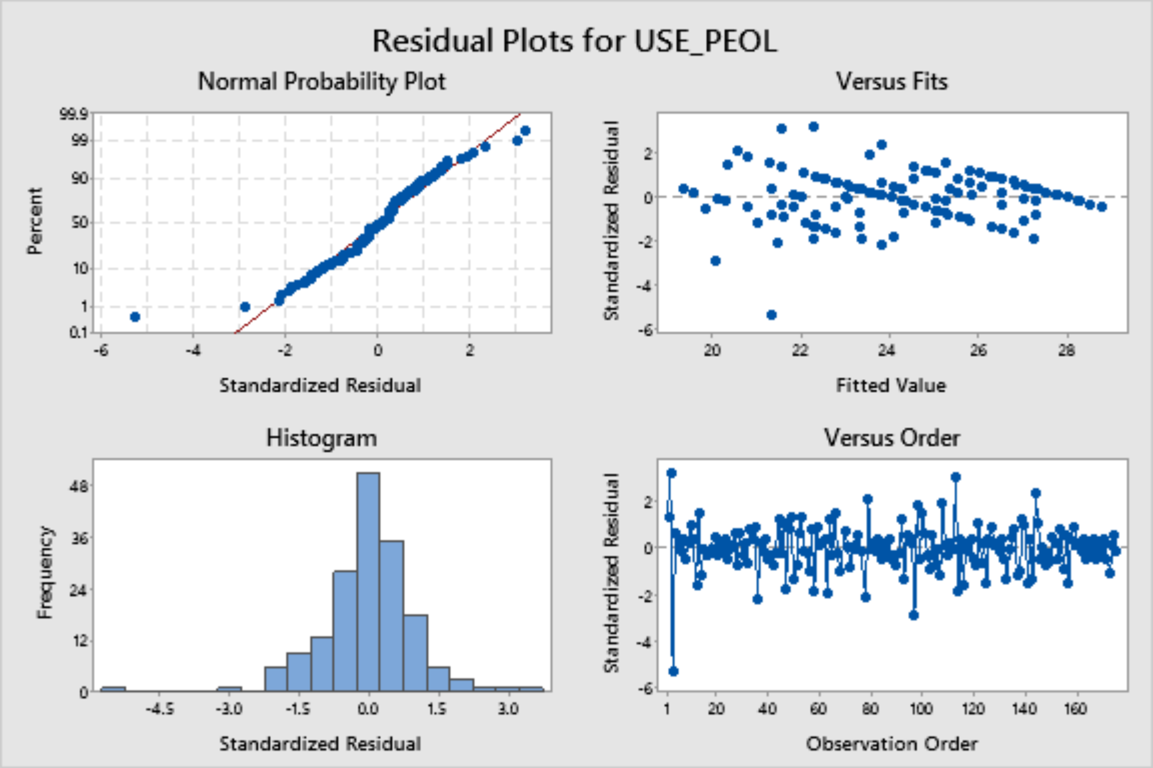


Figure B.3 Standardized Residual Plots for PEOL

General Linear Model: USE_Satisfication versus Search UI Layout, Nutrition Information format, Name

Analysis of Variance

Table B.5 ANOVA table for GLM of Satisfaction

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Searching Results Layout (Choice-based=1)	1	0.6	0.563	0.02	0.892
Nutrition Information Format (FSA-Nutri-scores=1)	1	119.9	119.929	3.95	0.049
Name	42	7175.2	170.839	5.62	0.000
Choice-based*FSA-Nutri-scores	1	14.2	14.179	0.47	0.496
Error	128	3891.0	30.399		
Lack-of-Fit	124	3812.0	30.742	1.56	0.367
Pure Error	4	79.0	19.750		
Total	173	11245.2			

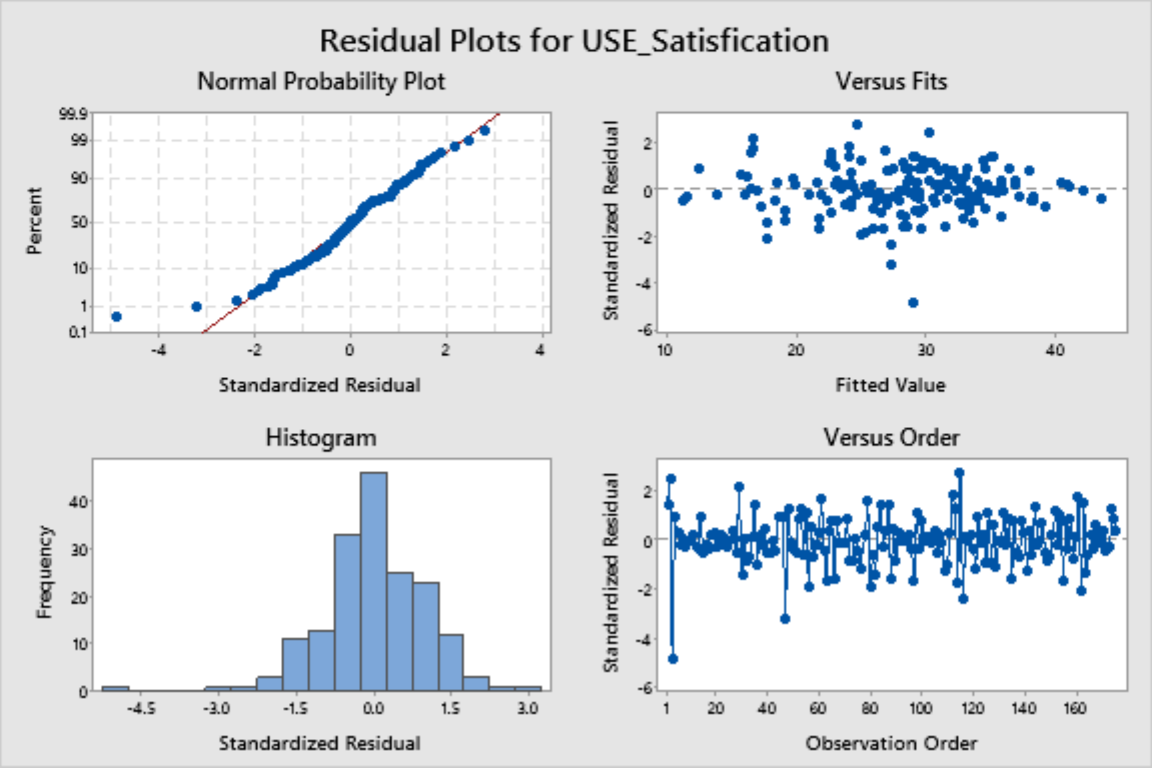


Figure B.4 Standardized Residual Plots for Satisfaction

General Linear Model: NASA_TLX versus Search UI Layout, Nutrition Information format, Name

Analysis of Variance

Table B.6 ANOVA table for GLM of NASA-TLX

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Searching Results Layout (Choice-based=1)	1	33.1	33.07	0.25	0.615
Nutrition Information Format (FSA-Nutri-scores=1)	1	3492.8	3492.76	26.83	0.000
Name	41	24484.7	597.19	4.59	0.000
Choice-based*FSA-Nutri-scores	1	389.3	389.30	2.99	0.086
Error	127	16536.0	130.20		
Lack-of-Fit	123	16418.1	133.48	4.53	0.074
Pure Error	4	117.8	29.46		
Total	171	45621.2			

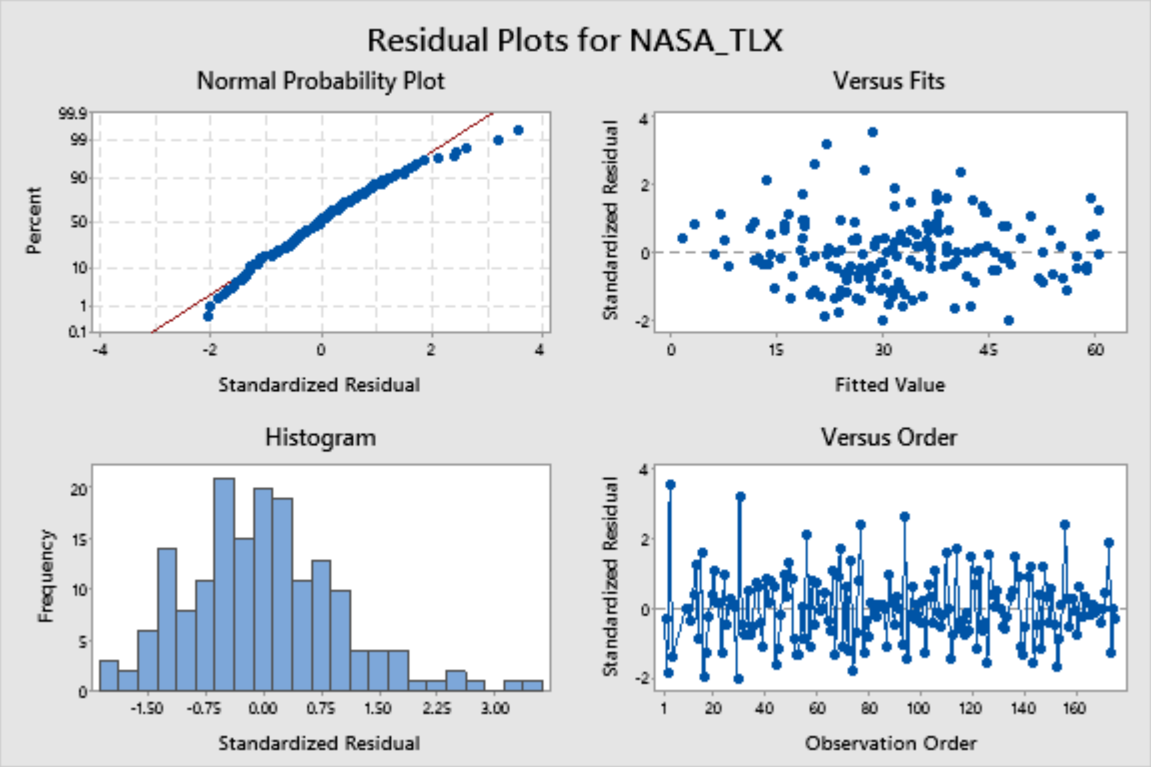


Figure B.5 Standardized Residual Plots for NASA-TLX

APPENDIX C. ANOVA TABLES FOR INDIVIDUAL DIFFERENCE IN STUDY PART 1

Output Corresponding to Table 5. Means of perceived usability and subjective workload between groups (6.3)

Table C.1 Method and Factors Information for One-Way ANOVA

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05
Rows unused	2

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
cluster_membership	3	1, 2, 3

Table C.2 One-Way ANOVA of Individual Difference for PU

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
cluster_membership	2	357.1	178.54	2.11	0.125
Error	171	14478.6	84.67		
Total	173	14835.7			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
9.20165	2.41%	1.27%	0.00%

Means

cluster_membership	N	Mean	StDev	95% CI
1	67	42.40	10.02	(40.18, 44.62)
2	15	42.40	8.98	(37.71, 47.09)
3	92	39.533	8.597	(37.639, 41.426)

Pooled StDev = 9.20165

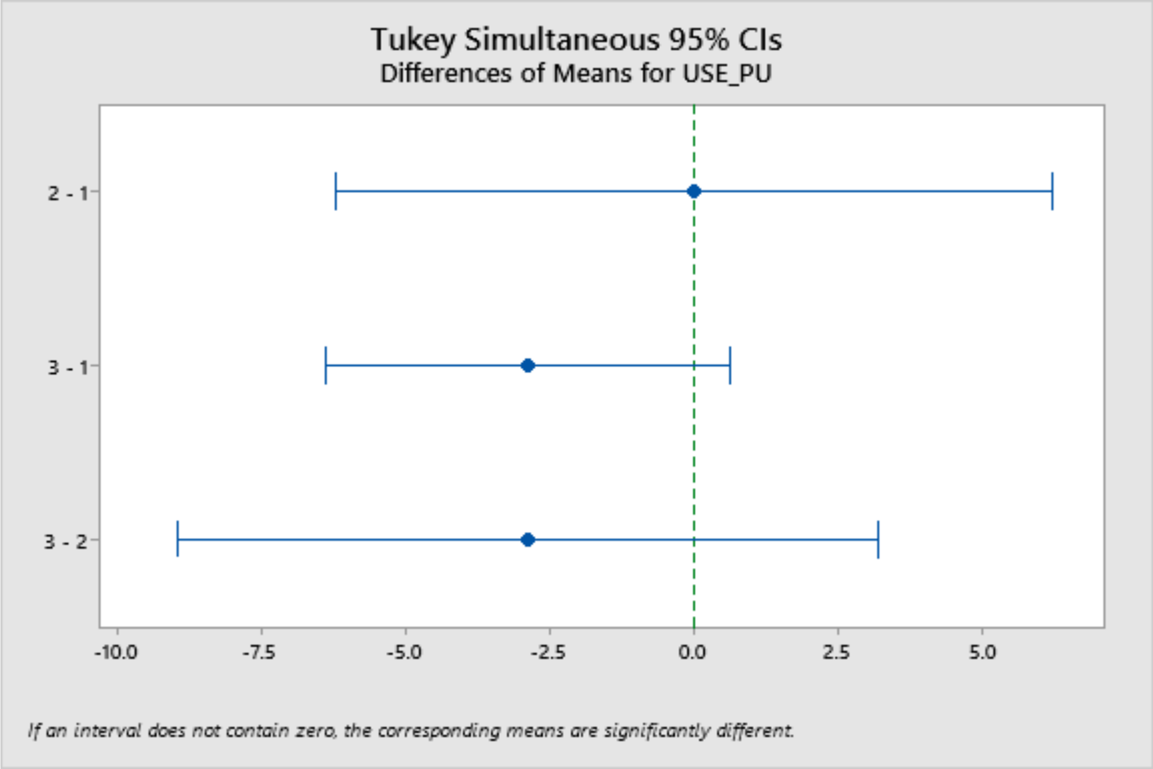
Table C.3 Tukey Post-Hoc Comparison between Groups for PU

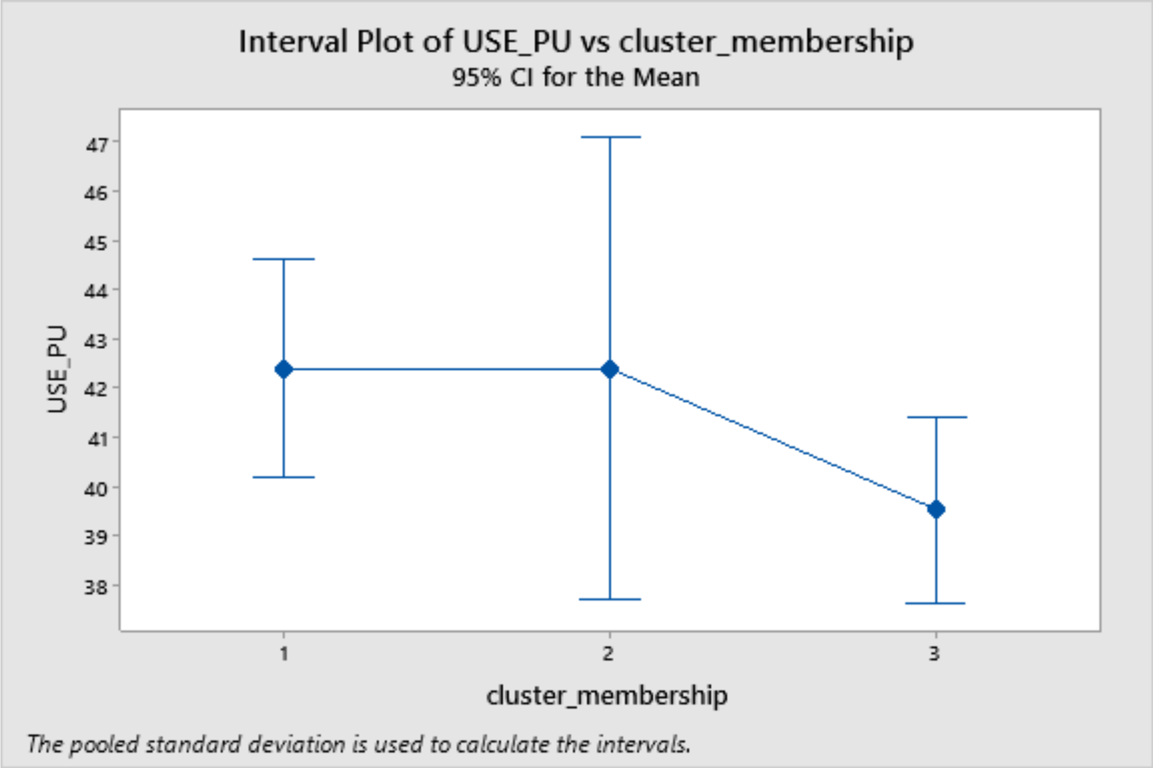
Tukey Pairwise Comparisons

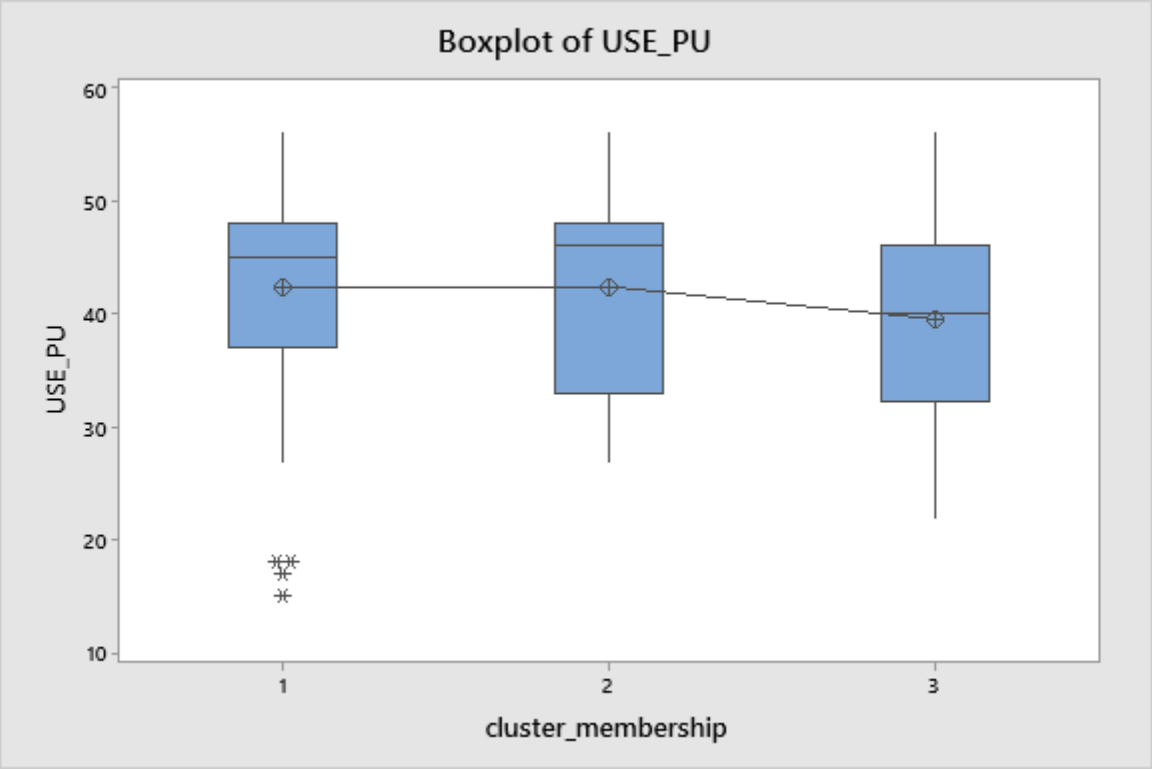
Grouping Information Using the Tukey Method and 95% Confidence

cluster_membership	N	Mean	Grouping
1	67	42.40	A
2	15	42.40	A
3	92	39.533	A

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for USE_PU
2 - 1
3 - 1
3 - 2
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of USE_PU vs cluster_membership
95% CI for the Mean
USE_PU
47
46
45
44
43
42
41
40
39
38
1
2
3
cluster_membership
The pooled standard deviation is used to calculate the intervals.



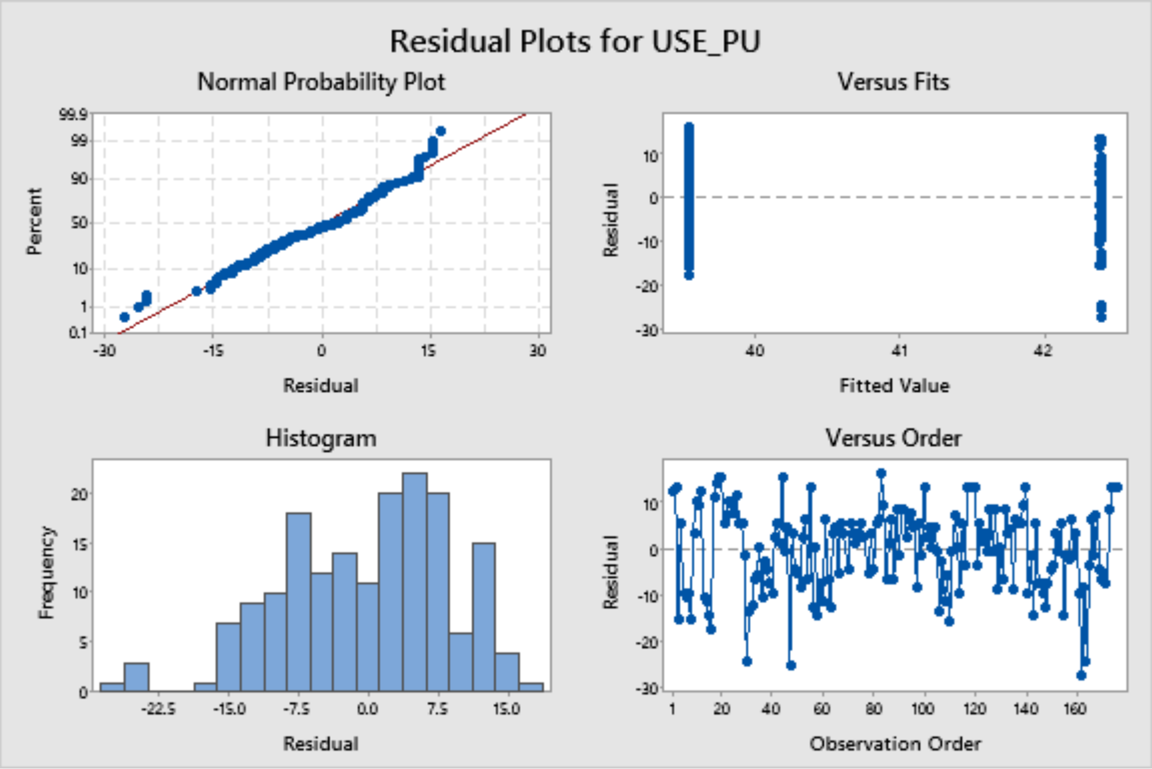


Figure C.1 Residual Plots of PU for Examining the Normality Assumptions

Table C.4 One-Way ANOVA of Individual Difference for PEOU

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
cluster_membership	2	515.8	257.9	2.27	0.106
Error	171	19405.7	113.5		
Total	173	19921.4			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.6529	2.59%	1.45%	0.00%

Means

cluster_membership	N	Mean	StDev	95% CI
1	67	61.49	10.72	(58.92, 64.06)
2	15	60.60	8.62	(55.17, 66.03)
3	92	57.91	10.88	(55.72, 60.11)

Pooled StDev = 10.6529

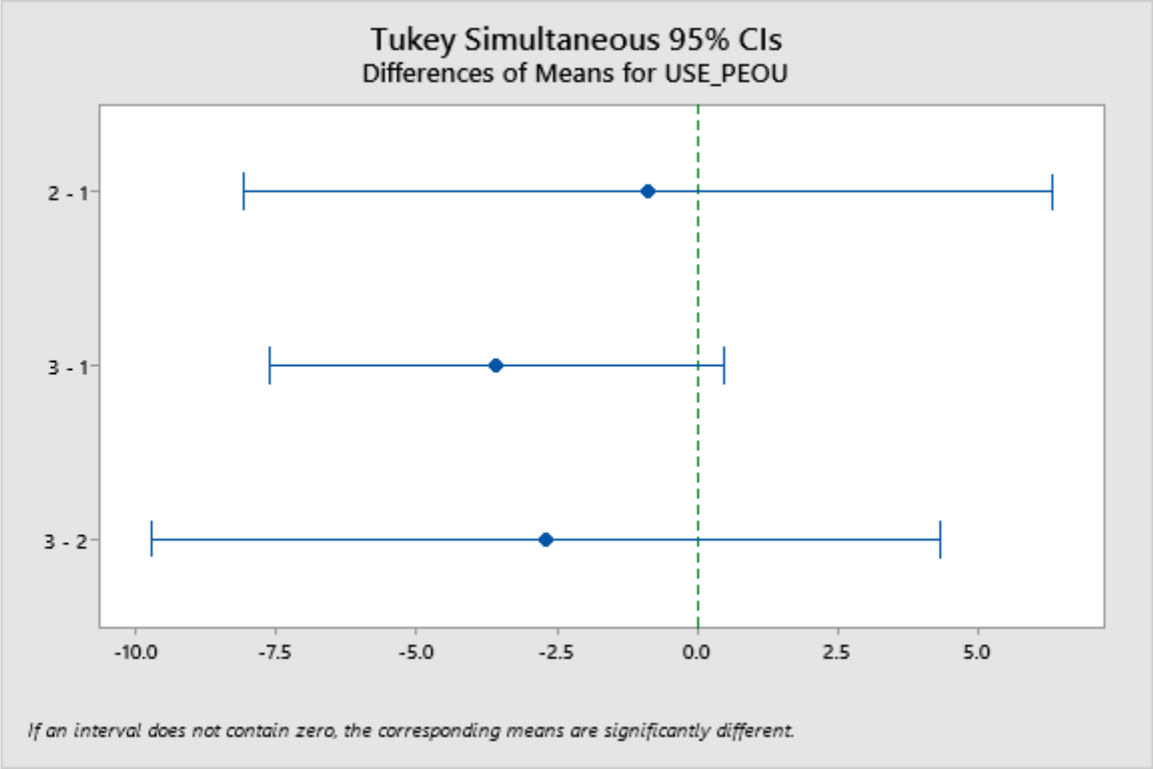
Table C.5 Tukey Post-Hoc Comparison between Groups for PEOU

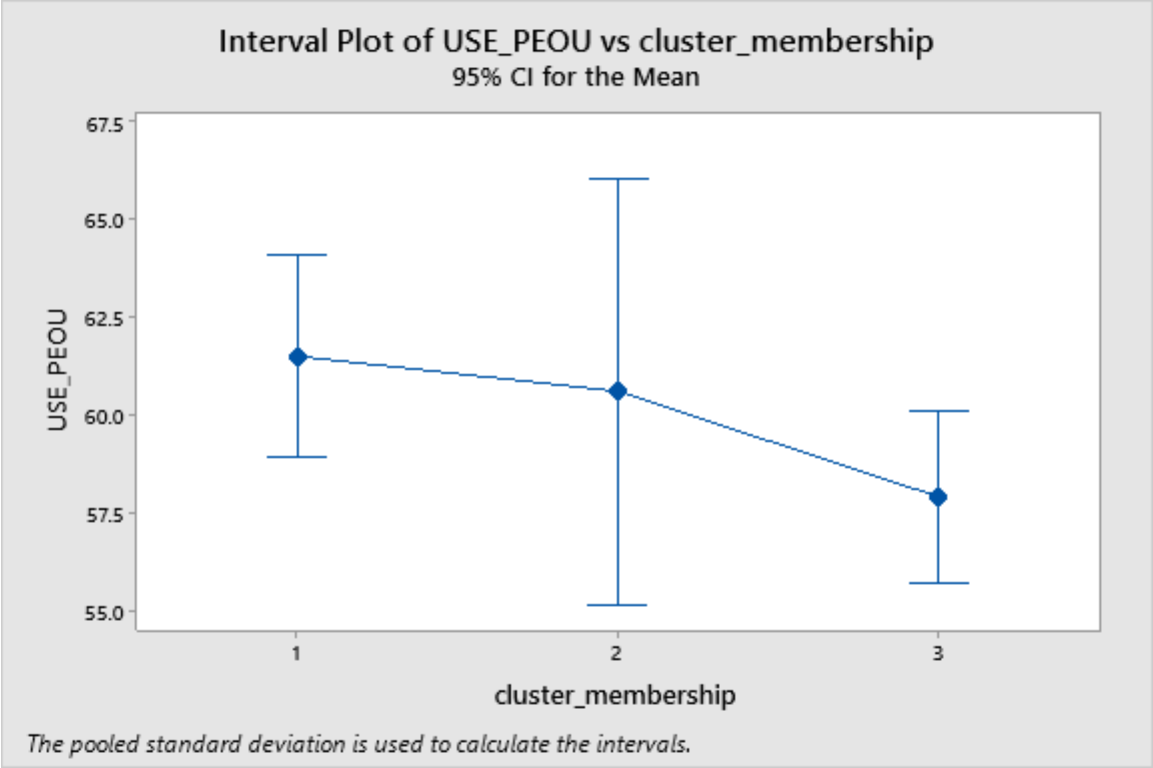
Tukey Pairwise Comparisons

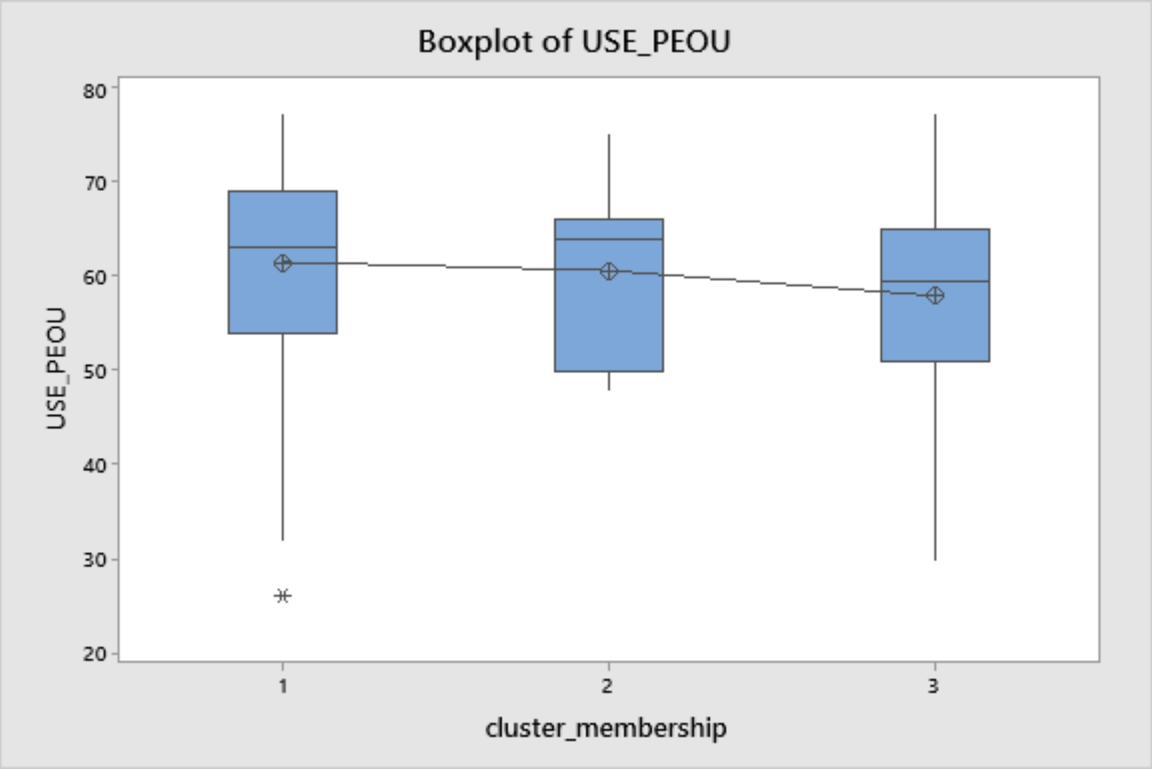
Grouping Information Using the Tukey Method and 95% Confidence

cluster_membership	N	Mean	Grouping
1	67	61.49	A
2	15	60.60	A
3	92	57.91	A

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for USE_PEOU
2 - 1
3 - 1
3 - 2
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of USE_PEOU vs cluster_membership
95% CI for the Mean
USE_PEOU
67.5
65.0
62.5
60.0
57.5
55.0
1
2
3
cluster_membership
The pooled standard deviation is used to calculate the intervals.



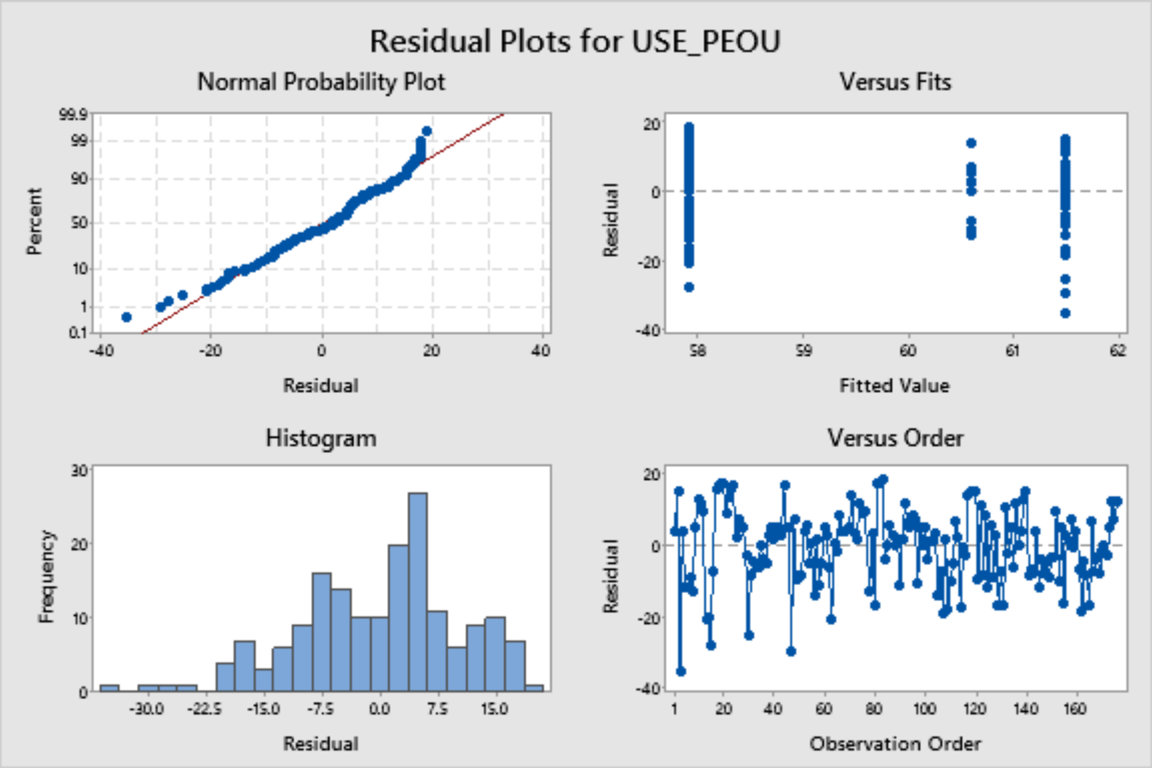


Figure C.2 Residual Plots of PEOU for Examining the Normality Assumptions

Table C.6 One-Way ANOVA of Individual Difference for PEOL

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
cluster_membership	2	83.20	41.601	5.14	0.007
Error	171	1384.71	8.098		
Total	173	1467.91			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.84564	5.67%	4.56%	1.85%

Means

cluster_membership	N	Mean	StDev	95% CI
1	67	24.925	3.178	(24.239, 25.612)
2	15	22.400	3.355	(20.950, 23.850)
3	92	24.804	2.482	(24.219, 25.390)

Pooled StDev = 2.84564

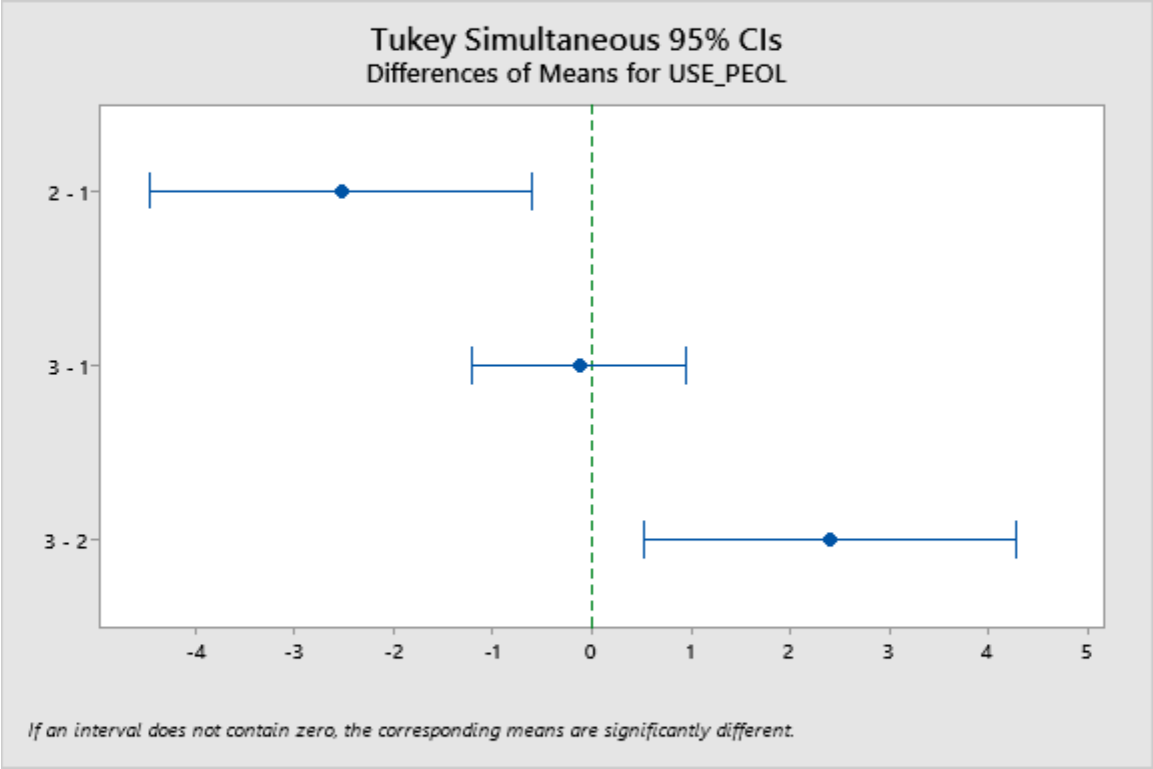
Table C.7 Tukey Post-Hoc Comparison between Groups for PEOL

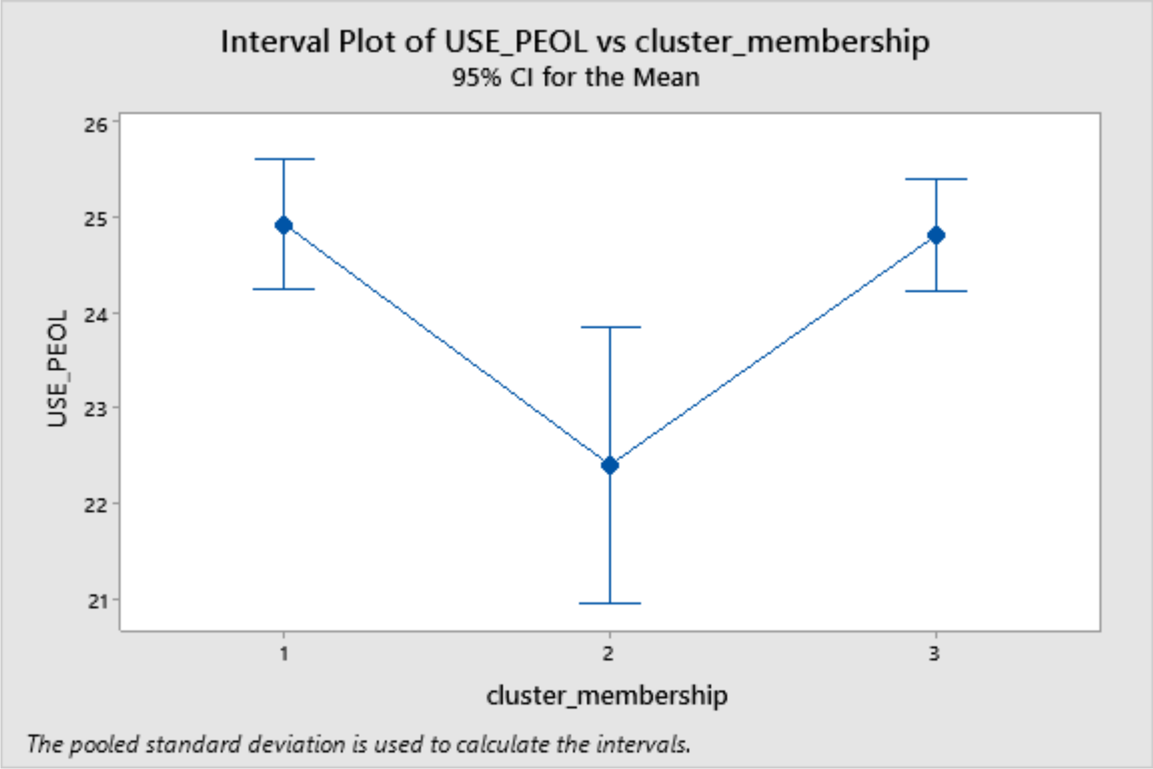
Tukey Pairwise Comparisons

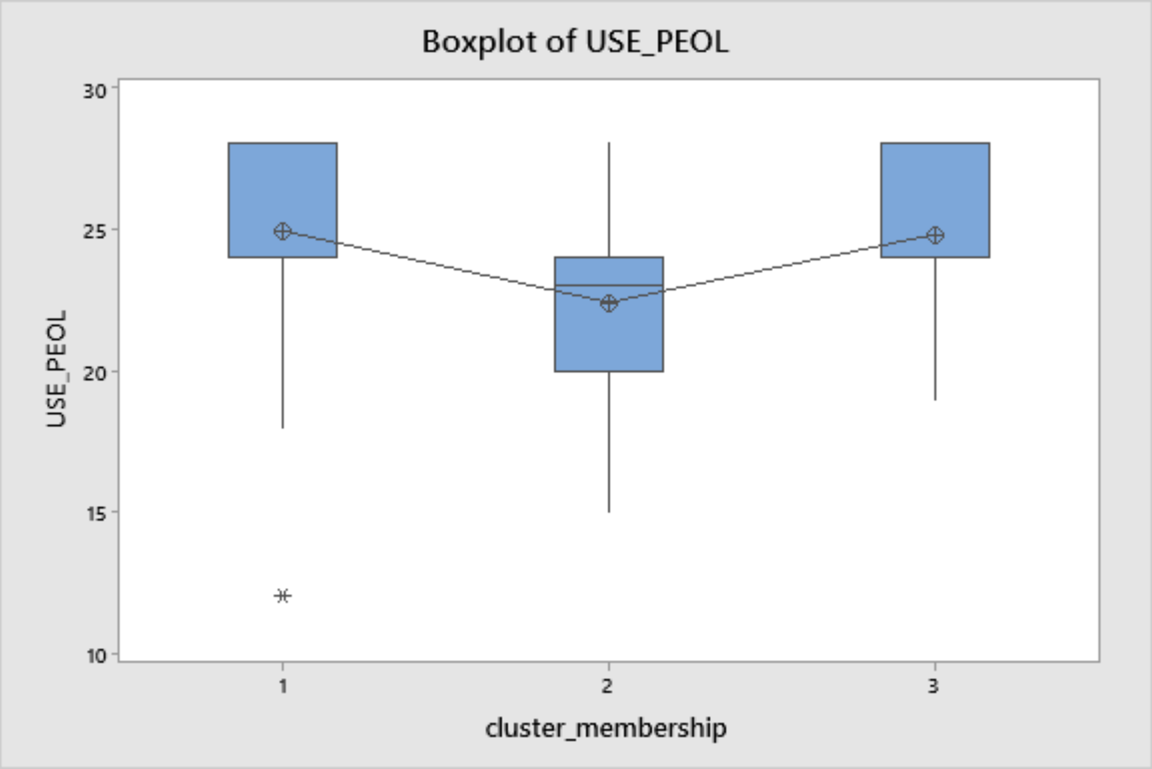
Grouping Information Using the Tukey Method and 95% Confidence

cluster_membership	N	Mean	Grouping	
1	67	24.925	A	
3	92	24.804	A	
2	15	22.400		B

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for USE_PEOL
2 - 1
3 - 1
3 - 2
-4
-3
-2
-1
0
1
2
3
4
5
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of USE_PEOL vs cluster_membership
95% CI for the Mean
USE_PEOL
26
25
24
23
22
21
1
2
3
cluster_membership
The pooled standard deviation is used to calculate the intervals.



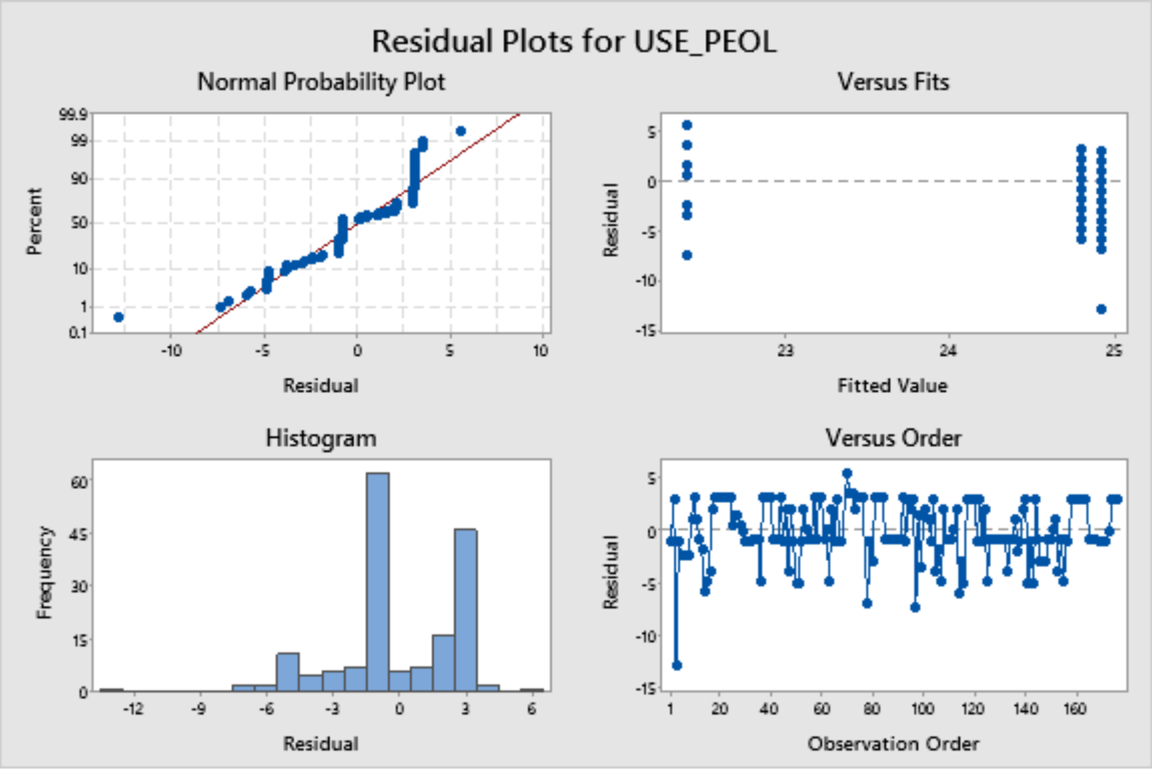


Figure C.3 Residual Plots of PEOL for Examining the Normality Assumptions

Table C.8 One-Way ANOVA of Individual Difference for Satisfaction

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
cluster_membership	2	395.1	197.53	3.11	0.047
Error	171	10850.2	63.45		
Total	173	11245.2			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.96563	3.51%	2.38%	0.45%

Means

cluster_membership	N	Mean	StDev	95% CI
1	67	29.80	8.94	(27.88, 31.72)
2	15	31.13	5.84	(27.07, 35.19)
3	92	27.109	7.485	(25.469, 28.748)

Pooled StDev = 7.96563

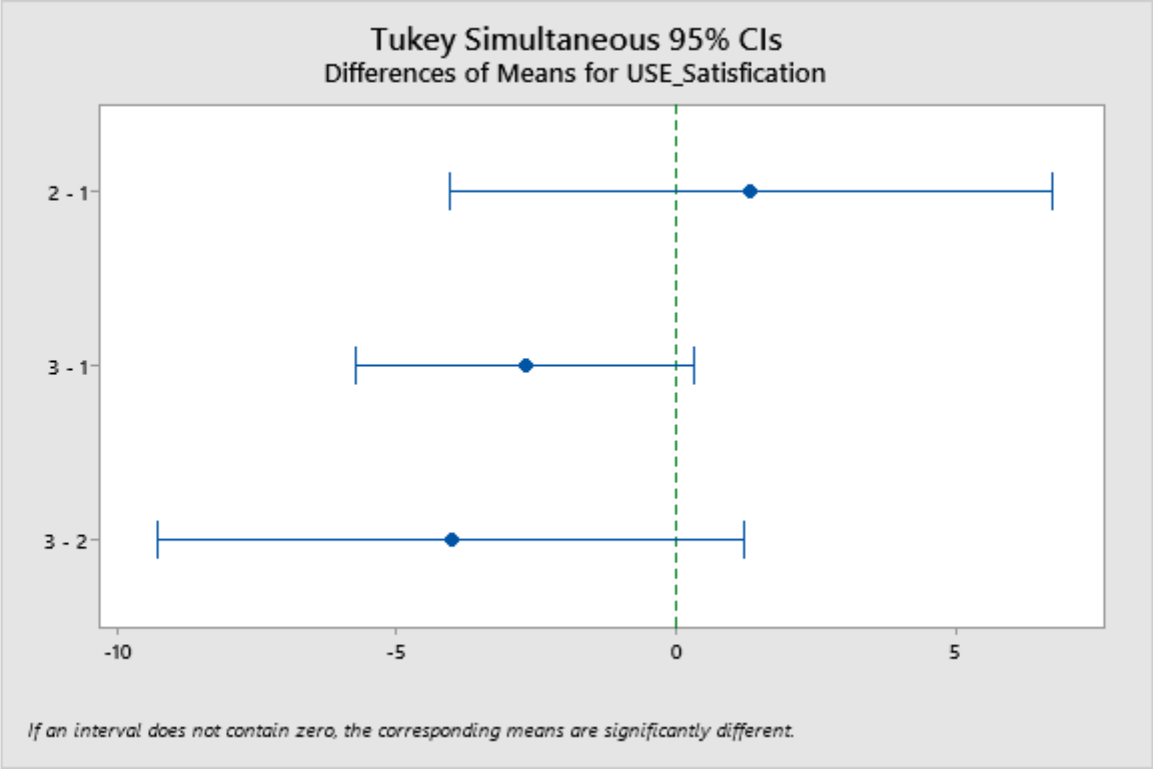
Table C.9 Tukey Post-Hoc Comparison between Groups for Satisfaction

Tukey Pairwise Comparisons

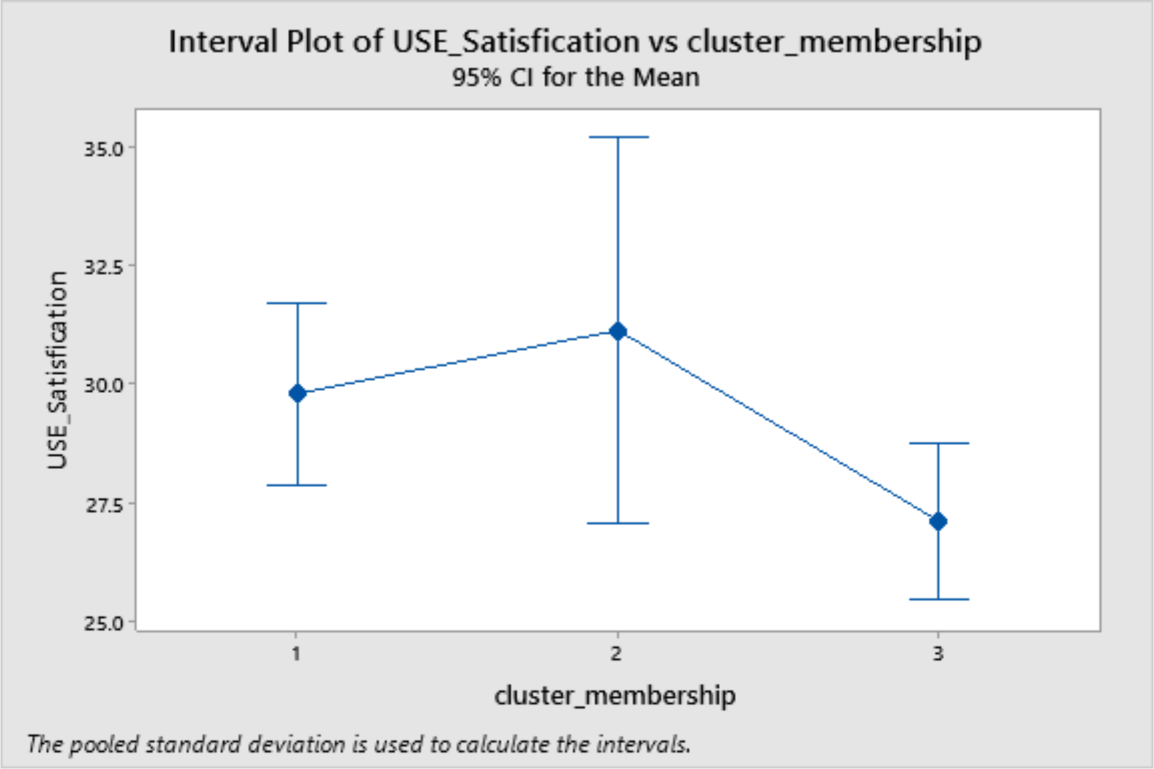
Grouping Information Using the Tukey Method and 95% Confidence

cluster_membership	N	Mean	Grouping
2	15	31.13	A
1	67	29.80	A
3	92	27.109	A

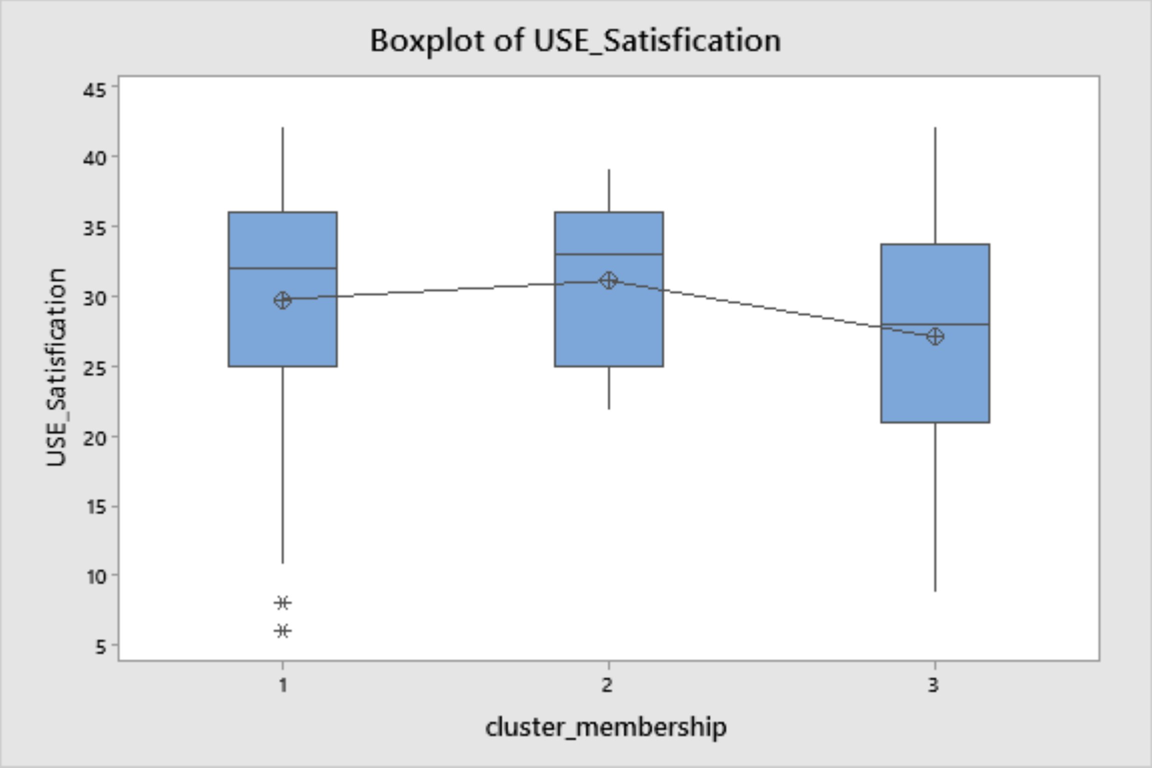
Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for USE_Satisfication

If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of USE_Satisfication vs cluster_membership
95% CI for the Mean

The pooled standard deviation is used to calculate the intervals.



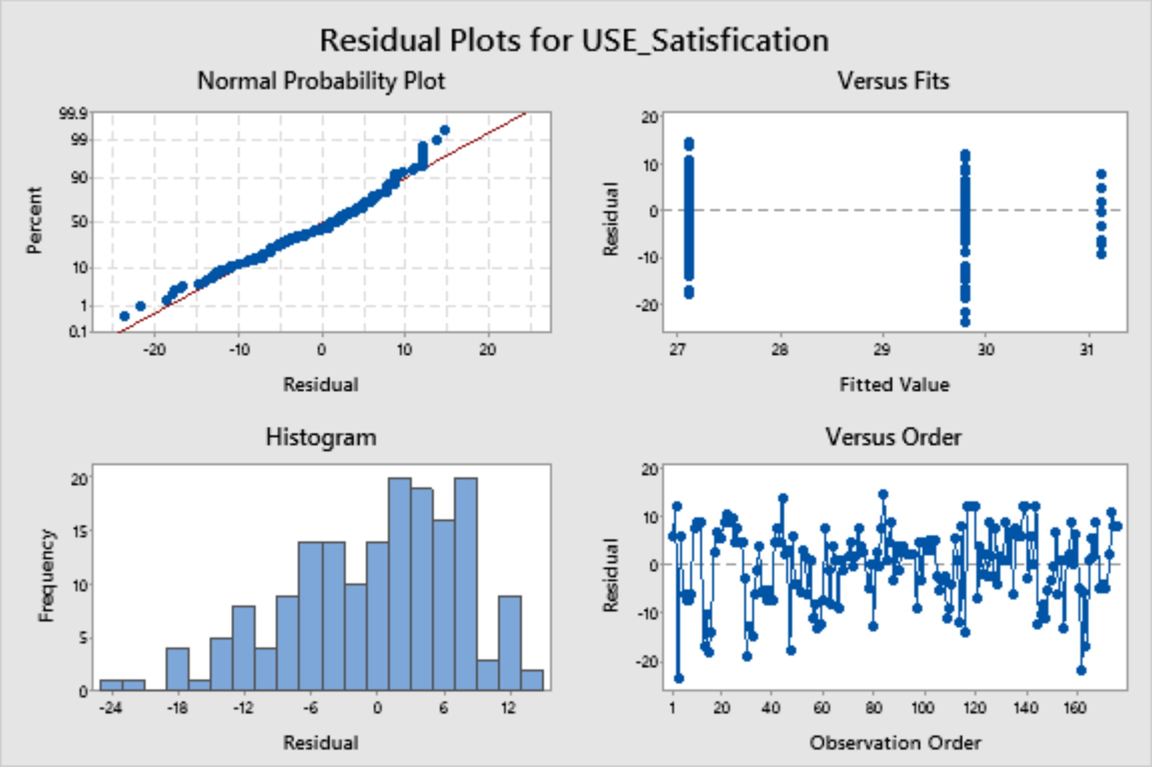
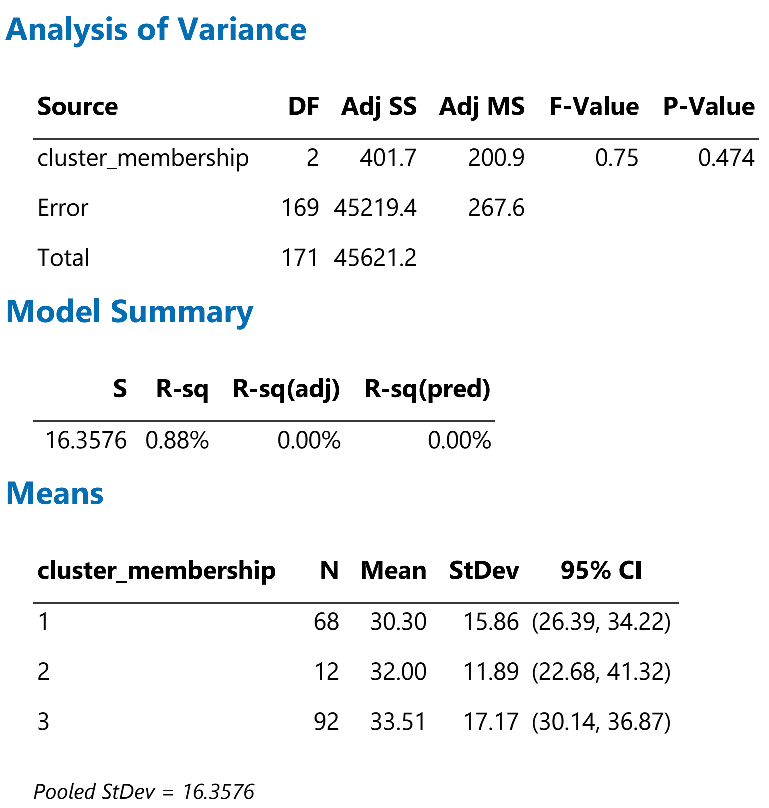


Figure C.4 Residual Plots of Satisfaction for Examining the Normality Assumptions

Table C.10 One-Way ANOVA of Individual Difference for NASA-TLX



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
cluster_membership	2	401.7	200.9	0.75	0.474
Error	169	45219.4	267.6		
Total	171	45621.2			

Model Summary

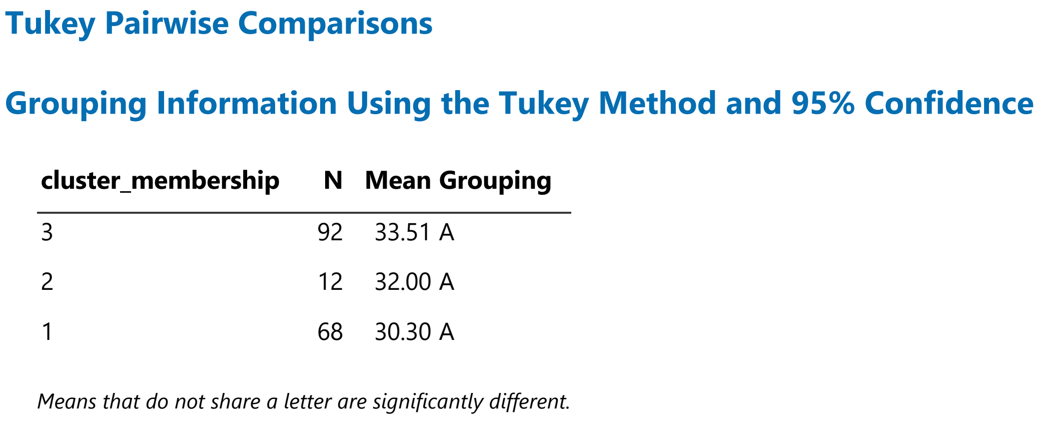
S	R-sq	R-sq(adj)	R-sq(pred)
16.3576	0.88%	0.00%	0.00%

Means

cluster_membership	N	Mean	StDev	95% CI
1	68	30.30	15.86	(26.39, 34.22)
2	12	32.00	11.89	(22.68, 41.32)
3	92	33.51	17.17	(30.14, 36.87)

Pooled StDev = 16.3576

Table C.11 Tukey Post-Hoc Comparison between Groups for NASA-TLX

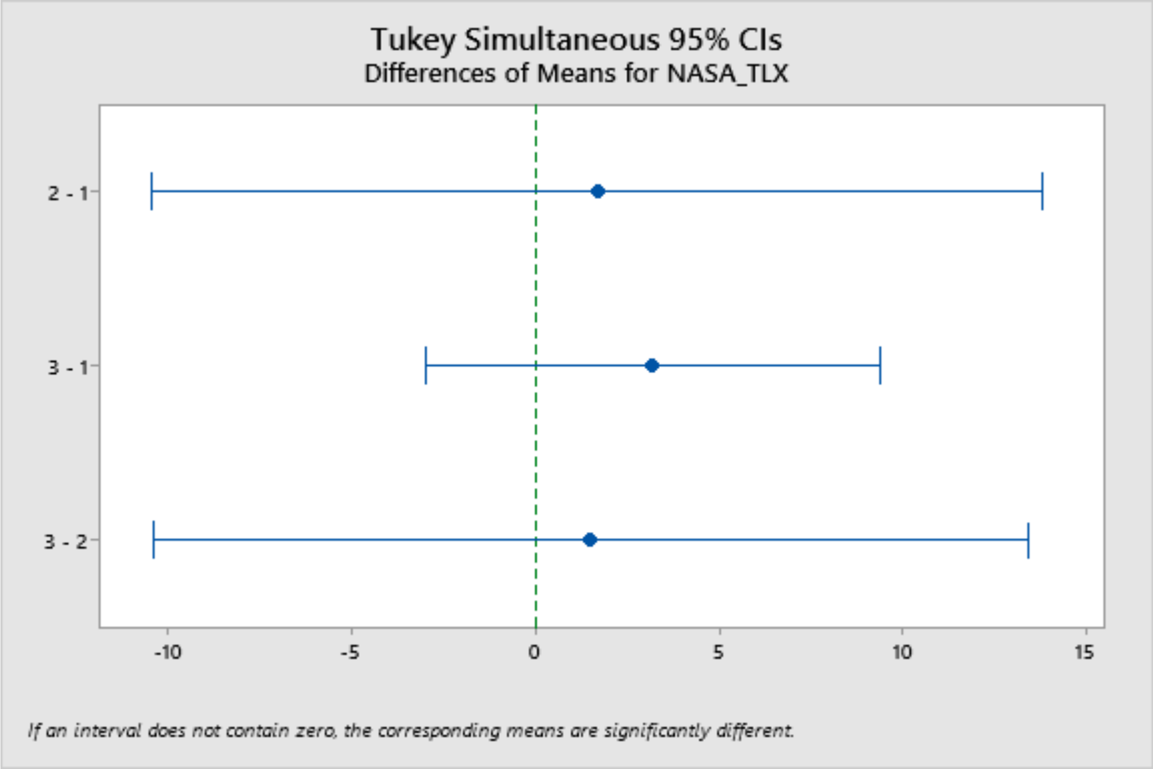


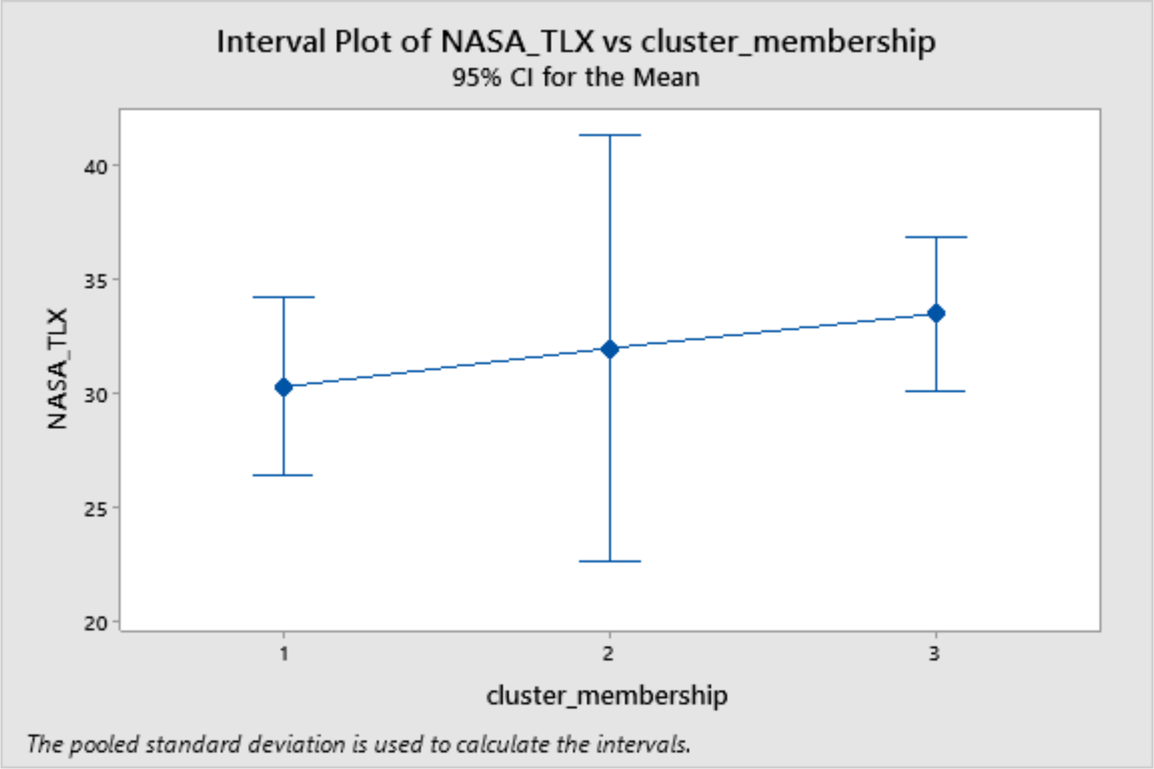
Tukey Pairwise Comparisons

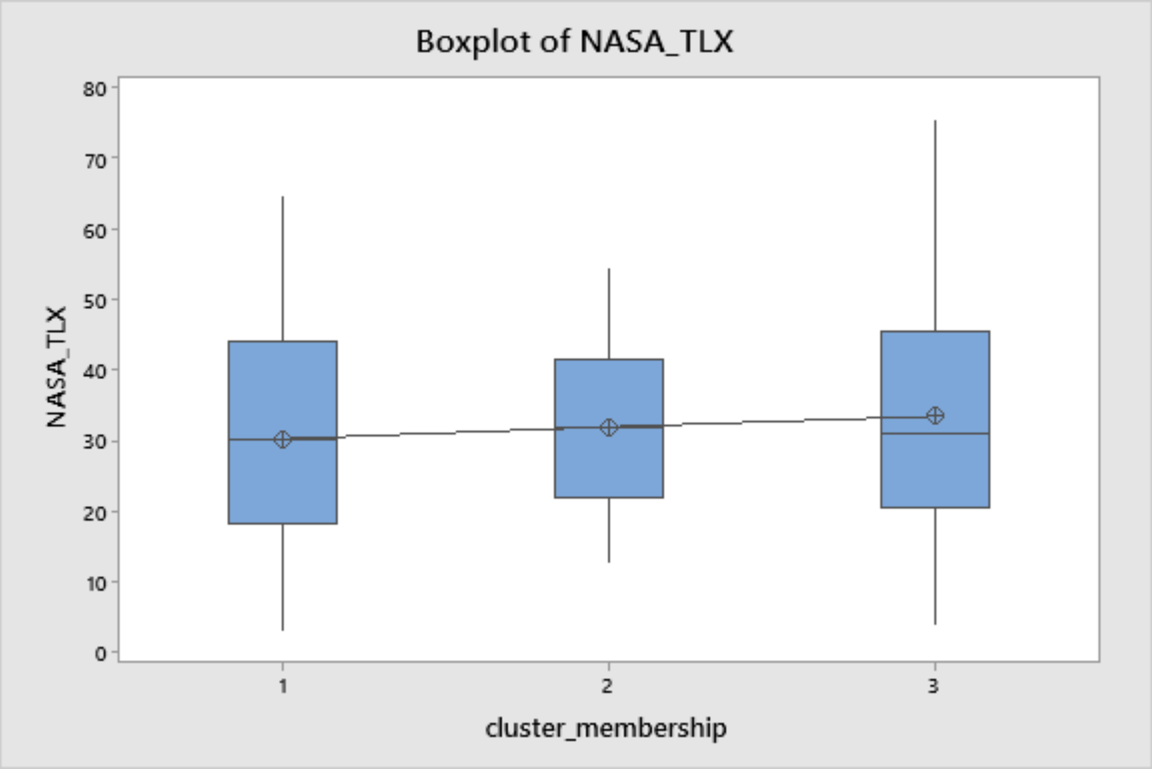
Grouping Information Using the Tukey Method and 95% Confidence

cluster_membership	N	Mean	Grouping
3	92	33.51	A
2	12	32.00	A
1	68	30.30	A

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for NASA_TLX
2 - 1
3 - 1
3 - 2
-10
-5
0
5
10
15
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of NASA_TLX vs cluster_membership
95% CI for the Mean
NASA_TLX
40
35
30
25
20
1
2
3
cluster_membership
The pooled standard deviation is used to calculate the intervals.



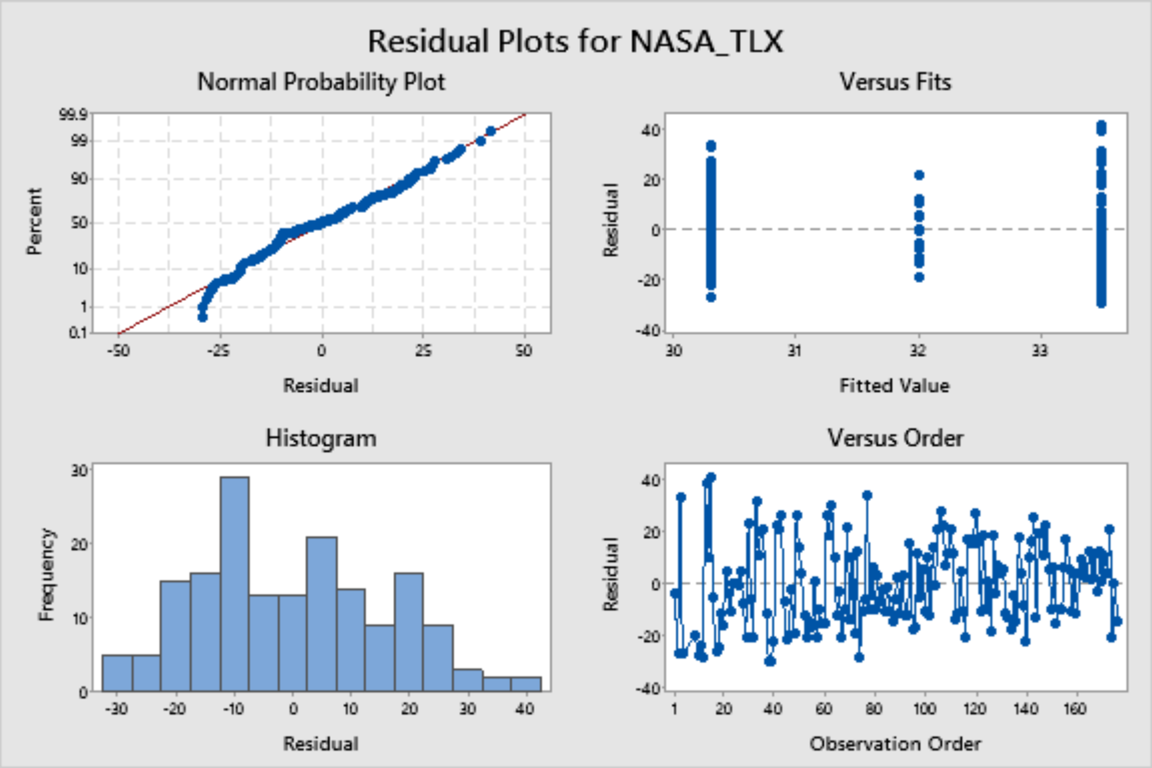


Figure C.5 Residual Plots of NASA-TLX for Examining the Normality Assumptions

APPENDIX D. ANOVA TABLES FOR GLMS OF STUDY PART 2

Appendix Corresponding to Table 6. Summary of GLMs for human decision performance ANOVA Tables for GLMs of Study Part 2 (7.2)

Table D.1 Factors Information

Factor Information

Factor	Type	Levels	Values
Decision Paradigm (2AFC = 1)	Fixed	2	0, 1
Nutrition Information Format (Nutri-scores =1)	Fixed	2	0, 1
System Default (Pre-selection=1)	Fixed	2	0, 1
Name	Random	36	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36

General Linear Model: d-prime versus Decision Paradigm, Nutrition Information Format , System Default , Name

Table D.2 ANOVA table for GLM of d-prime

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	0.8054	0.80544	6.09	0.014
Nutrition Information Format (Nutri-scores =1)	1	0.0201	0.02014	0.15	0.697
System Default (Pre-selection=1)	1	0.4570	0.45697	3.46	0.064
Name	35	6.1536	0.17582	1.33	0.112
2AFC* Nutri-scores	1	0.1478	0.14779	1.12	0.291
2AFC* pre-selection	1	0.2606	0.26062	1.97	0.162
Nutri-scores* pre-selection	1	0.0834	0.08345	0.63	0.428
2AFC* Nutri-scores* pre-selection	1	0.1637	0.16374	1.24	0.267
Error	237	31.3274	0.13218		
Lack-of-Fit	236	31.3051	0.13265	5.94	0.318
Pure Error	1	0.0223	0.02234		
Total	279	46.1658			

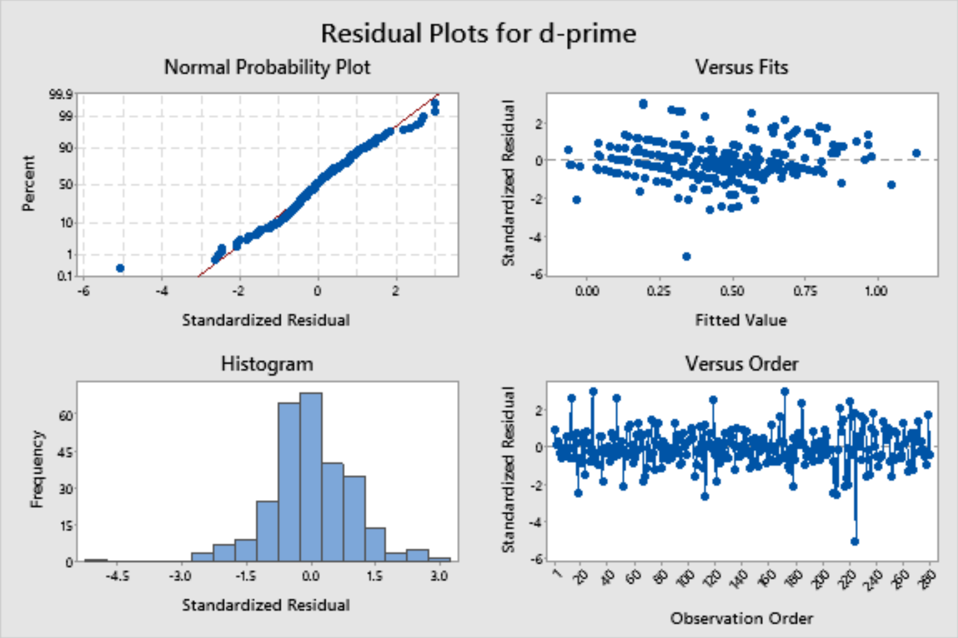


Figure D.1 Standardized Residual Plots for d-prime

General Linear Model: Accuracy versus Decision Paradigm, Nutrition Information Format, System Default , Name

Table D.3 ANOVA table for GLM of Accuracy

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	1.6524	1.65235	51.73	0.000
Nutrition Information Format (Nutri-scores =1)	1	0.0013	0.00132	0.04	0.839
System Default (Pre-selection=1)	1	0.2974	0.29739	9.31	0.003
Name	35	1.1227	0.03208	1.00	0.468
2AFC* Nutri-scores	1	0.0853	0.08530	2.67	0.104
2AFC* pre-selection	1	0.1914	0.19139	5.99	0.015
Nutri-scores* pre-selection	1	0.1035	0.10346	3.24	0.073
2AFC* Nutri-scores* pre-selection	1	0.0833	0.08331	2.61	0.108
Error	232	7.4106	0.03194		
Lack-of-Fit	231	7.3906	0.03199	1.60	0.570
Pure Error	1	0.0200	0.02000		
Total	274	20.3511			

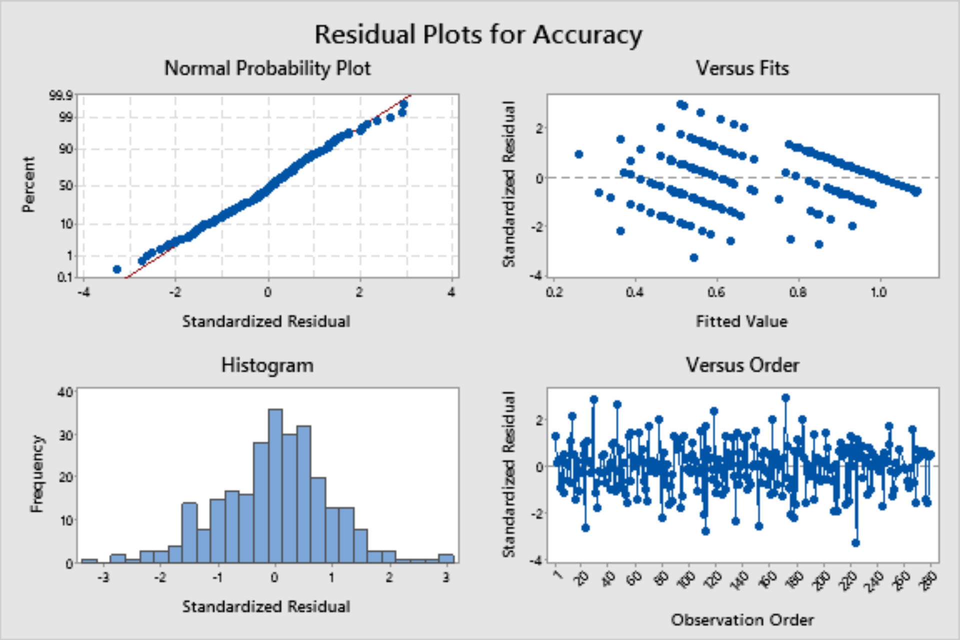


Figure D.2 Standardized Residual Plots for Accuracy

General Linear Model: c versus Decision Paradigm, Nutrition Information Format, System Default , Name

Table D.4 ANOVA table for GLM of response criterion, c

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	1.4214	1.42142	67.31	0.000
Nutrition Information Format (Nutri-scores =1)	1	0.0739	0.07390	3.50	0.063
System Default (Pre-selection=1)	1	0.0807	0.08066	3.82	0.052
Name	35	3.6480	0.10423	4.94	0.000
2AFC* Nutri-scores	1	0.0544	0.05439	2.58	0.110
2AFC* pre-selection	1	0.0564	0.05638	2.67	0.104
Nutri-scores* pre-selection	1	0.0680	0.06798	3.22	0.074
2AFC* Nutri-scores* pre-selection	1	0.0821	0.08214	3.89	0.050
Error	237	5.0047	0.02112		
Lack-of-Fit	236	5.0040	0.02120	29.96	0.145
Pure Error	1	0.0007	0.00071		
Total	279	12.9136			

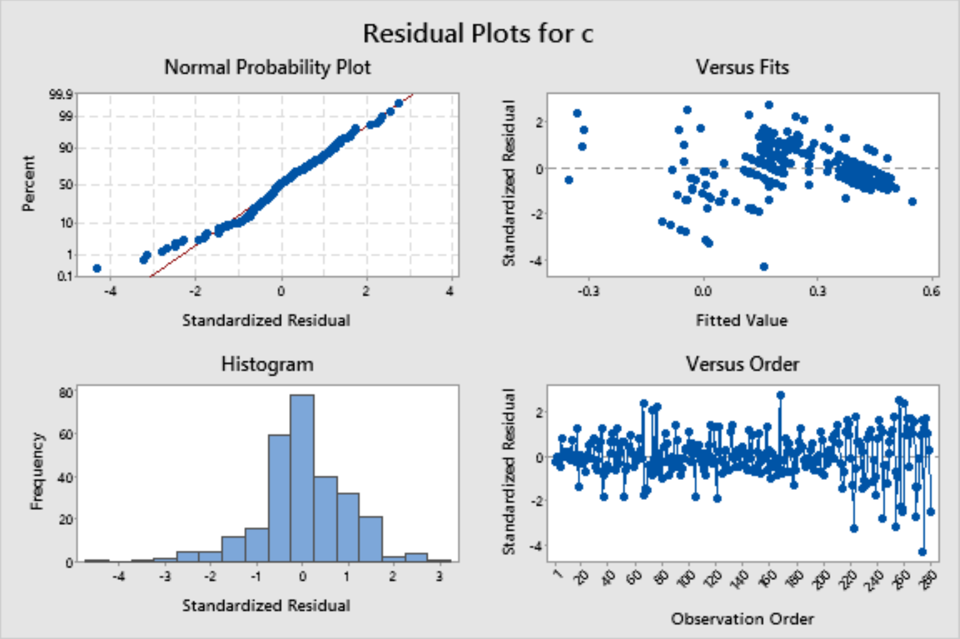


Figure D.3 Standardized Residual Plots for c

General Linear Model: Time of Response_overall versus Decision Paradigm, Nutrition Information Format, System Default , Name

Table D.5 ANOVA table for GLM of Time of Response

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	36	35.6	0.03	0.865
Nutrition Information Format (Nutri-scores =1)	1	16096	16095.5	13.13	0.000
System Default (Pre-selection=1)	1	11322	11321.6	9.23	0.003
Name	35	156955	4484.4	3.66	0.000
2AFC* Nutri-scores	1	35	34.9	0.03	0.866
2AFC* pre-selection	1	197	197.5	0.16	0.689
Nutri-scores* pre-selection	1	9856	9855.6	8.04	0.005
2AFC* Nutri-scores* pre-selection	1	763	763.1	0.62	0.431
Error	205	251365	1226.2		
Lack-of-Fit	204	251365	1232.2	*	*
Pure Error	1	0	0.0		
Total	247	445939			

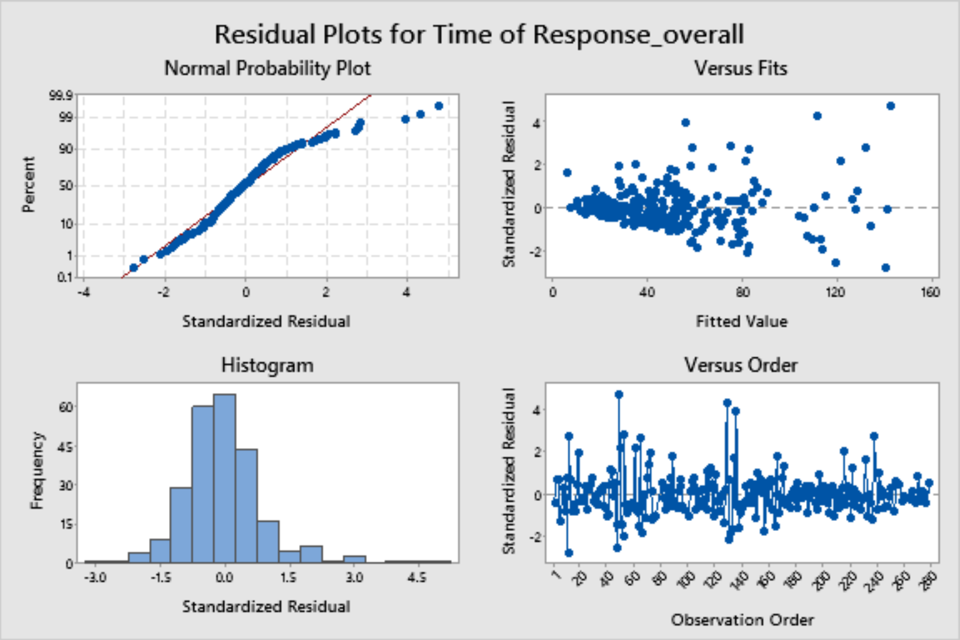


Figure D.4 Standardized Residual Plots for Time of Response

General Linear Model: NASA_TLX versus Decision Paradigm, Nutrition Information Format, System Default, Name

Table D.6 ANOVA table for GLM of NASA-TLX

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	39.0	39.00	0.32	0.575
Nutrition Information Format (Nutri-scores =1)	1	2048.0	2048.04	16.57	0.000
System Default (Pre-selection=1)	1	99.3	99.29	0.80	0.371
Name	35	53072.8	1516.37	12.27	0.000
2AFC* Nutri-scores	1	15.0	14.97	0.12	0.728
2AFC* pre-selection	1	192.5	192.54	1.56	0.213
Nutri-scores* pre-selection	1	204.2	204.16	1.65	0.200
2AFC* Nutri-scores* pre-selection	1	59.5	59.47	0.48	0.489
Error	233	28799.0	123.60		
Lack-of-Fit	232	28790.3	124.10	14.30	0.208
Pure Error	1	8.7	8.68		
Total	275	87677.6			

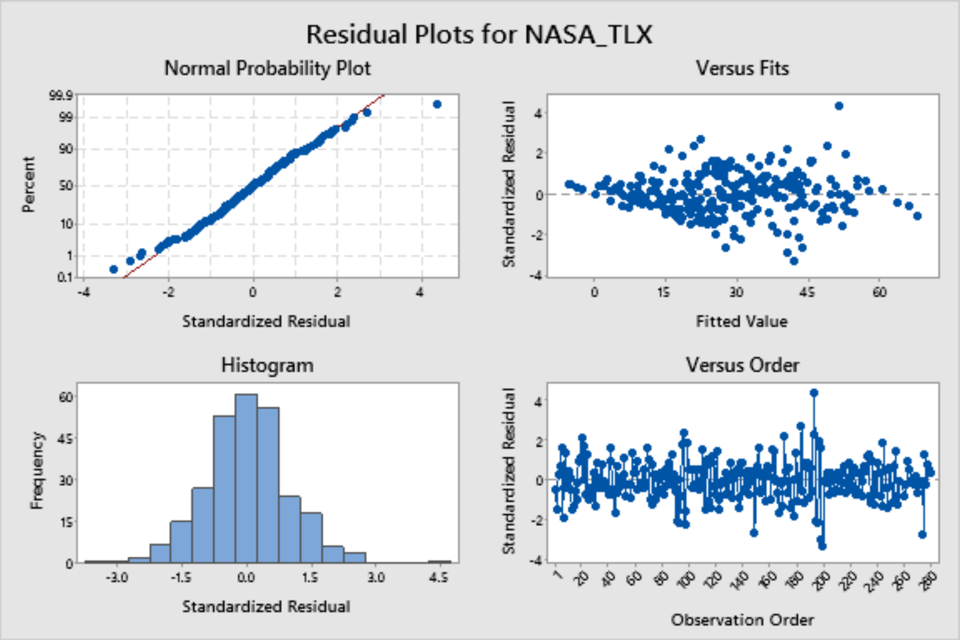


Figure D.5 Standardized Residual Plots for NASA-TLX

APPENDIX E. ANOVA TABLES FOR INDIVIDUAL DIFFERENCE IN STUDY PART 2

Appendix Outputs Corresponding to Table 7. Means of task performance and subjective workload (7.3)

Table E.1 Method and Factors Information

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Cluster_Membership	3	1, 2, 3

Table E.2 One-Way ANOVA of Individual Difference for d-prime

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Cluster_Membership	2	2.093	1.047	0.42	0.658
Error	277	692.003	2.498		
Total	279	694.096			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.58057	0.30%	0.00%	0.00%

Means

Cluster_Membership	N	Mean	StDev	95% CI
1	88	0.944	1.686	(0.612, 1.276)
2	56	0.918	1.354	(0.503, 1.334)
3	136	1.106	1.596	(0.839, 1.373)

Pooled StDev = 1.58057

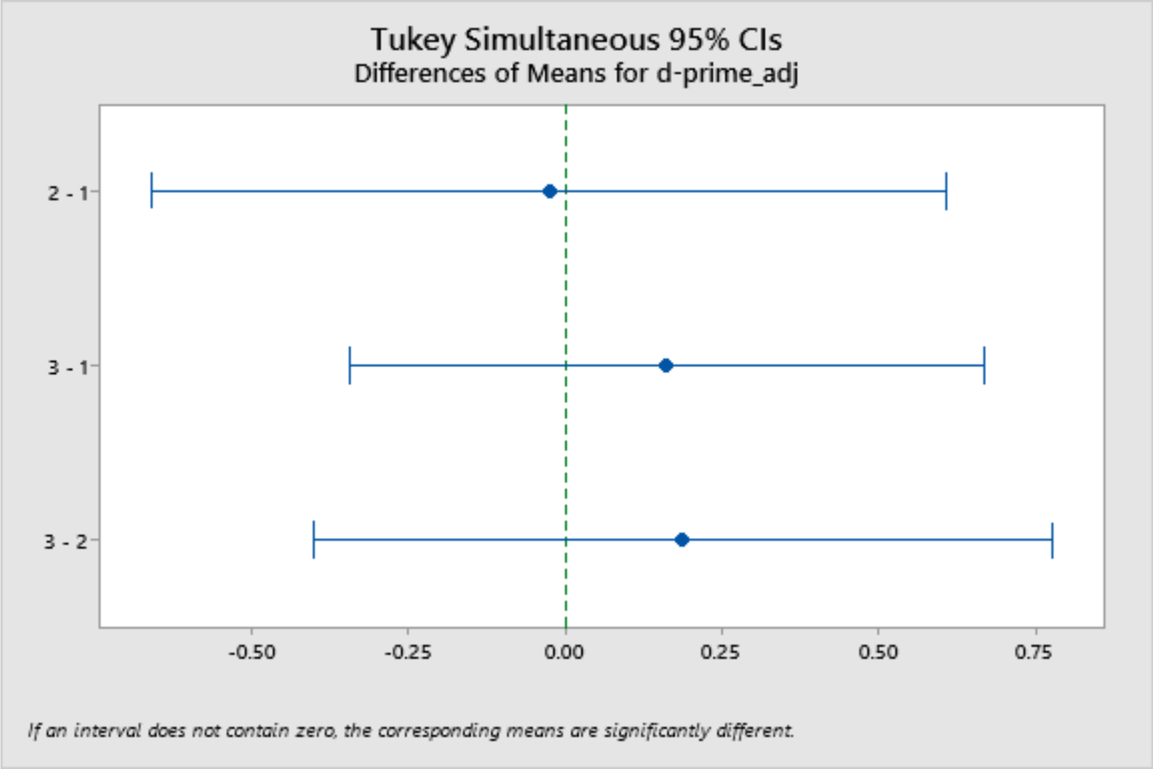
Table E.3 Tukey Post-Hoc Comparison between Groups for d-prime

Tukey Pairwise Comparisons

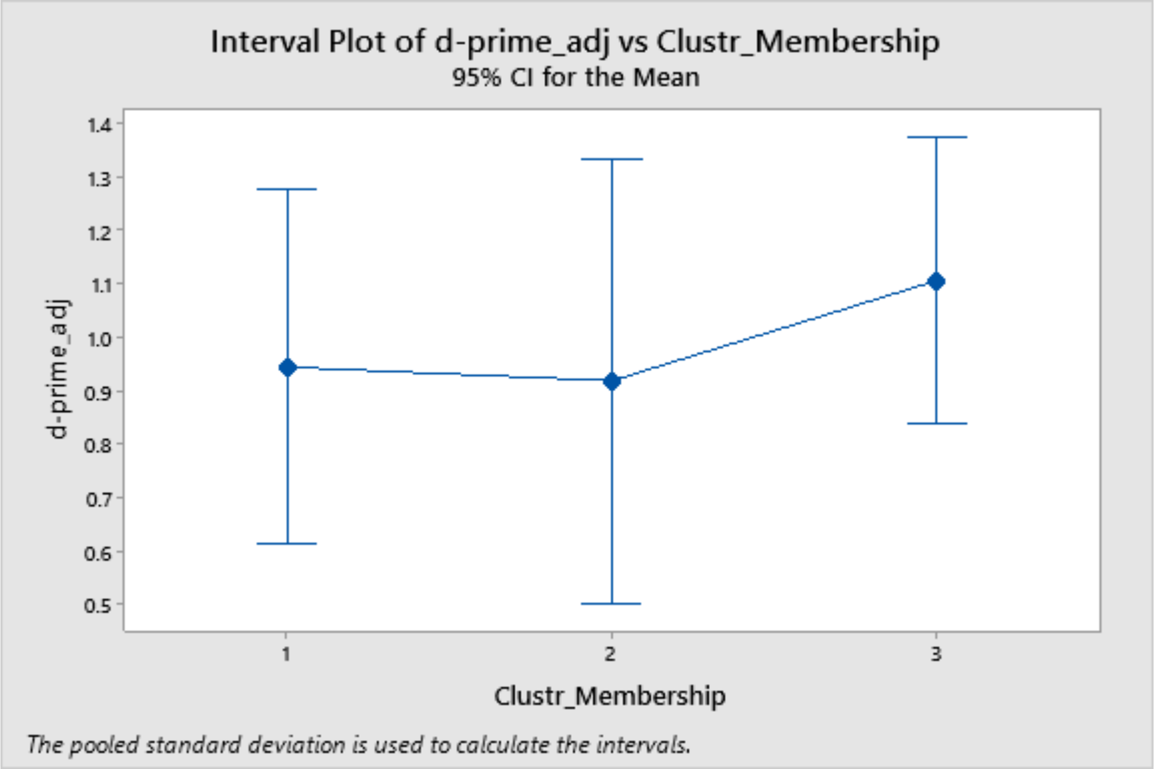
Grouping Information Using the Tukey Method and 95% Confidence

Cluster_Membership	N	Mean	Grouping
3	136	1.106	A
1	88	0.944	A
2	56	0.918	A

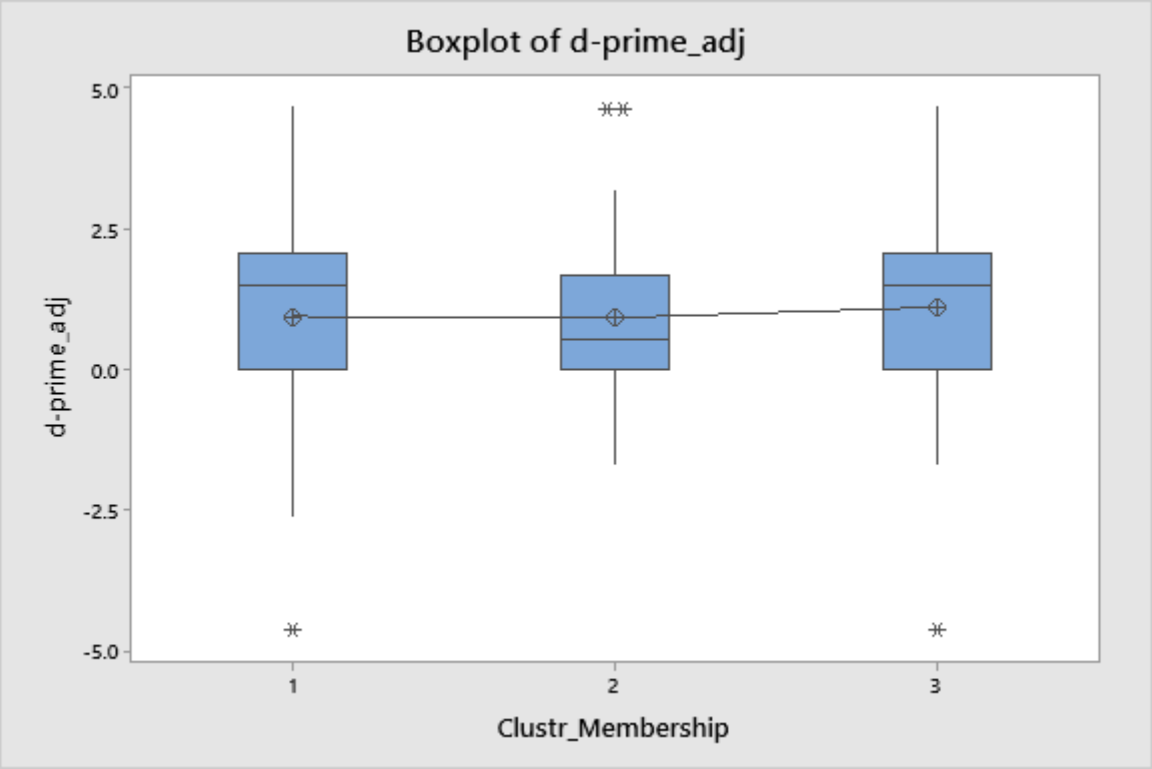
Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for d-prime_adj

If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of d-prime_adj vs Clustr_Membership
95% CI for the Mean

The pooled standard deviation is used to calculate the intervals.



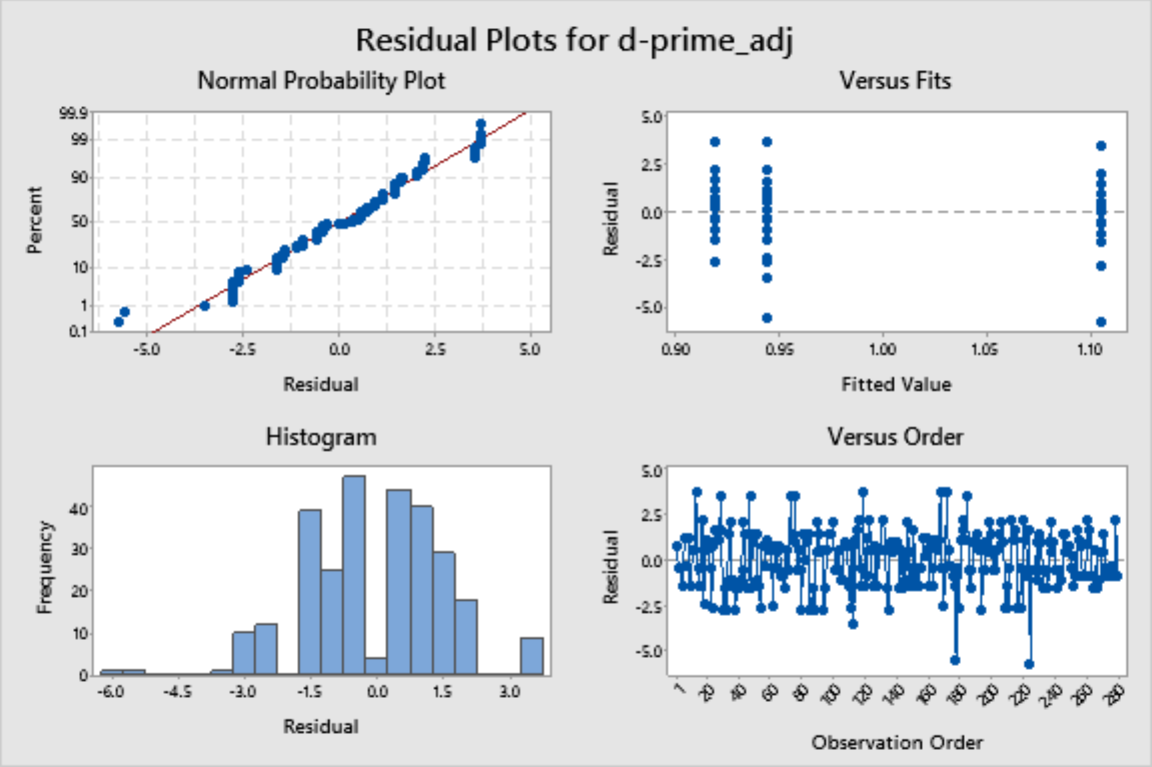


Figure E.1 Residual Plots of d-prime for Examining the Normality Assumptions

Table E.4 One-Way ANOVA of Individual Difference for Accuracy

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Cluster_Membership	2	0.0633	0.03164	0.39	0.675
Error	272	21.8358	0.08028		
Total	274	21.8991			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.283335	0.29%	0.00%	0.00%

Means

Cluster_Membership	N	Mean	StDev	95% CI
1	83	0.7542	0.3221	(0.6930, 0.8154)
2	56	0.7107	0.2440	(0.6362, 0.7853)
3	136	0.7368	0.2729	(0.6889, 0.7846)

Pooled StDev = 0.283335

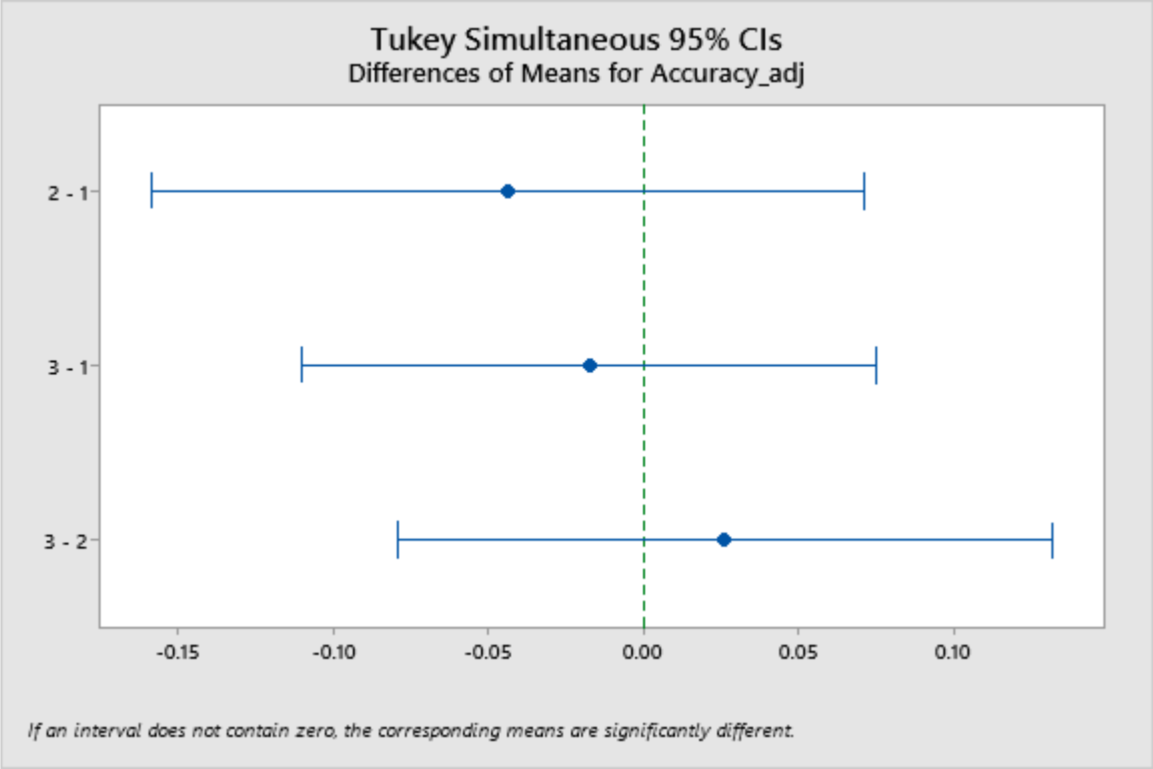
Table E.5 Tukey Post-Hoc Comparison between Groups for Accuracy

Tukey Pairwise Comparisons

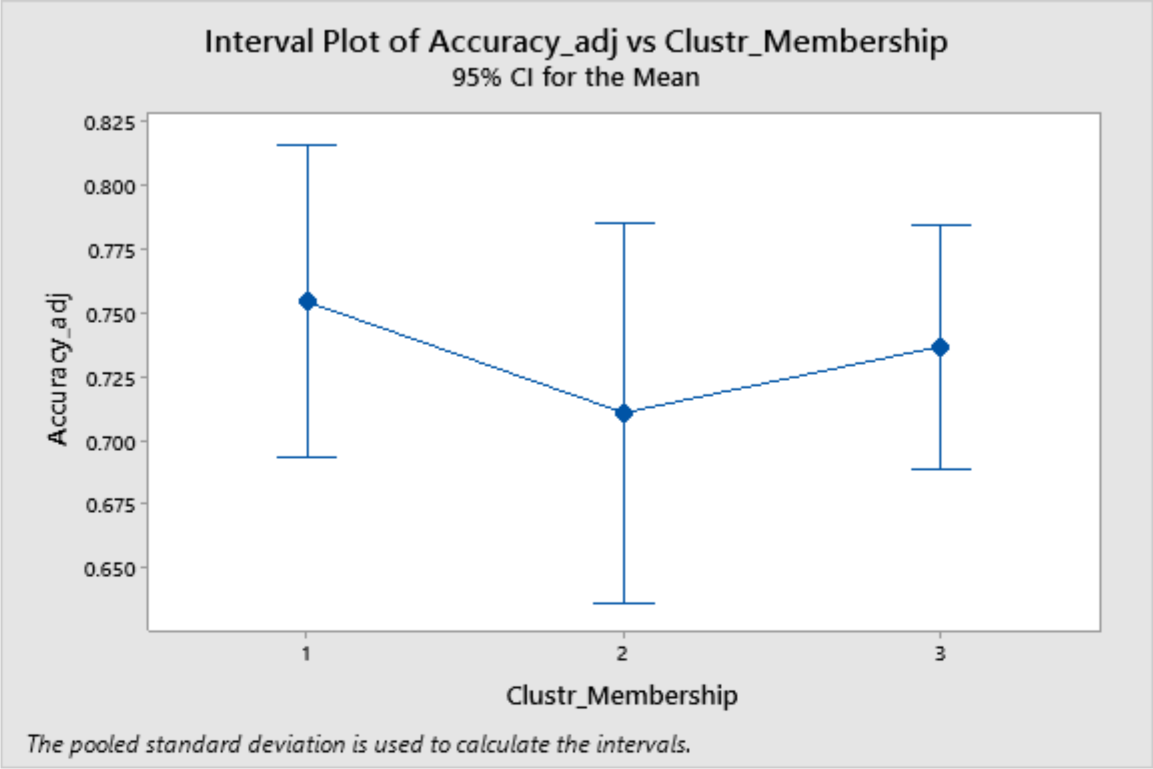
Grouping Information Using the Tukey Method and 95% Confidence

Cluster_Membership	N	Mean	Grouping
1	83	0.7542	A
3	136	0.7368	A
2	56	0.7107	A

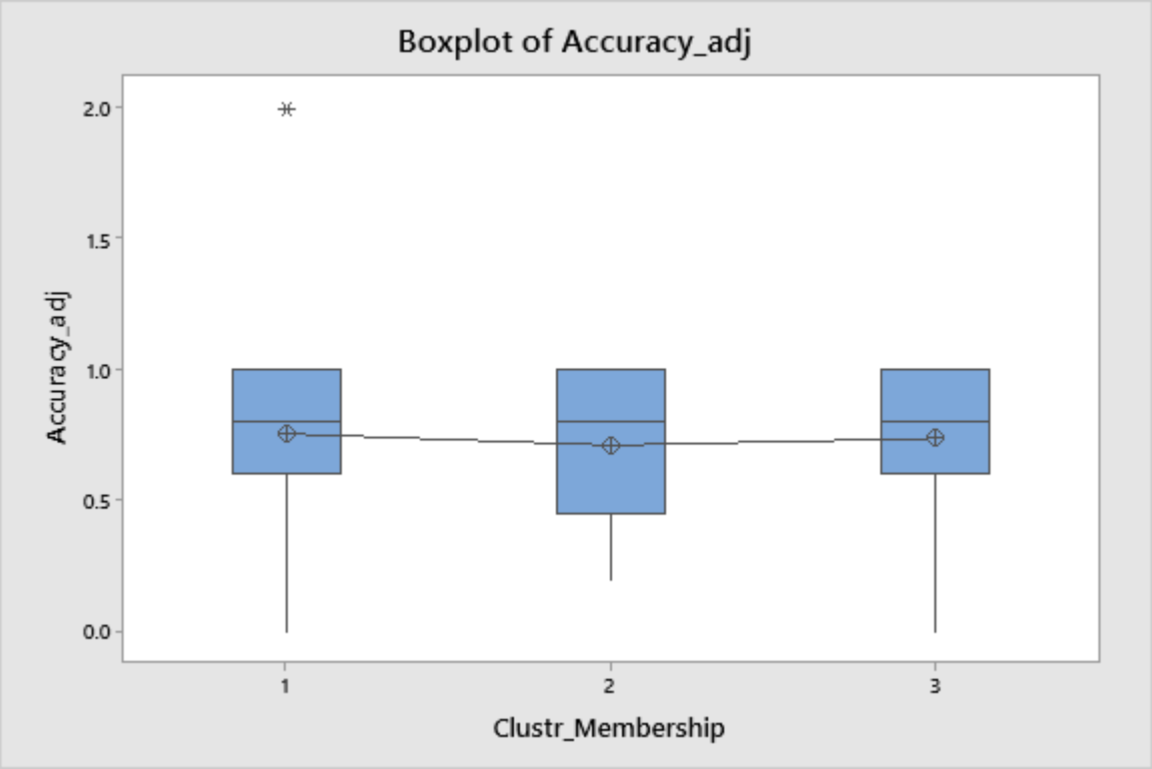
Means that do not share a letter are significantly different.



If an interval does not contain zero, the corresponding means are significantly different.



The pooled standard deviation is used to calculate the intervals.



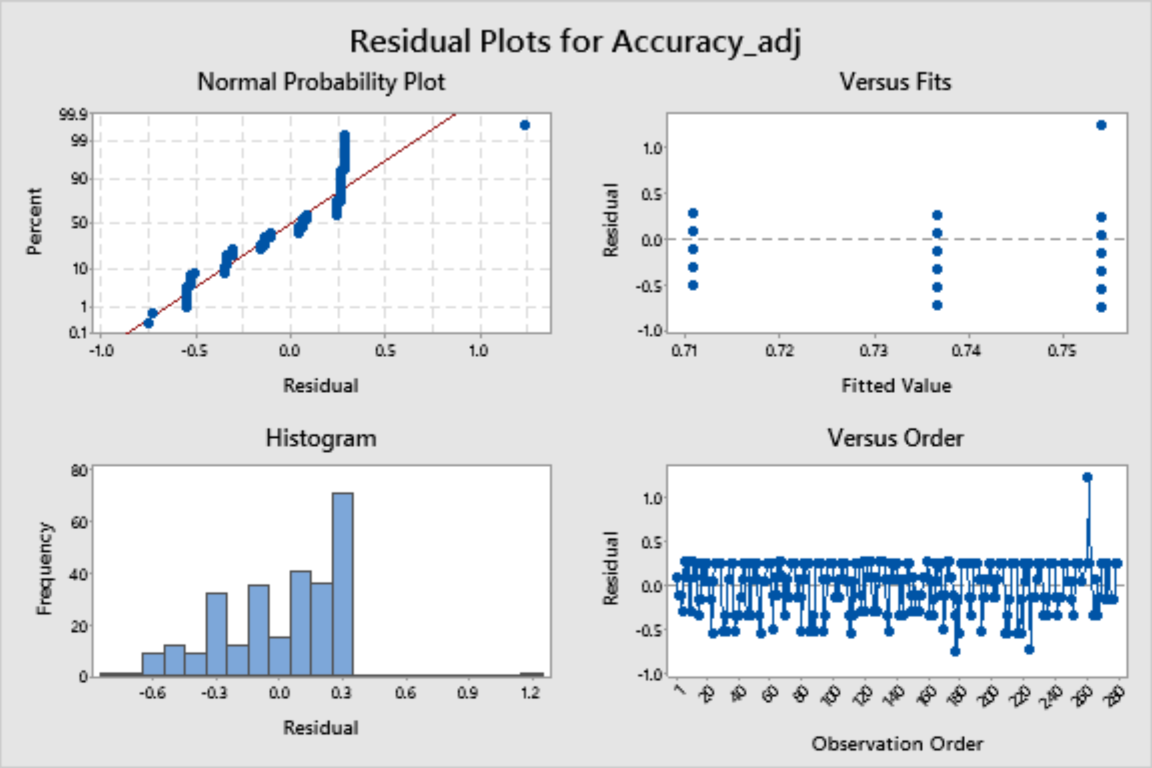
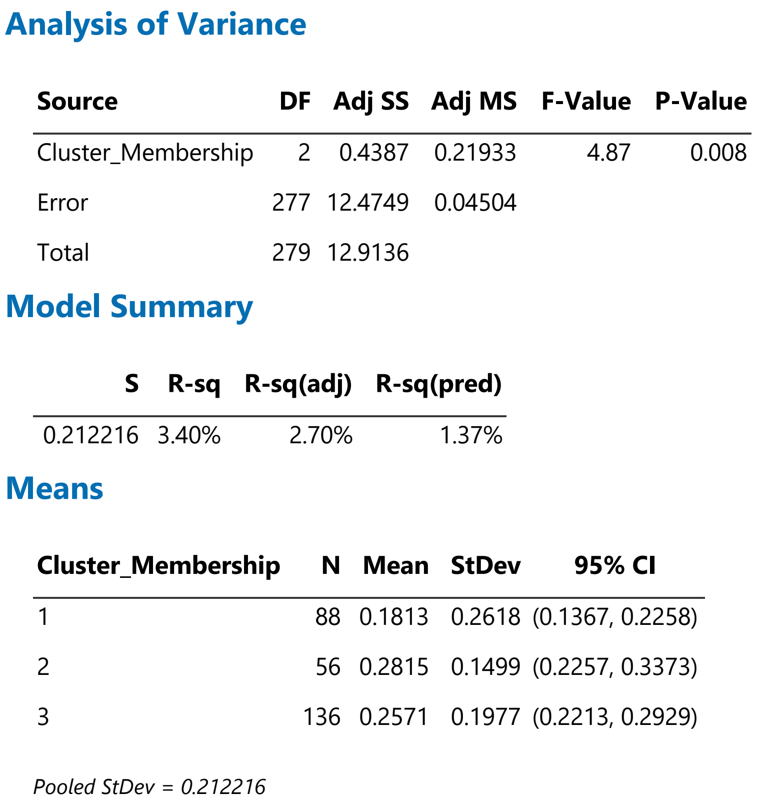


Figure E.2 Residual Plots of Accuracy for Examining the Normality Assumptions

Table E.6 One-Way ANOVA of Individual Difference for c



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Cluster_Membership	2	0.4387	0.21933	4.87	0.008
Error	277	12.4749	0.04504		
Total	279	12.9136			

Model Summary

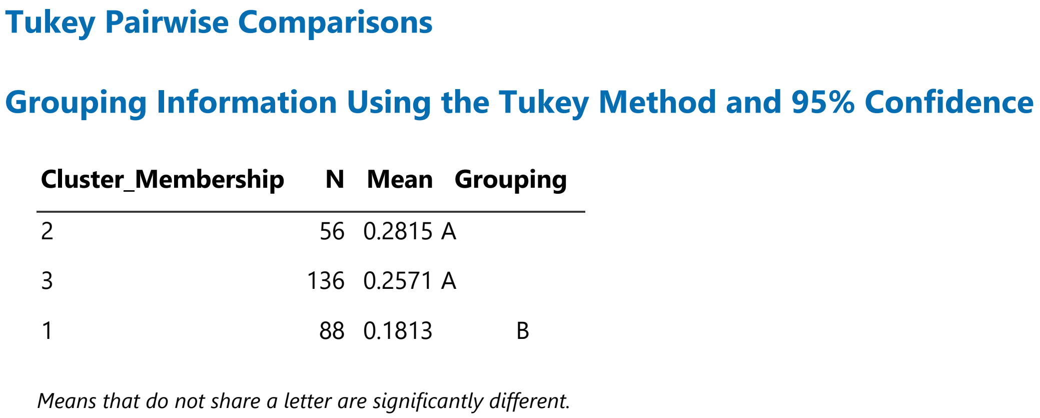
S	R-sq	R-sq(adj)	R-sq(pred)
0.212216	3.40%	2.70%	1.37%

Means

Cluster_Membership	N	Mean	StDev	95% CI
1	88	0.1813	0.2618	(0.1367, 0.2258)
2	56	0.2815	0.1499	(0.2257, 0.3373)
3	136	0.2571	0.1977	(0.2213, 0.2929)

Pooled StDev = 0.212216

Table E.7 Tukey Post-Hoc Comparison between Groups for c

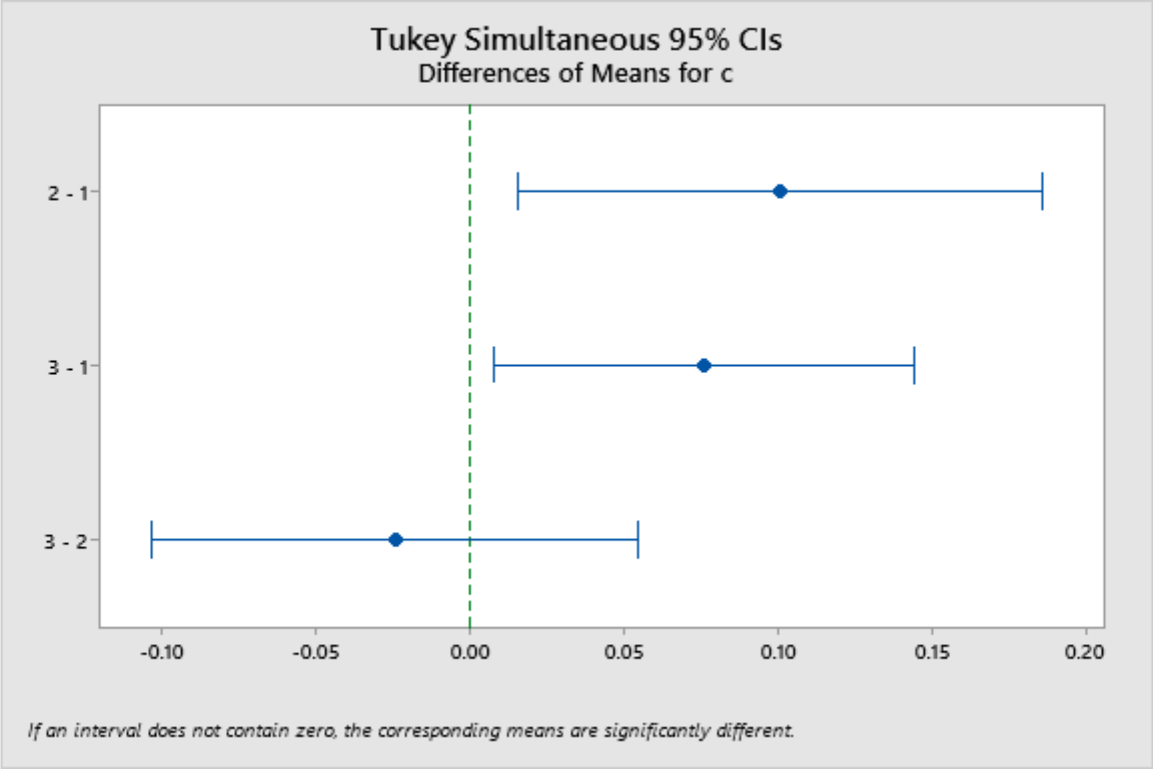


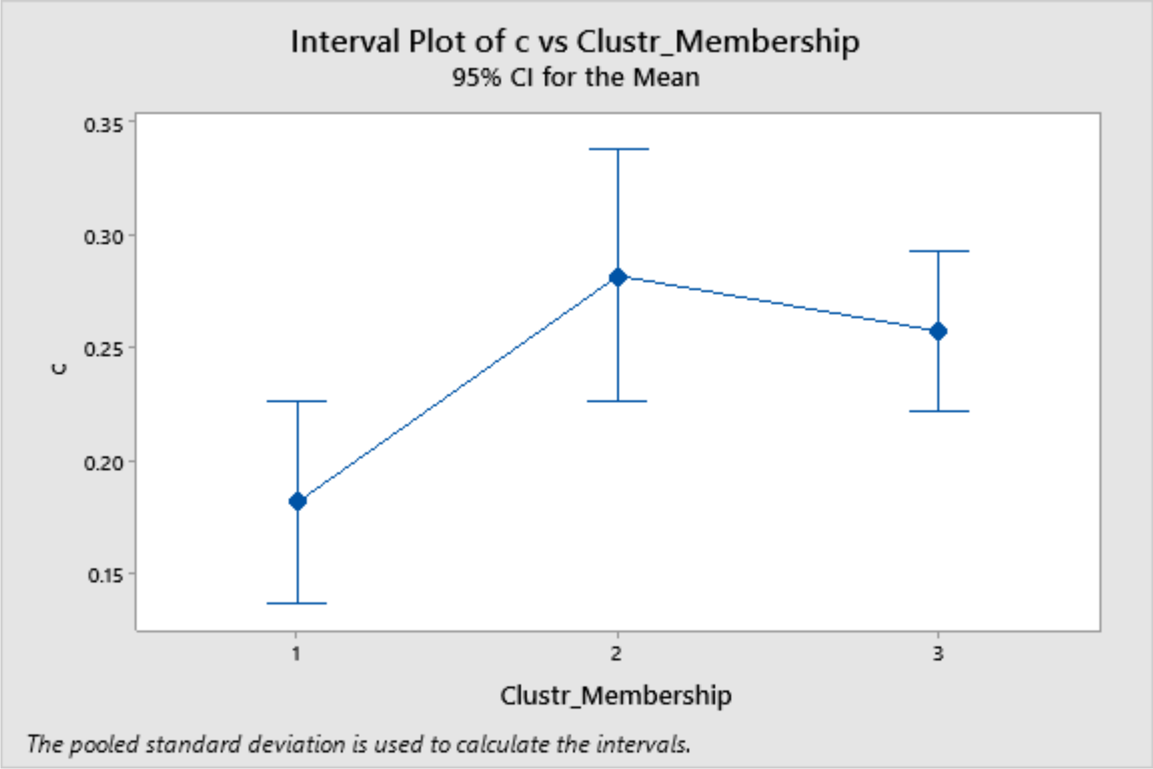
Tukey Pairwise Comparisons

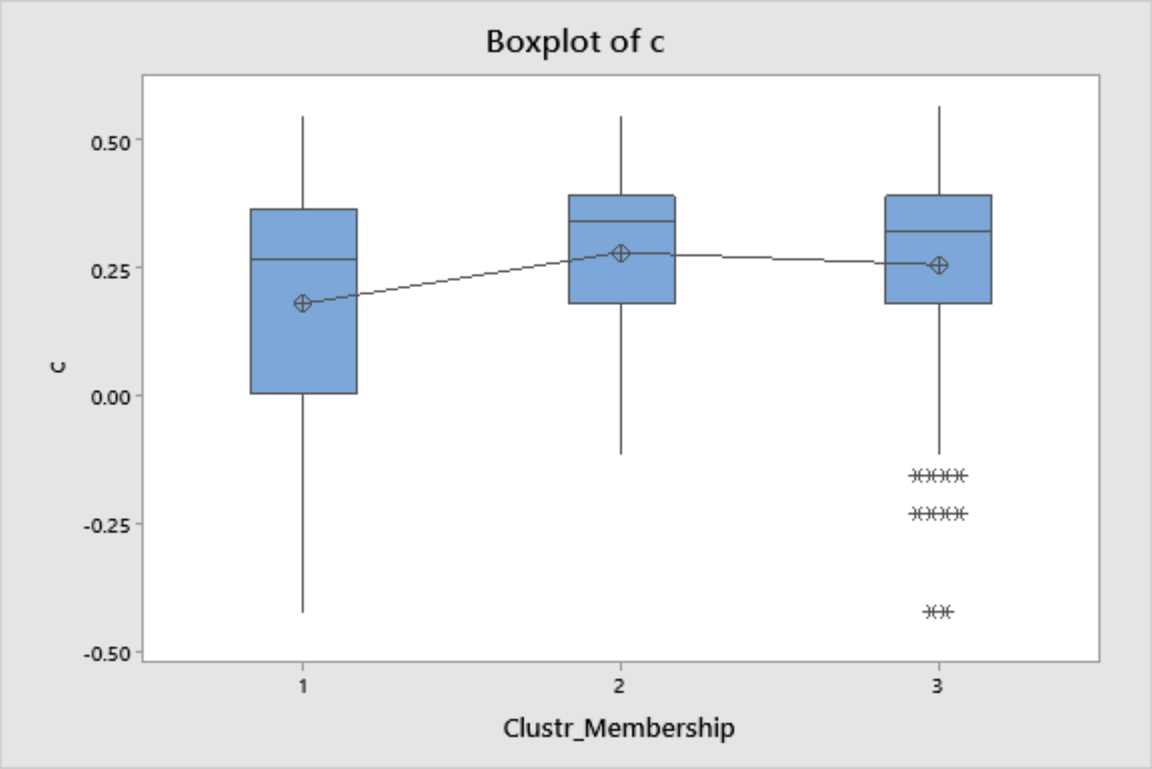
Grouping Information Using the Tukey Method and 95% Confidence

Cluster_Membership	N	Mean	Grouping	
2	56	0.2815	A	
3	136	0.2571	A	
1	88	0.1813		B

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for c
2 - 1
3 - 1
3 - 2
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of c vs Clustr_Membership
95% CI for the Mean
0.35
0.30
0.25
0.20
0.15
c
1
2
3
Clustr_Membership
The pooled standard deviation is used to calculate the intervals.



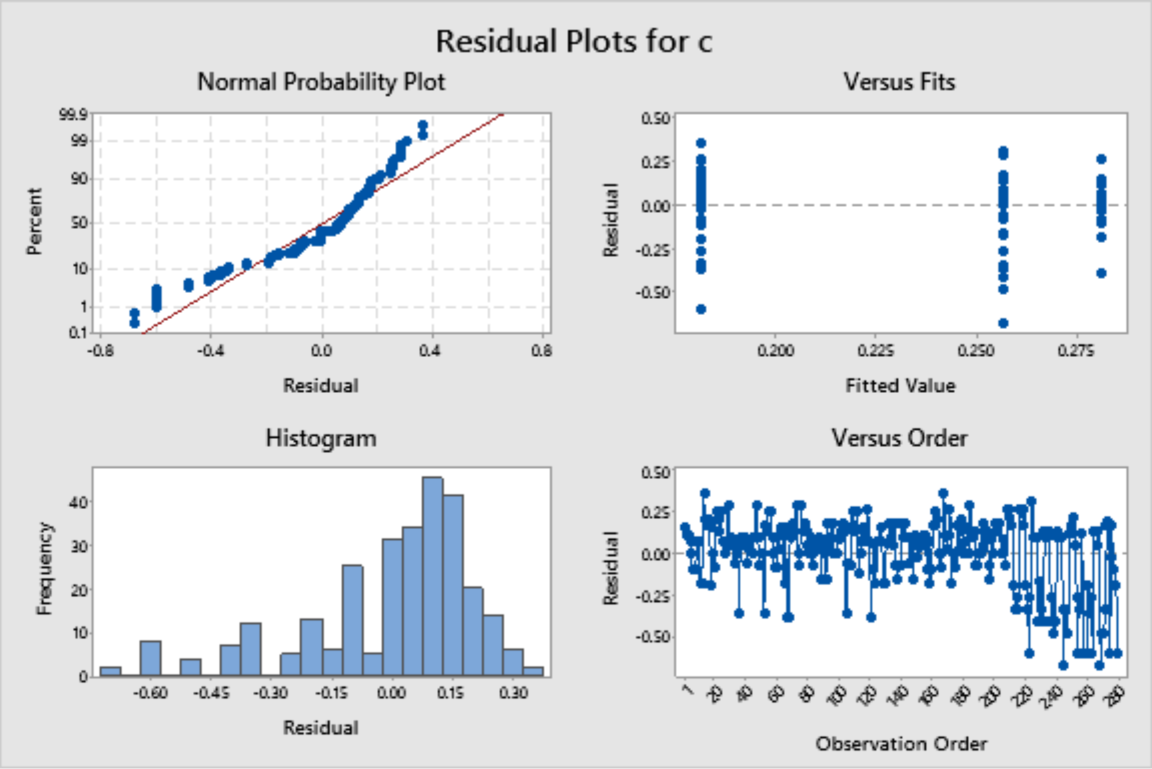


Figure E.3 Residual Plots of c for Examining the Normality Assumptions

Table E.8 One-Way ANOVA of Individual Difference for Time of Response

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Cluster_Membership	2	29005	14503	8.52	0.000
Error	245	416934	1702		
Total	247	445939			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
41.2525	6.50%	5.74%	4.09%

Means

Cluster_Membership	N	Mean	StDev	95% CI
1	76	44.10	31.15	(34.78, 53.42)
2	52	71.79	50.89	(60.52, 83.06)
3	120	46.06	42.22	(38.65, 53.48)

Pooled StDev = 41.2525

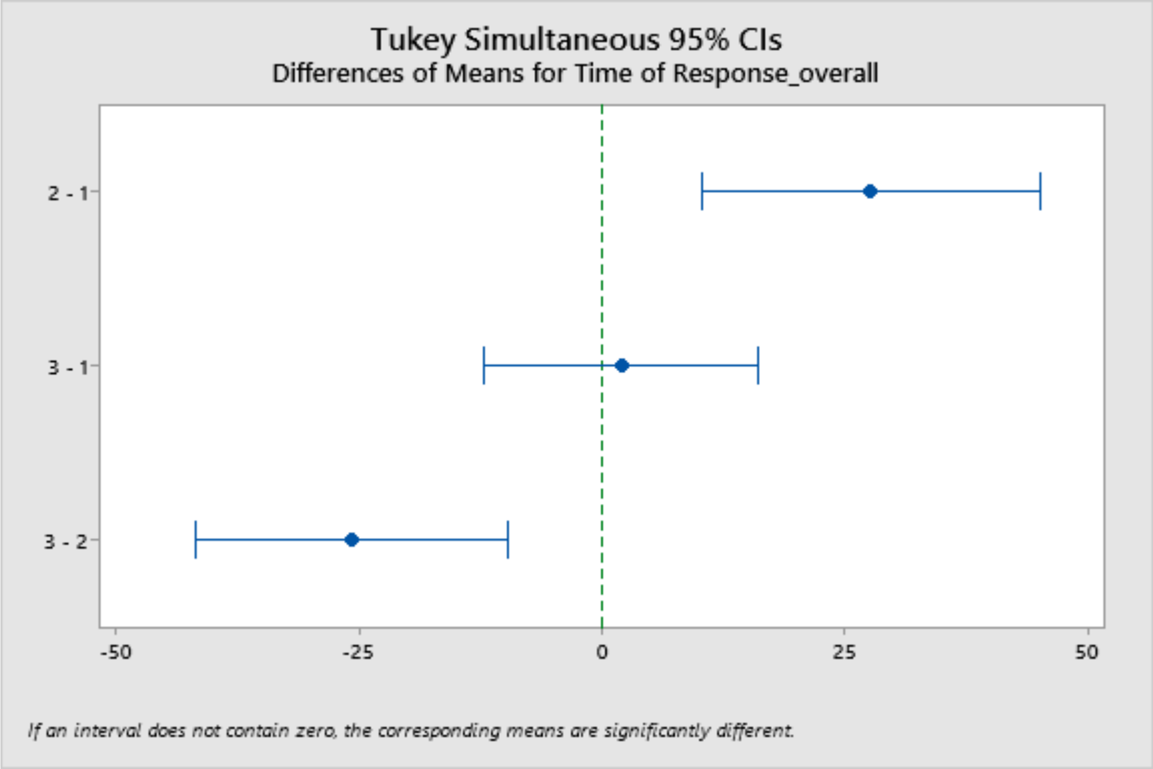
Table E.9 Tukey Post-Hoc Comparison between Groups for Time of Response

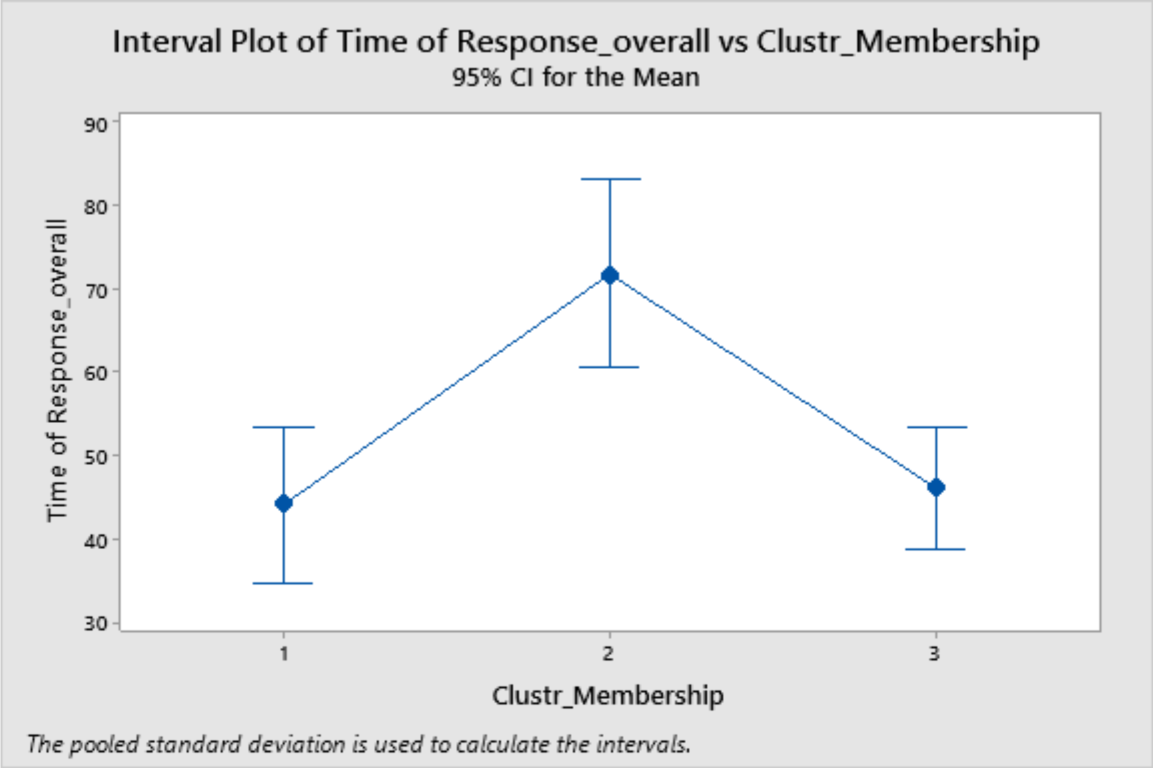
Tukey Pairwise Comparisons

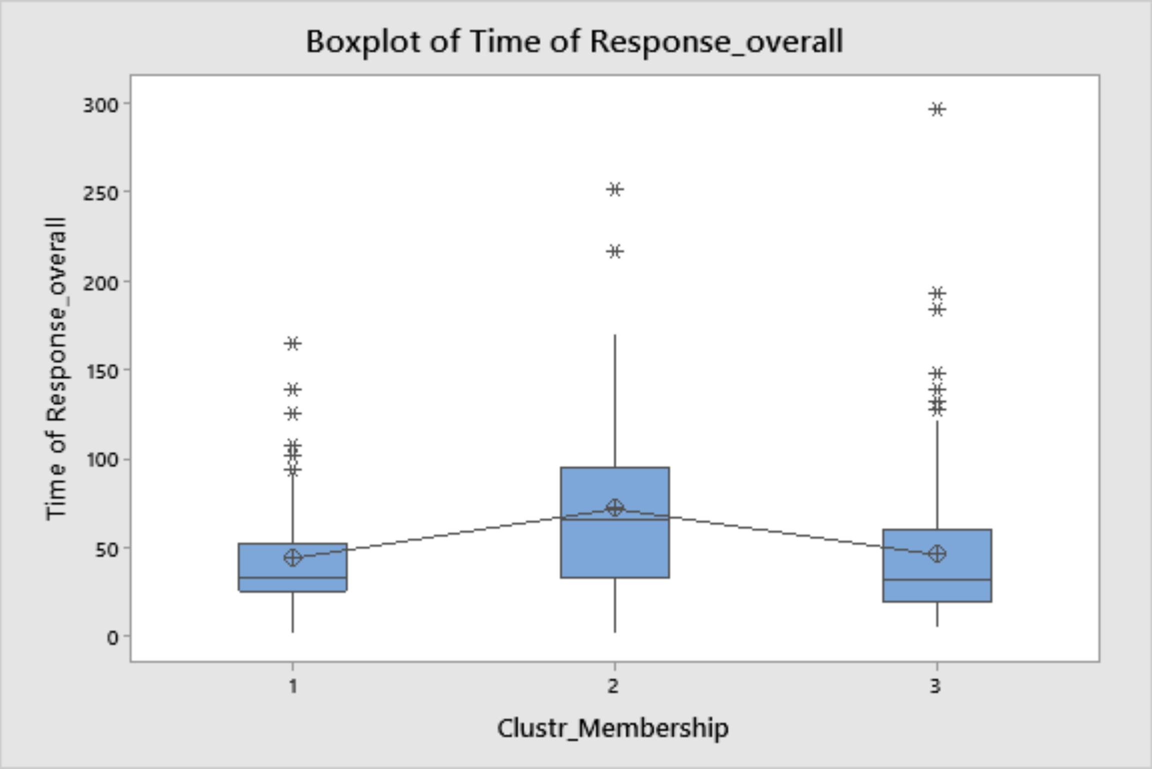
Grouping Information Using the Tukey Method and 95% Confidence

Cluster_Membership	N	Mean	Grouping	
2	52	71.79	A	
3	120	46.06		B
1	76	44.10		B

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for Time of Response_overall
2 - 1
3 - 1
3 - 2
-50
-25
0
25
50
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of Time of Response_overall vs Clustr_Membership
95% CI for the Mean
Time of Response_overall
90
80
70
60
50
40
30
1
2
3
Clustr_Membership
The pooled standard deviation is used to calculate the intervals.



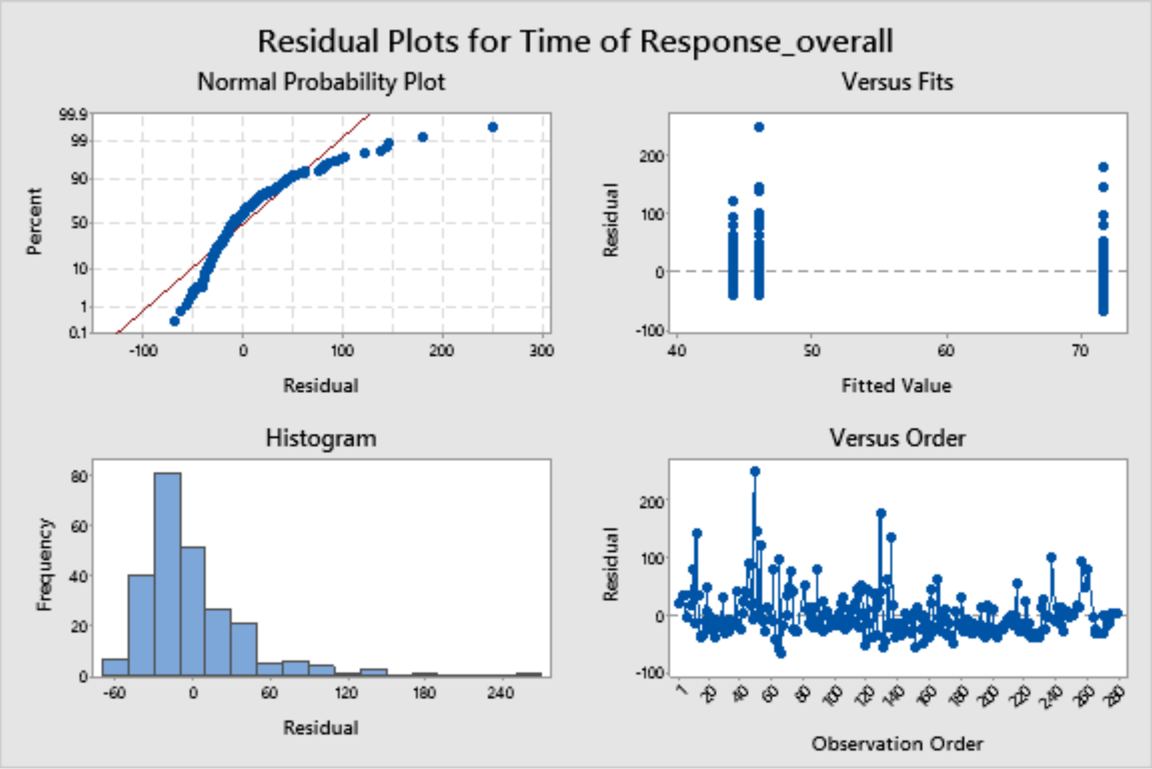


Figure E.4 Residual Plots of Time of Response for Examining the Normality Assumptions

Table E.10 One-Way ANOVA of Individual Difference for NASA-TLX

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Cluster_Membership	2	2733	1366.5	4.39	0.013
Error	273	84945	311.2		
Total	275	87678			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
17.6395	3.12%	2.41%	1.01%

Means

Cluster_Membership	N	Mean	StDev	95% CI
1	84	26.96	15.55	(23.17, 30.75)
2	56	33.85	18.16	(29.21, 38.49)
3	136	25.65	18.60	(22.67, 28.63)

Pooled StDev = 17.6395

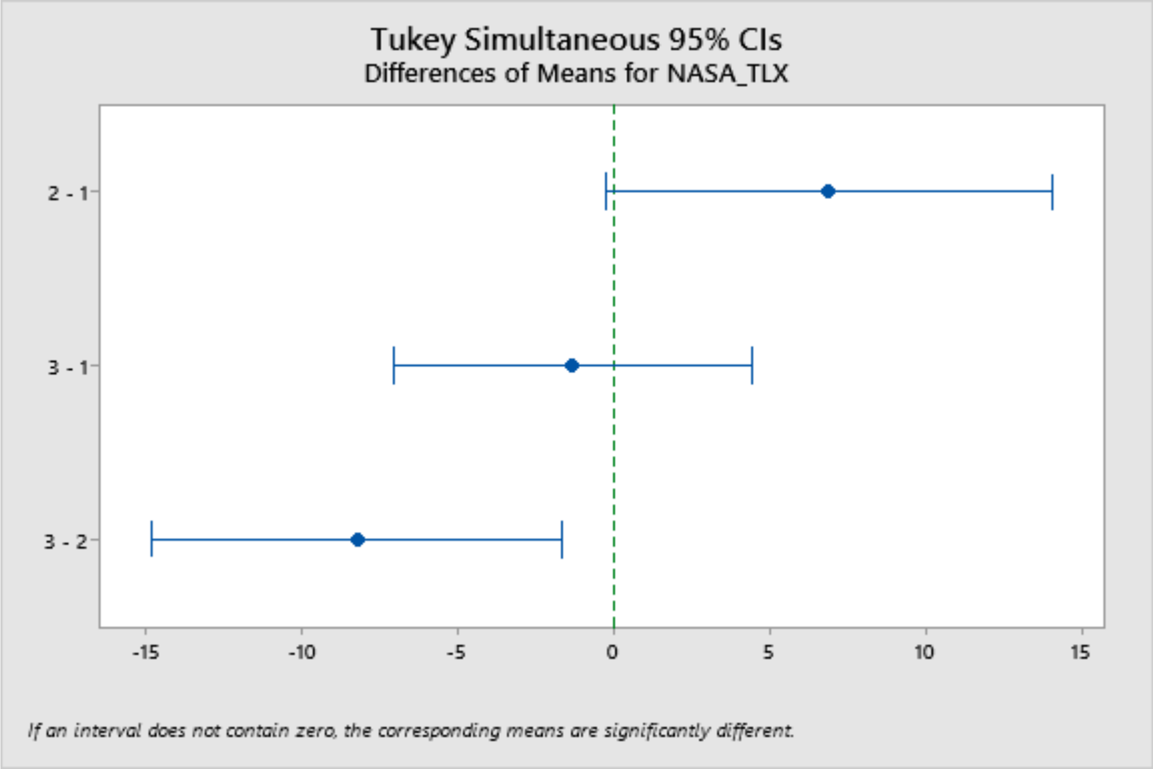
Table E.11 Tukey Post-Hoc Comparison between Groups for NASA-TLX

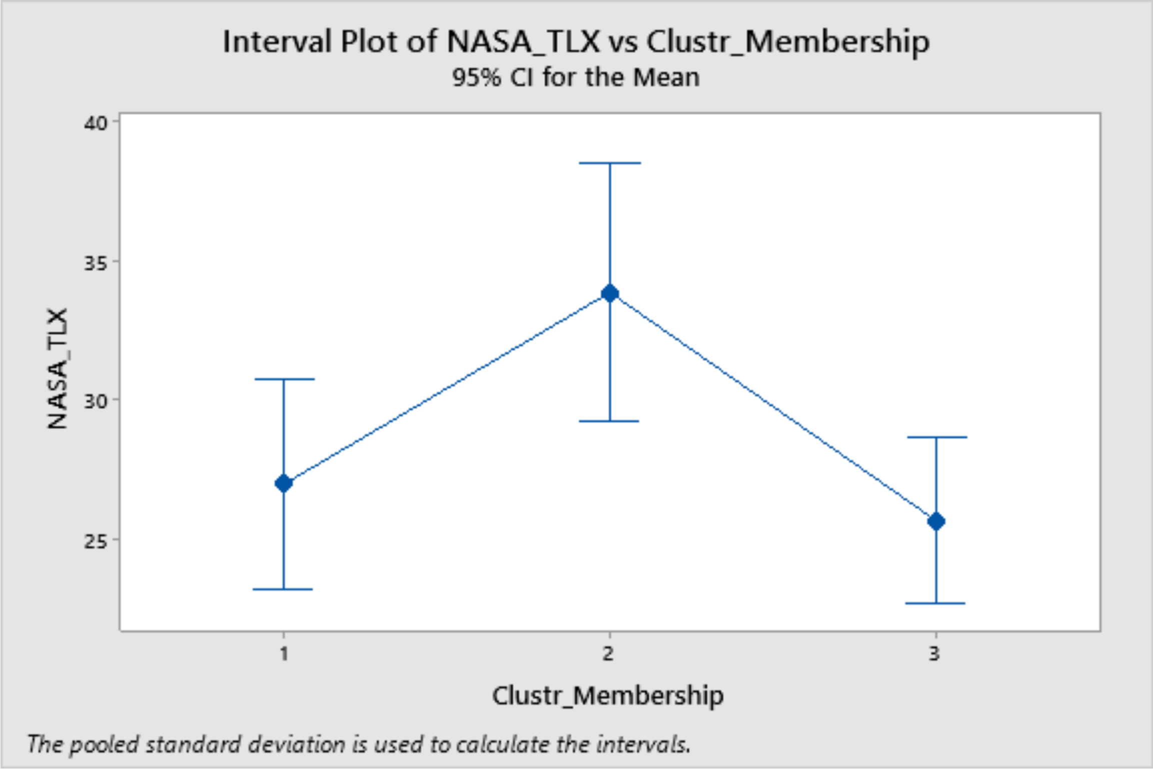
Tukey Pairwise Comparisons

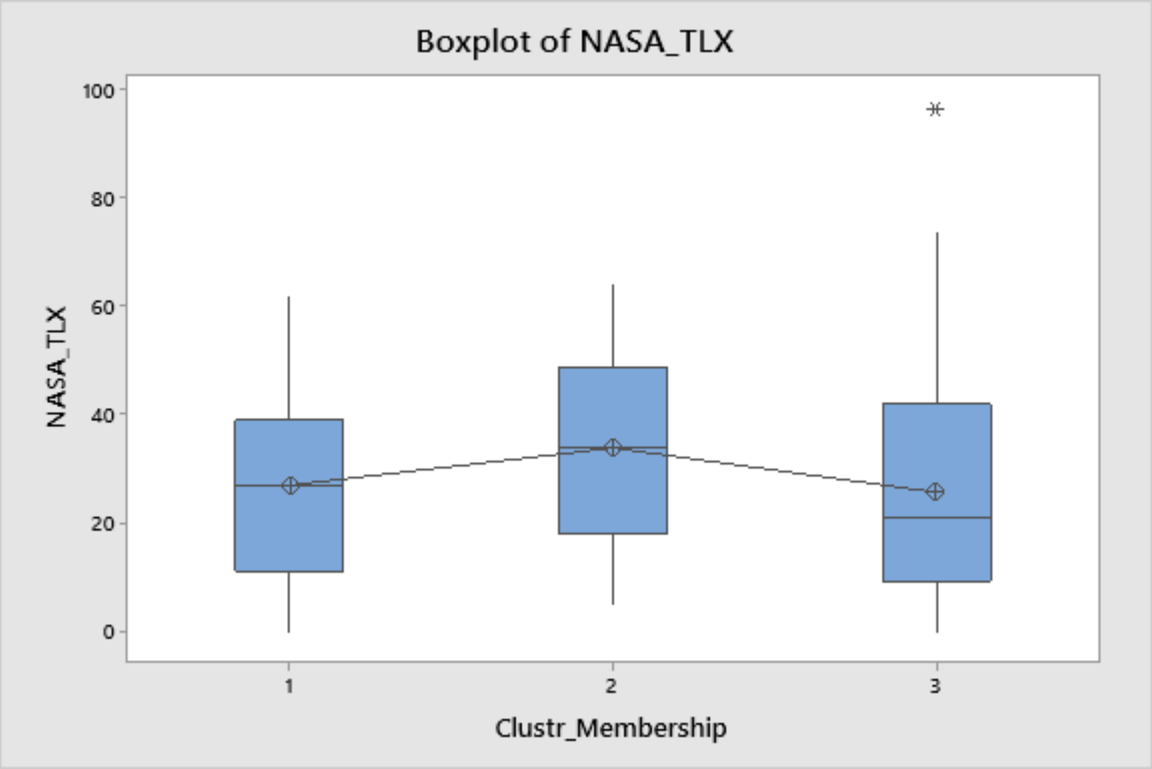
Grouping Information Using the Tukey Method and 95% Confidence

Cluster_Membership	N	Mean	Grouping	
2	56	33.85	A	
1	84	26.96	A	B
3	136	25.65		B

Means that do not share a letter are significantly different.


Tukey Simultaneous 95% CIs
Differences of Means for NASA_TLX
2 - 1
3 - 1
3 - 2
-15
-10
-5
0
5
10
15
If an interval does not contain zero, the corresponding means are significantly different.


Interval Plot of NASA_TLX vs Clustr_Membership
95% CI for the Mean
NASA_TLX
40
35
30
25
1
2
3
Clustr_Membership
The pooled standard deviation is used to calculate the intervals.



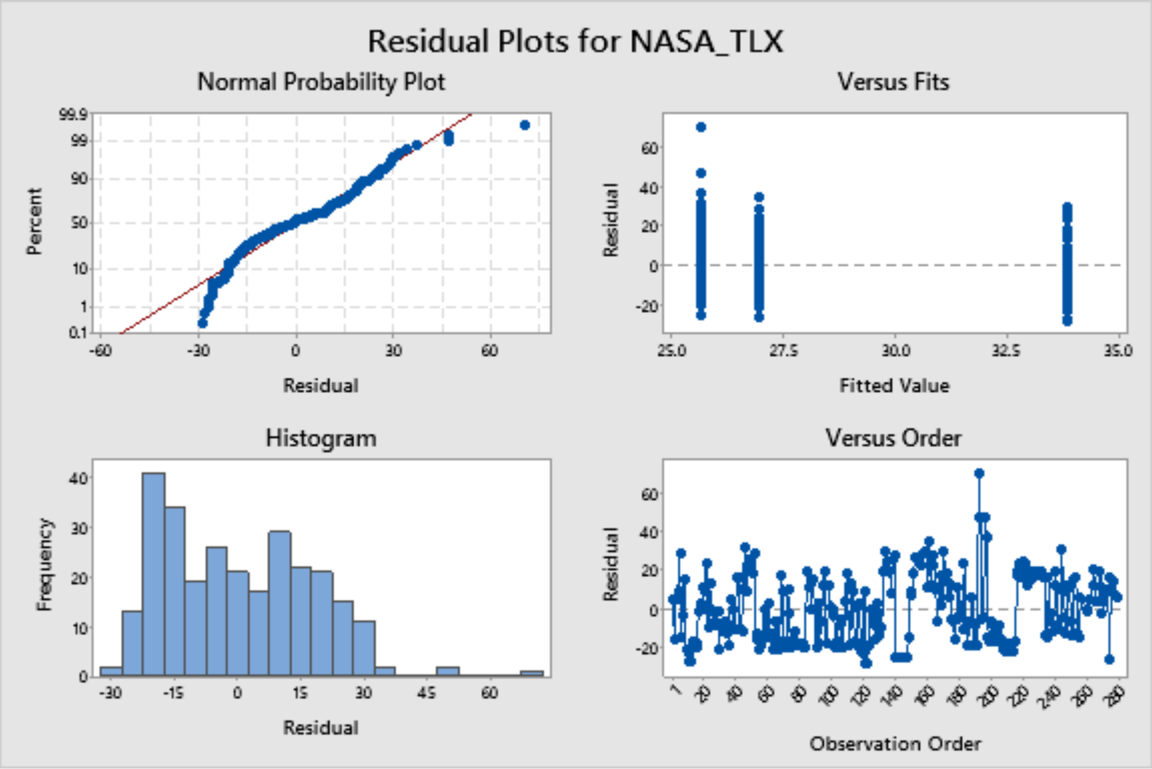


Figure E.5 Residual Plots of NASA-TLX for Examining the Normality Assumptions

APPENDIX F. ANOVA TABLES FOR CONFUSION MATRIX METRICS

Appendix Outputs Corresponding to Table 10. Confusion matrix metrics. (8.1.1)

General Linear Model: balanced_accuracy versus Decision Paradigm, Nutrition Information Format, System Default, Name

Table F.1 ANOVA table for GLM of balanced accuracy

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	1.7655	1.76550	98.22	0.000
Nutrition Information Format (Nutri-scores = 1)	1	0.0020	0.00201	0.11	0.739
System Default (Pre-selection = 1)	1	0.0015	0.00153	0.08	0.771
Name	35	0.6363	0.01818	1.01	0.457
2AFC*Nutri-Scores	1	0.0896	0.08958	4.98	0.027
2AFC*pre-selection	1	0.0112	0.01121	0.62	0.430
Nutri-scores*pre-selection	1	0.0000	0.00003	0.00	0.969
2AFC*Nutri-Scores*pre-selection	1	0.0383	0.03826	2.13	0.146
Error	237	4.2602	0.01798		
Lack-of-Fit	236	4.2602	0.01805	13215.05	0.007
Pure Error	1	0.0000	0.00000		
Total	279	14.6246			

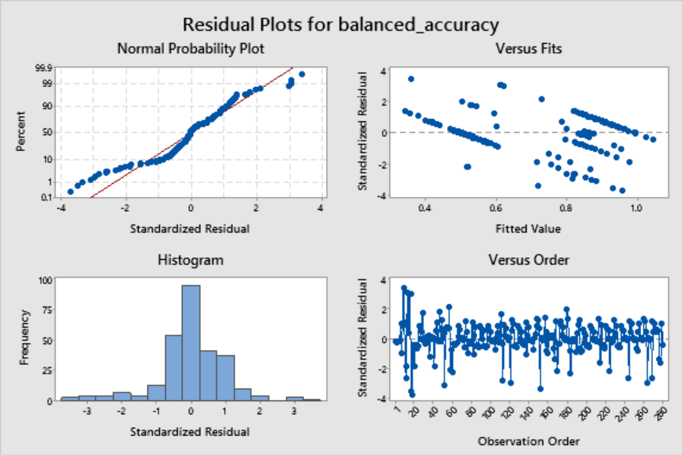


Figure F.1 Standardized Residual Plots for balanced accuracy

General Linear Model: Recall versus yes/no_2AFC, Nutrition Information Format, System Default, Name

Table F.2 ANOVA table for GLM of recall

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	0.3638	0.363812	10.73	0.001
Nutrition Information Format (Nutri-scores = 1)	1	0.0041	0.004143	0.12	0.727
System Default (Pre-selection = 1)	1	0.0133	0.013254	0.39	0.532
Name	35	1.2494	0.035697	1.05	0.395
2AFC*Nutri-Scores	1	0.1239	0.123904	3.65	0.057
2AFC*pre-selection	1	0.0728	0.072774	2.15	0.144
Nutri-scores*pre-selection	1	0.0177	0.017693	0.52	0.471
2AFC*Nutri-Scores*pre-selection	1	0.1533	0.153310	4.52	0.035
Error	237	8.0360	0.033907		
Lack-of-Fit	236	8.0360	0.034051	24927.71	0.005
Pure Error	1	0.0000	0.000001		
Total	279	10.0516			

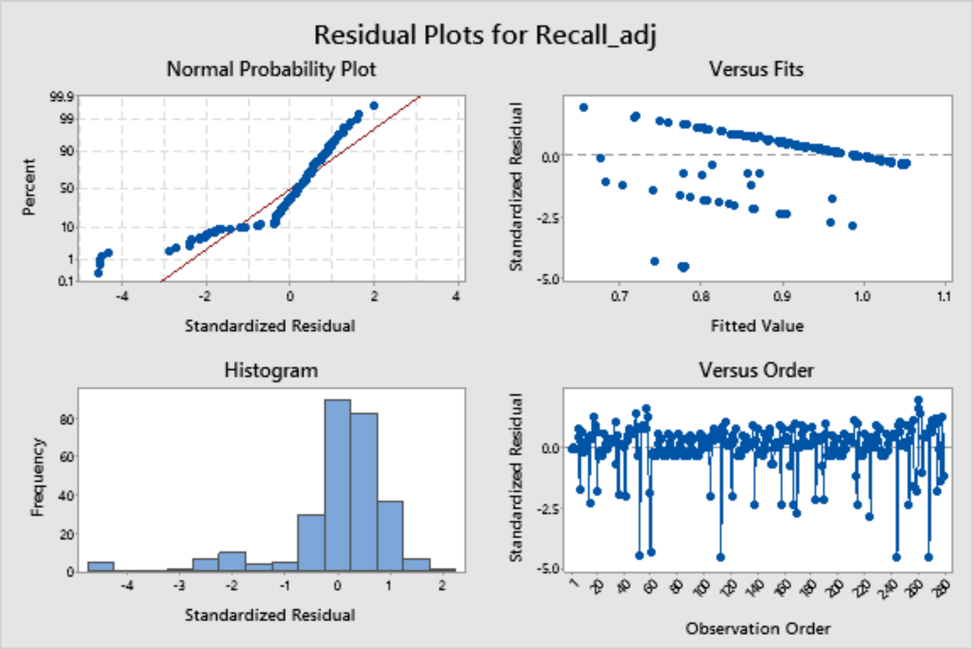


Figure F.2 Standardized Residual Plots for recall

General Linear Model: Precision versus yes/no_2AFC, Nutrition Information Format, System Default, Name

Table F.3 ANOVA table for GLM of precision

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	0.9890	0.988984	19.87	0.000
Nutrition Information Format (Nutri-scores = 1)	1	0.0011	0.001076	0.02	0.883
System Default (Pre-selection = 1)	1	0.3194	0.319394	6.42	0.012
Name	35	1.3032	0.037233	0.75	0.848
2AFC*Nutri-Scores	1	0.1384	0.138380	2.78	0.097
2AFC*pre-selection	1	0.2562	0.256221	5.15	0.024
Nutri-scores*pre-selection	1	0.0879	0.087860	1.77	0.185
2AFC*Nutri-Scores*pre-selection	1	0.1578	0.157759	3.17	0.076
Error	237	11.7933	0.049761		
Lack-of-Fit	236	11.7733	0.049887	2.49	0.473
Pure Error	1	0.0200	0.020000		
Total	279	22.2254			

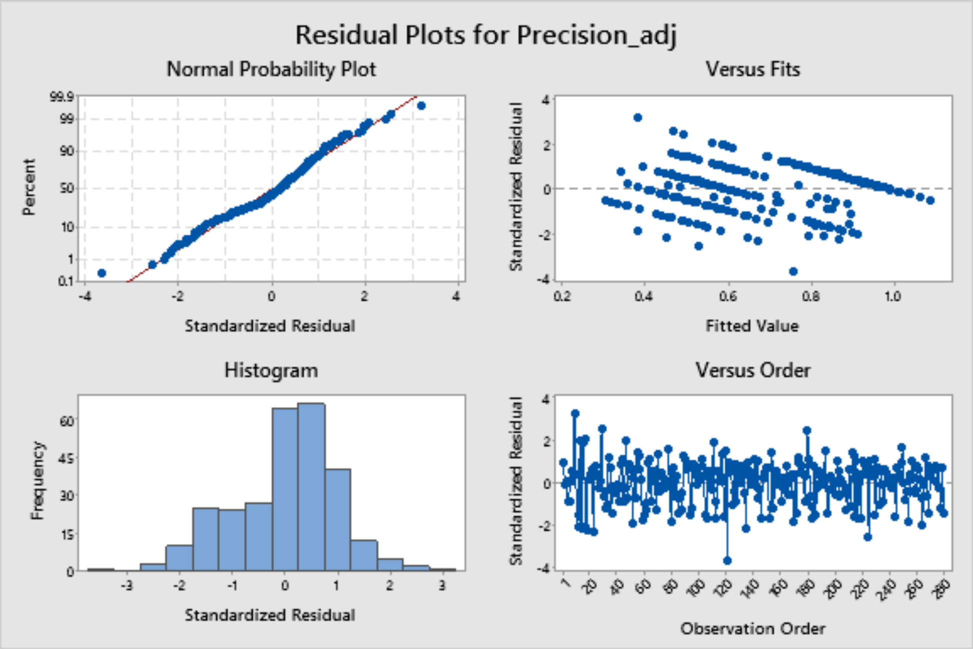


Figure F.3 Standardized Residual Plots for precision

General Linear Model: F1-Scores versus yes/no_2AFC, Nutrition Information Format, System Default, Name

Table F.4 ANOVA table for GLM of F1-scores

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	0.1386	0.138612	2.65	0.105
Nutrition Information Format (Nutri-scores = 1)	1	0.0004	0.000430	0.01	0.928
System Default (Pre-selection = 1)	1	0.2981	0.298129	5.69	0.018
Name	35	1.3703	0.039151	0.75	0.848
2AFC*Nutri-Scores	1	0.1880	0.187960	3.59	0.059
2AFC*pre-selection	1	0.3277	0.327726	6.26	0.013
Nutri-scores*pre-selection	1	0.1005	0.100460	1.92	0.167
2AFC*Nutri-Scores*pre-selection	1	0.2658	0.265798	5.08	0.025
Error	237	12.4074	0.052352		
Lack-of-Fit	236	12.3914	0.052506	3.30	0.418
Pure Error	1	0.0159	0.015922		
Total	279	17.4646			

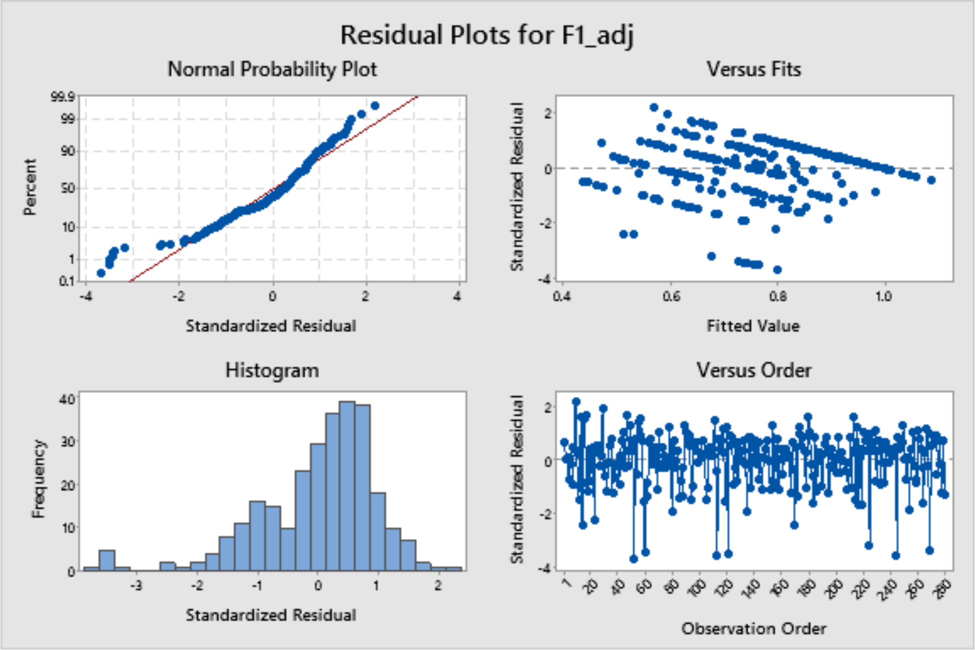


Figure F.4 Standardized Residual Plots for F1-scores

General Linear Model: Informedness versus yes/no_2AFC, Nutrition Information Format, System Default, Name

Table F.5 ANOVA table for GLM of Informedness

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	0.6061	0.60612	4.50	0.035
Nutrition Information Format (Nutri-scores = 1)	1	0.0123	0.01227	0.09	0.763
System Default (Pre-selection = 1)	1	1.2833	1.28332	9.54	0.002
Name	35	3.8436	0.10982	0.82	0.761
2AFC*Nutri-Scores	1	0.2958	0.29580	2.20	0.140
2AFC*pre-selection	1	0.6683	0.66833	4.97	0.027
Nutri-scores*pre-selection	1	0.4257	0.42570	3.16	0.077
2AFC*Nutri-Scores*pre-selection	1	0.4810	0.48099	3.57	0.060
Error	237	31.8924	0.13457		
Lack-of-Fit	236	31.8124	0.13480	1.68	0.558
Pure Error	1	0.0800	0.08000		
Total	279	46.0124			

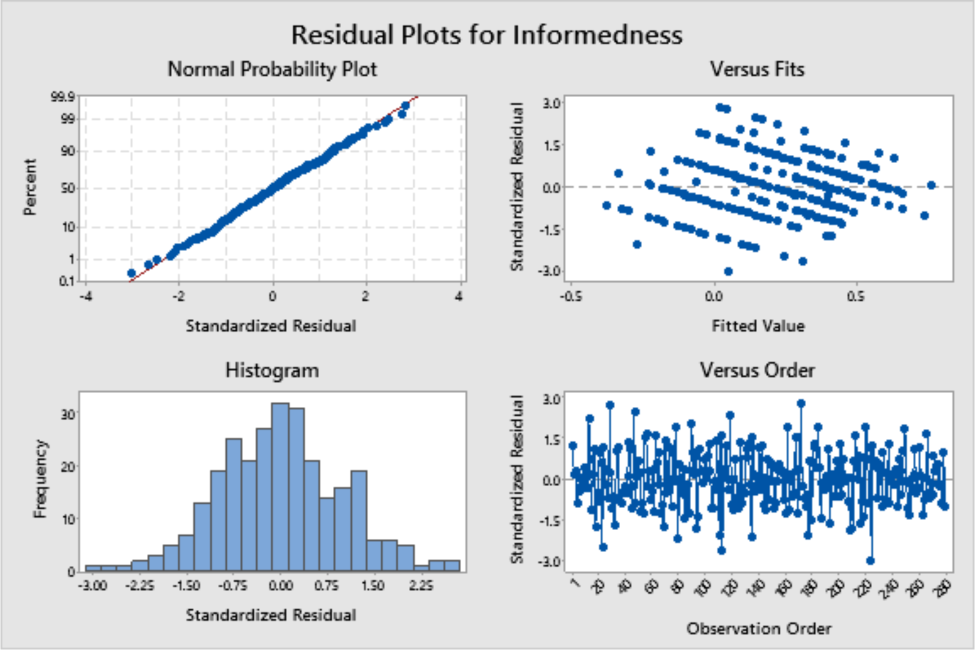


Figure F.5 Standardized Residual Plots for Informedness

General Linear Model: Markedness versus yes/no_2AFC, Nutrition Information Format, System Default, Name

Table F.6 ANOVA table for GLM of Markedness

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Decision Paradigm (2AFC = 1)	1	6.5396	6.53957	78.88	0.000
Nutrition Information Format (Nutri-scores = 1)	1	0.0035	0.00346	0.04	0.838
System Default (Pre-selection = 1)	1	0.2485	0.24849	3.00	0.085
Name	35	3.0065	0.08590	1.04	0.420
2AFC*Nutri-Scores	1	0.3260	0.32600	3.93	0.049
2AFC*pre-selection	1	0.2048	0.20479	2.47	0.117
Nutri-scores*pre-selection	1	0.0669	0.06691	0.81	0.370
2AFC*Nutri-Scores*pre-selection	1	0.2407	0.24071	2.90	0.090
Error	237	19.6482	0.08290		
Lack-of-Fit	236	19.6282	0.08317	4.16	0.376
Pure Error	1	0.0200	0.02000		
Total	279	61.9379			

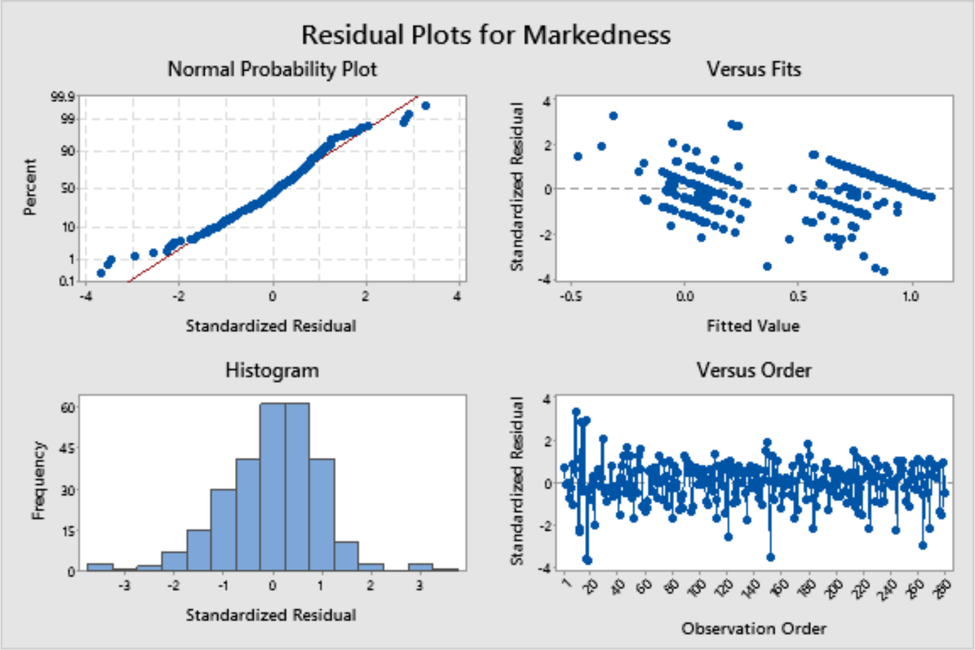


Figure F.6 Standardized Residual Plots for Markedness

REFERENCES

Abdi, J., Al-Hindawi, A., Ng, T., & Vizcaychipi, M. P. (2018). Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open*, *8*(2), 18815. https://doi.org/10.1136/bmjopen-2017-018815

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *27*(3), 379–387. https://doi.org/10.1037/0278-6133.27.3.379

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal*, *11*(4), 325–338. https://doi.org/10.1023/A:1025869312943

Agapito, G., Simeoni, M., Calabrese, B., Caré, I., Lamprinoudi, T., Guzzi, P. H., … Cannataro, M. (2018). DIETOS: A dietary recommender system for chronic diseases monitoring and management. *Computer Methods and Programs in Biomedicine*, *153*, 93–104. https://doi.org/10.1016/j.cmpb.2017.10.014

Ali, E. E., Chew, L., & Yap, K. Y.-L. (2016). Evolution and current status of mhealth research: A systematic review. *BMJ Innovations*, *2*(1), 33–40. https://doi.org/10.1136/bmjinnov-2015-000096

Arning, K., & Ziefle, M. (2009). Effects of age, cognitive, and personal factors on PDA menu navigation performance. *Behaviour & Information Technology*, *28*(3), 251–268. https://doi.org/10.1080/01449290701679395

Arno, A., & Thomas, S. (2016). The efficacy of nudge theory strategies in influencing adult dietary behaviour: A systematic review and meta-analysis. *BMC Public Health*, *16*(1), 676. https://doi.org/10.1186/s12889-016-3272-x

Astell, A. J., Hwang, F., Brown, L. J. E., Timon, C., Maclean, L. M., Smith, T., … Williams, E. A. (2014). Validation of the NANA (Novel Assessment of Nutrition and Ageing) touch screen system for use at home by older adults. *Experimental Gerontology*, *60*, 100–107. https://doi.org/10.1016/j.exger.2014.10.008

Athilingam, P., Labrador, M. A., Remo, E. F. J., Mack, L., San Juan, A. B., & Elliott, A. F. (2016). Features and usability assessment of a patient-centered mobile application (HeartMapp) for self-management of heart failure. *Applied Nursing Research*, *32*, 156–163. https://doi.org/10.1016/j.apnr.2016.07.001

Aula, A. (2005). User study on older adults' use of the Web and search engines. *Universal Access in the Information Society*, *4*(1), 67–81. https://doi.org/10.1007/s10209-004-0097-7

Azar, K. M. J., Lesser, L. I., Laing, B. Y., Stephens, J., Aurora, M. S., Burke, L. E., & Palaniappan, L. P. (2013). Mobile applications for weight management: Theory-based content analysis. *American Journal of Preventive Medicine*, *45*(5), 583–589. https://doi.org/10.1016/j.amepre.2013.07.005

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual Review of Psychology*, *52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Bardus, M., van Beurden, S. B., Smith, J. R., & Abraham, C. (2016). A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1). https://doi.org/10.1186/s12966-016-0359-9

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, Vol. 13, pp. 407–424. https://doi.org/10.1108/eb024320

Bauer, J. M., & Reisch, L. A. (2018). Behavioural Insights and (Un)healthy Dietary Choices: a Review of Current Evidence. *Journal of Consumer Policy*. https://doi.org/10.1007/s10603-018-9387-y

Baumgartl, H., Sauter, D., Roessler, P., & Buettner, R. (2020). Measuring social desirability using a novel machine learning approach based on EEG data. *Proceedings of the 24th Pacific Asia Conference on Information Systems: Information Systems (IS) for the Future, PACIS 2020*. Retrieved from https://www.researchgate.net/publication/341293453_Measuring_Social_Desirability_Using_a_Novel_Machine_Learning_Approach_Based_on_EEG_Data

Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., & Nürnberger, A. (2013). Research paper recommender system evaluation: A quantitative literature survey. *ACM International Conference Proceeding Series*, 15–22. https://doi.org/10.1145/2532508.2532512

Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G. D., & Essa, I. (2015). Leveraging context to support automated food recognition in restaurants. *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, 580–587. https://doi.org/10.1109/WACV.2015.83

Bidmon, S., Terlutter, R., & Röttl, J. (2014). What explains usage of mobile physician-rating apps? Results from a web-based questionnaire. *Journal of Medical Internet Research*, *16*(6), e148. https://doi.org/10.2196/jmir.3122

Binstock, R. H. (1985). The oldest old: a fresh perspective or compassionate ageism revisited? *The Milbank Memorial Fund Quarterly. Health and Society*, Vol. 63, pp. 420–451. https://doi.org/10.2307/3349887

Bleich, S. N., Economos, C. D., Spiker, M. L., Vercammen, K. A., VanEpps, E. M., Block, J. P., … Roberto, C. A. (2017). A Systematic Review of Calorie Labeling and Modified Calorie Labeling Interventions: Impact on Consumer and Restaurant Behavior. *Obesity*, *25*(12), 2018–2044. https://doi.org/10.1002/oby.21940

Boland, M. R., Rusanov, A., So, Y., Lopez-Jimenez, C., Busacca, L., Steinman, R. C., … Weng, C. (2014). From expert-derived user needs to user-perceived ease of use and usefulness: A two-phase mixed-methods evaluation framework. *Journal of Biomedical Informatics*, *52*, 141–150. https://doi.org/10.1016/j.jbi.2013.12.004

Boot, W. R., Charness, N., Czaja, S. J., Sharit, J., Rogers, W. A., Fisk, A. D., … Nair, S. (2015a). Computer proficiency questionnaire: Assessing low and high computer proficient seniors. *Gerontologist*, *55*(3), 404–411. https://doi.org/10.1093/geront/gnt117

Boot, W. R., Charness, N., Czaja, S. J., Sharit, J., Rogers, W. A., Fisk, A. D., … Nair, S. (2015b). Computer Proficiency Questionnaire: Assessing Low and High Computer Proficient Seniors. *The Gerontologist*, *55*(3), 404–411. https://doi.org/10.1093/geront/gnt117

Bosch, M., Zhu, F., Khanna, N., Boushey, C. J., & Delp, E. J. (2011). Combining global and local features for food identification in dietary assessment. *Proceedings - International Conference on Image Processing, ICIP*, 1789–1792. https://doi.org/10.1109/ICIP.2011.6115809

Bovens, L. (2009). The ethics of nudge. In *Preference Change: Approaches from Philosophy, conomics and Psychology* (pp. 207–219). https://doi.org/10.4337/9781786430557.00012

Brauner, P., Calero Valdez, A., Schroeder, U., & Ziefle, M. (2013). Increase physical fitness and create health awareness through exergames and gamification: The role of individual factors, motivation and acceptance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7946 LNCS*, 349–362. https://doi.org/10.1007/978-3-642-39062-3_22

Briz-Ponce, L., & García-Peñalvo, F. J. (2015). An Empirical Assessment of a Technology Acceptance Model for Apps in Medical Education. *Journal of Medical Systems*, *39*(11), 1–5. https://doi.org/10.1007/s10916-015-0352-x

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *Proceedings - International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Brooke, J. (1996). The system usability scale: a quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 189–194. Retrieved from https://books.google.com/books?hl=en&lr=&id=ujFRDwAAQBAJ&oi=fnd&pg=PA189&dq=Brooke+SUS&ots=Z9TPv1ZRBG&sig=AphoSga5g8NMvjgH15cdWSfJGhE#v=onepage&q=Brooke SUS&f=false

Brown, W., Yen, P.-Y., Rojas, M., & Schnall, R. (2013). Assessment of the Health IT Usability Evaluation Model (Health-ITUEM) for evaluating mobile health (mHealth) technology. *Journal of Biomedical Informatics*, *46*(6), 1080–1087. https://doi.org/10.1016/j.jbi.2013.08.001

Bucher, T., Collins, C., Rollo, M. E., McCaffrey, T. A., De Vlieger, N., Van der Bend, D., … Perez-Cueto, F. J. A. (2016). Nudging consumers towards healthier choices: a systematic review of positional influences on food choice. *The British Journal of Nutrition*, *115*(12), 2252–2263. https://doi.org/10.1017/S0007114516001653

Burgess, A. (2012). "Nudging" healthy lifestyles: The UK experiments with the behavioural alternative to regulation and the market. *European Journal of Risk Regulation*, *3*(1), 3–16. https://doi.org/10.1017/S1867299X00001756

Burke, L. E., Wang, J., & Sevick, M. A. (2011). Self-Monitoring in Weight Loss: A Systematic Review of the Literature. *Journal of the American Dietetic Association*, *111*(1), 92–102. https://doi.org/10.1016/j.jada.2010.10.008

Burton-Jones, A., Grange, C., & Student, D. (2011). *From Use to Effective Use: A Representation Theory Perspective*.

Byrne, K., Frazee, K., Sims-Gould, J., & Martin-Matthews, A. (2012). Valuing the Older Person in the Context of Delivery and Receipt of Home Support. *Journal of Applied Gerontology*, *31*(3), 377–401. https://doi.org/10.1177/0733464810387578

Campos, S., Doxey, J., & Hammond, D. (2011). Nutrition labels on pre-packaged foods: a systematic review. *Public Health Nutrition*, *14*(8), 1496–1506. https://doi.org/10.1017/S1368980010003290

Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 Ways to Nudge: A review of technology-mediated nudging in human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1–15. https://doi.org/10.1145/3290605.3300733

Carter, Michelle C, Albar, S. A., Morris, M. A., Mulla, U. Z., Hancock, N., Evans, C. E., … Cade, J. E. (2015). Development of a UK Online 24-h Dietary Assessment Tool: myfood24. *Nutrients*, *7*(6), 4016–4032. https://doi.org/10.3390/nu7064016

Carter, Michelle Clare, Burley, V. J., Nykjaer, C., & Cade, J. E. (2013). Adherence to a smartphone application for weight loss compared to website and paper diary: Pilot randomized controlled trial. *Journal of Medical Internet Research*, *15*(4). https://doi.org/10.2196/jmir.2283

Castelnuovo, G., Pietrabissa, G., Manzoni, G. M., Cattivelli, R., Rossi, A., Novelli, M., … Molinari, E. (2017). Cognitive behavioral therapy to aid weight loss in obese patients: Current perspectives. *Psychology Research and Behavior Management*, *10*, 165–173. https://doi.org/10.2147/PRBM.S113278

Cecchini, M., & Warin, L. (2016). Impact of food labelling systems on food choices and eating behaviours: a systematic review and meta-analysis of randomized studies. *Obesity Reviews*, *17*(3), 201–210. https://doi.org/10.1111/obr.12364

Chao, W. Y., & Hass, Z. (2020). Choice-Based User Interface Design of a Smart Healthy Food Recommender System for Nudging Eating Behavior of Older Adult Patients with Newly Diagnosed Type II Diabetes. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12208 LNCS*, 221–234. https://doi.org/10.1007/978-3-030-50249-2_17

Chao, W. Y., Lehto, M., Pitts, B., & Hass, Z. (2021). Evaluation of the Effectiveness of an Interpretive Nutrition Label Format in Improving Healthy Food Discrimination Using Signal Detection Theory. *Advances in Intelligent Systems and Computing*, *1201 AISC*, 342–348. https://doi.org/10.1007/978-3-030-51041-1_45

Chao, W. Y., Qu, Q. X., Zhang, L., & Duffy, V. G. (2017). Age and computer skill level difference in aging-centered design: A case study of a social type website. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10287 LNCS*, 132–141. https://doi.org/10.1007/978-3-319-58466-9_13

Charness, N., & Boot, W. R. (2009). Aging and information technology use: Potential and barriers. *Current Directions in Psychological Science*, *18*(5), 253–258. https://doi.org/10.1111/j.1467-8721.2009.01647.x

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, A. (2015). New data shows losing 80% of mobile users is normal, and why the best apps do better. Retrieved August 16, 2020, from Uncategorized website: https://andrewchen.co/new-data-shows-why-losing-80-of-your-mobile-users-is-normal-and-that-the-best-apps-do-much-better/

Chen, J., Lieffers, J., Bauman, A., Hanning, R., & Allman-Farinelli, M. (2017). Designing Health Apps to Support Dietetic Professional Practice and Their Patients: Qualitative Results From an International Survey. *JMIR MHealth and UHealth*, *5*(3). https://doi.org/10.2196/mhealth.6945

Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., & Yang, J. (2009). PFID: Pittsburgh Fast-food Image Dataset. *Proceedings - International Conference on Image Processing, ICIP*, 289–292. https://doi.org/10.1109/ICIP.2009.5413511

Cho, J. (2016). The impact of post-adoption beliefs on the continued use of health apps. *International Journal of Medical Informatics*, *87*, 75–83. https://doi.org/10.1016/j.ijmedinf.2015.12.016

Cho, Jaehee. (2016). The impact of post-adoption beliefs on the continued use of health apps. *International Journal of Medical Informatics*, *87*, 75–83. https://doi.org/10.1016/j.ijmedinf.2015.12.016

Cho, Jaehee, Quinlan, M. M., Park, D., & Noh, G.-Y. (2014). Determinants of Adoption of Smartphone Health Apps among College Students. *American Journal of Health Behavior*, *38*(6), 860–870. https://doi.org/10.5993/AJHB.38.6.8

Chung, M. K., Kim, D., Na, S., & Lee, D. (2010). Usability evaluation of numeric entry tasks on keypad type and age. *International Journal of Industrial Ergonomics*, *40*(1), 97–105. https://doi.org/10.1016/j.ergon.2009.08.001

Cimperman, M., Makovec Brencic, M., & Trkman, P. (2016). Analyzing older users' home telehealth services acceptance behavior—applying an Extended UTAUT model. *International Journal of Medical Informatics*, *90*, 22–31. https://doi.org/10.1016/j.ijmedinf.2016.03.002

Cimperman, M., Makovec Brenčič, M., & Trkman, P. (2016). Analyzing older users' home telehealth services acceptance behavior-applying an Extended UTAUT model. *International Journal of Medical Informatics*, *90*, 22–31. https://doi.org/10.1016/j.ijmedinf.2016.03.002

Cordeiro, F., Bales, E., Cherry, E., & Fogarty, J. (2015). Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*, 3207–3216. https://doi.org/10.1145/2702123.2702154

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Czaja, S. J., Boot, W. R., Charness, N., & Rogers, W. A. (2017). Designing for Older Adults: Principles and Creative Human Factors Approaches. In *Designing for Older Adults: Principles and Creative Human Factors Approaches*. https://doi.org/10.4324/9780203485729

Czaja, S. J., Lee, C. C., Branham, J., & Remis, P. (2012). OASIS connections: Results from an evaluation study. *Gerontologist*, *52*(5), 712–721. https://doi.org/10.1093/geront/gns004

Czaja, S. J., Sharit, J., Lee, C. C., Nair, S. N., Hernández, M. A., Arana, N., & Fu, S. H. (2013). Factors influencing use of an e-health website in a community sample of older adults. *Journal of the American Medical Informatics Association*, *20*(2), 277–284. https://doi.org/10.1136/amiajnl-2012-000876

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems : theory and results* (Massachusetts Institute of Technology). Retrieved from http://dspace.mit.edu/handle/1721.1/15192

De Jong, N., Wentzel, J., Kelders, S., Oinas-Kukkonen, H., & Van Gemert-Pijnen, J. (2014). Evaluation of perceived persuasiveness constructs by combining user tests and expert assessments. *CEUR Workshop Proceedings*, *1153*, 7–15. Retrieved from https://www.researchgate.net/publication/262604684

Deng, Z. (2013). Understanding public users' adoption of mobile health service. *International Journal of Mobile Communications*, *11*(4), 351–373. https://doi.org/10.1504/IJMC.2013.055748

Dijkstra, W., Smit, J. H., & Comijs, H. C. (2001). Using social desirability scales in research among the elderly. *Quality and Quantity*, *35*(1), 107–115. https://doi.org/10.1023/A:1004816210439

Dix, A., Finlay, J. E., Abowd, G. D., & Beale, R. (2003). *Human-Computer Interaction* (3 edition). Retrieved from https://www.amazon.com/Human-Computer-Interaction-3rd-Alan-Dix/dp/0130461091

Dunbar-Jacob, J., Erlen, J. A., Schlenk, E. A., Ryan, C. M., Sereika, S. M., & Doswell, W. M. (2000). Adherence in chronic disease. *Annual Review of Nursing Research*, Vol. 18, pp. 48–90. https://doi.org/10.1891/0739-6686.18.1.48

Egnell, M., Ducrot, P., Touvier, M., Allès, B., Hercberg, S., Kesse-Guyot, E., & Julia, C. (2018). Objective understanding of Nutri-Score Front-Of-Package nutrition label according to individual characteristics of subjects: Comparisons with other format labels. *PLOS ONE*, *13*(8), e0202095. https://doi.org/10.1371/journal.pone.0202095

Elbert, S. P., Dijkstra, A., & Oenema, A. (2016). A Mobile Phone App Intervention Targeting Fruit and Vegetable Consumption: The Efficacy of Textual and Auditory Tailored Health Information Tested in a Randomized Controlled Trial. *Journal of Medical Internet Research*, *18*(6). https://doi.org/10.2196/jmir.5056

Elsweiler, D., Harvey, M., Ludwig, B., & Said, A. (2015). Bringing the "healthy" into food recommenders. *CEUR Workshop Proceedings*, *1533*, 33–36.

Eng, D. S., & Lee, J. M. (2013). The Promise and Peril of Mobile Health Applications for Diabetes and Endocrinology. *Pediatric Diabetes*, *14*(4), 231–238. https://doi.org/10.1111/pedi.12034

European Commission. (2020). *Commission staff working document - evaluation of the Regulation (EC) No 1924/2006 on nutrition and health claims made on foods with regard to nutrient profiles and health claims made on plants and their preparations and of the general regulatory framewor*. Retrieved from https://ec.europa.eu/food/sites/food/files/safety/docs/labelling_nutrition-claims_swd_2020-95_part-1.pdf

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Eyles, H., McLean, R., Neal, B., Jiang, Y., Doughty, R. N., McLean, R., & Ni Mhurchu, C. (2017). A salt-reduction smartphone app supports lower-salt food purchases for people with cardiovascular disease: Findings from the SaltSwitch randomised controlled trial. *European Journal of Preventive Cardiology*, *24*(13), 1435–1444. https://doi.org/10.1177/2047487317715713

Farage, M. A., Miller, K. W., Ajayi, F., & Hutchins, D. (2012). Design principles to accommodate older adults. *Global Journal of Health Science*, Vol. 4, pp. 2–25. https://doi.org/10.5539/gjhs.v4n2p2

Farber, N., Shinkle, D., Lynott, J., Fox-Grage, W., & Harrell, R. (2011). Aging in place: A state survey of livability policies and practices. In *Journal of aging research* (Vol. 2012). https://doi.org/10.1155/2012/120952

Fasola, J., & Mataric, M. (2013). A Socially Assistive Robot Exercise Coach for the Elderly. *Journal of Human-Robot Interaction*, *2*(2). https://doi.org/10.5898/jhri.2.2.fasola

Fastame, M. C., & Penna, M. P. (2012). Does Social Desirability Confound the Assessment of Self-Reported Measures of Well-Being and Metacognitive Efficiency in Young and Older Adults? *Clinical Gerontologist*, *35*(3), 239–256. https://doi.org/10.1080/07317115.2012.660411

Fisk, A. D., Czaja, S. J., Rogers, W. A., Charness, N., & Sharit, J. (2009). Designing for Older Adults: Principles and Creative Human Factors Approaches. In *Designing for Older Adults: Principles and Creative Human Factors Approaches* (Second). https://doi.org/10.4324/9780203485729

Flaherty, S.-J., McCarthy, M., Collins, A., & McAuliffe, F. (2018). Can existing mobile apps support healthier food purchasing behaviour? Content analysis of nutrition content, behaviour change theory and user quality integration. *Public Health Nutrition*, *21*(2), 288–298. https://doi.org/10.1017/S1368980017002889

Fletcher, J., & Jensen, R. (2015). Mobile health: Barriers to Mobile Phone Use in the Aging Population. *On-Line Journal of Nursing Informatics*, *19*(3). Retrieved from https://www.himss.org/mobile-health-barriers-mobile-phone-use-aging-population

Fogg, B. (2009). A behavior model for persuasive design. *ACM International Conference Proceeding Series*, *350*, 1. https://doi.org/10.1145/1541948.1541999

Fogg, B. J. (2002). Persuasive technology. *Ubiquity*, *2002*(December), 2. https://doi.org/10.1145/764008.763957

Fogg, B. J. (2003). Persuasive Technology: Using Computers to Change What We Think and Do. In *Persuasive Technology: Using Computers to Change What We Think and Do*. https://doi.org/10.1016/B978-1-55860-643-2.X5000-8

Franco, R. Z., Fallaize, R., Lovegrove, J. A., & Hwang, F. (2018). Online dietary intake assessment using a graphical food frequency app (eNutri): Usability metrics from the EatWellUK study. *PLOS ONE*, *13*(8), e0202006. https://doi.org/10.1371/journal.pone.0202006

Free, C., Phillips, G., Galli, L., Watson, L., Felix, L., Edwards, P., … Haines, A. (2013). The Effectiveness of Mobile-Health Technology-Based Health Behaviour Change or Disease Management Interventions for Health Care Consumers: A Systematic Review. *PLoS Medicine*, *10*(1), e1001362. https://doi.org/10.1371/journal.pmed.1001362

Freid, V. M., Bernstein, A. B., & Bush, M. A. (2012). Multiple chronic conditions among adults aged 45 and over: trends over the past 10 years. *NCHS Data Brief*, (100), 1–8. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db100_tables.pdf#3.

Freyne, J., & Berkovsky, S. (2010). Recommending food: Reasoning on recipes and ingredients. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6075 LNCS*, 381–386. https://doi.org/10.1007/978-3-642-13470-8_36

Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., … McBurnie, M. A. (2001). Frailty in older adults: Evidence for a phenotype. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *56*(3), 146–156. https://doi.org/10.1093/gerona/56.3.m146

Friis, R., Skov, L. R., Olsen, A., Appleton, K. M., Saulais, L., Dinnella, C., … Perez-Cueto, F. J. A. (2017). Comparison of three nudge interventions (priming, default option, and perceived variety) to promote vegetable consumption in a self-service buffet setting. *PLOS ONE*, *12*(5), e0176028. https://doi.org/10.1371/journal.pone.0176028

Funk, K. L., Stevens, V. J., Appe, L. J., Bauck, A., Brantley, P. J., Champagne, C. M., … Vollmer, W. M. (2010). Associations of internet website use with weight change in a long-term weight loss maintenance program. *Journal of Medical Internet Research*, *12*(3), e29. https://doi.org/10.2196/jmir.1504

Gao, C., Zhou, L., Liu, Z., Wang, H., & Bowers, B. (2017). Mobile application for diabetes self-management in China: Do they fit for older adults? *International Journal of Medical Informatics*, *101*, 68–74. https://doi.org/10.1016/j.ijmedinf.2017.02.005

Gao, M., Kortum, P., & Oswald, F. (2018). Psychometric Evaluation of the USE (Usefulness, Satisfaction, and Ease of use) Questionnaire for Reliability and Validity. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1414–1418. https://doi.org/10.1177/1541931218621322

Gavin, J. R., Stolar, M. W., Freeman, J. S., & Spellman, C. W. (2010). Improving outcomes in patients with type 2 diabetes mellitus: practical solutions for clinical challenges. *The Journal of the American Osteopathic Association*, *110*(5 Suppl 6). Retrieved from https://jaoa.org/article.aspx?articleid=2093906

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. Retrieved from https://pure.mpg.de/rest/items/item_2102373/component/file_2102372/content

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Retrieved from http://webs.wofford.edu/pechwj/Heuristics and Biases - The Psychology of Intuitive Judgment.pdf

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., … Turner, M. B. (2013, January 1). Executive summary: Heart disease and stroke statistics-2013 update: A Report from the American Heart Association. *Circulation*, Vol. 127, pp. 143–152. https://doi.org/10.1161/CIR.0b013e318282ab8f

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula Pub.

Gregory, A., Mackintosh, S., Kumar, S., & Grech, C. (2017). Experiences of health care for older people who need support to live at home: A systematic review of the qualitative literature. *Geriatric Nursing*, *38*(4), 315–324. https://doi.org/10.1016/j.gerinurse.2016.12.001

Grindrod, K. A., Li, M., & Gates, A. (2014). Evaluating user perceptions of mobile medication management applications with older adults: a usability study. *JMIR MHealth and UHealth*, *2*(1), e11. https://doi.org/10.2196/mhealth.3048

Guigoz, Y., Lauque, S., & Vellas, B. J. (2002). Identifying the elderly at risk for malnutrition the mini nutritional assessment. *Clinics in Geriatric Medicine*, Vol. 18, pp. 737–757. https://doi.org/10.1016/S0749-0690(02)00059-9

Guo, X., Sun, Y., Wang, N., Peng, Z., & Yan, Z. (2013). The dark side of elderly acceptance of preventive mobile health services in China. *Electronic Markets*, *23*(1), 49–61. https://doi.org/10.1007/s12525-012-0112-4

Hakobyan, L., Lumsden, J., Shaw, R., & O'Sullivan, D. (2016). A longitudinal evaluation of the acceptability and impact of a diet diary app for older adults with age-related macular degeneration. *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 124–134. https://doi.org/10.1145/2935334.2935356

Hales, S., Dunn, C., Wilcox, S., & Turner-McGrievy, G. M. (2016). Is a Picture Worth a Thousand Words? Few Evidence-Based Features of Dietary Interventions Included in Photo Diet Tracking Mobile Apps for Weight Loss. *Journal of Diabetes Science and Technology*, *10*(6), 1399–1405. https://doi.org/10.1177/1932296816651451

Hamari, J., Koivisto, J., & Pakkanen, T. (2014). Do persuasive technologies persuade? - A review of empirical studies. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8462 LNCS*, 118–136. https://doi.org/10.1007/978-3-319-07127-5_11

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hänninen, R., Taipale, S., & Luostari, R. (2020). Exploring heterogeneous ICT use among older adults:The warm experts' perspective. *New Media and Society*, 146144482091735. https://doi.org/10.1177/1461444820917353

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the Manipulation of Choice. *European Journal of Risk Regulation*, *4*(1), 3–28. https://doi.org/10.1017/s1867299x00002762

Hanson, V. L. (2011). Technology skill and age: What will be the same 20 years from now? *Universal Access in the Information Society*, *10*(4), 443–452. https://doi.org/10.1007/s10209-011-0224-1

Harden, A., Peersman, G., Oliver, S., Mauthner, M., & Oakley, A. (1999). A systematic review of the effectiveness of health promotion interventions in the workplace. *Occupational Medicine*, *49*(8), 540–548. https://doi.org/10.1093/occmed/49.8.540

Harris, M. I. (2000). Health care and health status and outcomes for patients with type 2 diabetes. *Diabetes Care*, *23*(6), 754–758. https://doi.org/10.2337/diacare.23.6.754

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Harte, R., Glynn, L., Broderick, B., Rodriguez-Molinero, A., Baker, P., McGuiness, B., … ÓLaighin, G. (2014). Human Centred Design Considerations for Connected Health Devices for the Older Adult. *Journal of Personalized Medicine*, *4*(2), 245–281. https://doi.org/10.3390/jpm4020245

Hickson, M. (2006, January 1). Malnutrition and ageing. *Postgraduate Medical Journal*, Vol. 82, pp. 2–8. https://doi.org/10.1136/pgmj.2005.037564

Hillier-Brown, F. C., Summerbell, C. D., Moore, H. J., Routen, A., Lake, A. A., Adams, J., … Brown, T. J. (2017). The impact of interventions to promote healthier ready-to-eat meals (to eat in, to take away or to be delivered) sold by specific food outlets open to the general public: a systematic review. *Obesity Reviews*, *18*(2), 227–246. https://doi.org/10.1111/obr.12479

Hingle, M., & Patrick, H. (2016). There Are Thousands of Apps for That: Navigating Mobile Technology for Nutrition Education and Behavior. *Journal of Nutrition Education and Behavior*, *48*(3), 213-218.e1. https://doi.org/10.1016/j.jneb.2015.12.009

Hodgkins, C., Barnett, J., Wasowicz-Kirylo, G., Stysko-Kunkowska, M., Gulcan, Y., Kustepeli, Y., … Raats, M. (2012). Understanding how consumers categorise nutritional labels: A consumer derived typology for front-of-pack nutrition labelling. *Appetite*, *59*(3), 806–817. https://doi.org/10.1016/j.appet.2012.08.014

Holden, S. S., Zlatevska, N., & Dubelaar, C. (2016). Whether Smaller Plates Reduce Consumption Depends on Who's Serving and Who's Looking: A Meta-Analysis. *Journal of the Association for Consumer Research*, *1*(1), 134–146. https://doi.org/10.1086/684441

Hollis, J. F., Gullion, C. M., Stevens, V. J., Brantley, P. J., Appel, L. J., Ard, J. D., … Svetkey, L. P. (2008). Weight Loss During the Intensive Intervention Phase of the Weight-Loss Maintenance Trial. *American Journal of Preventive Medicine*, *35*(2), 118–126. https://doi.org/10.1016/j.amepre.2008.04.013

Holzinger, A. (2005, January 1). Usability engineering methods for software developers. *Communications of the ACM*, Vol. 48, pp. 71–74. https://doi.org/10.1145/1039539.1039541

Holzinger, A., Searle, G., & Witzer, A. N. (2007). On some aspects of improving mobile applications for the elderly. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4554 LNCS*(PART 1), 923–932. https://doi.org/10.1007/978-3-540-73279-2_103

Hongu, N., Pope, B. T., Bilgiç, P., Orr, B. J., Suzuki, A., Kim, A. S., … Roe, D. J. (2015). Usability of a smartphone food picture app for assisting 24-hour dietary recall: a pilot study. *Nutrition Research and Practice*, *9*(2), 207–212. https://doi.org/10.4162/nrp.2015.9.2.207

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human Computer Studies*, *64*(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

Hu, P. J., Chau, P. Y. K., Liu Sheng, O. R., & Tam, K. Y. (1999). Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology. *Journal of Management Information Systems*, *16*(2), 91–112. https://doi.org/10.1080/07421222.1999.11518247

International Organization for Standardization [ISO]. (2018). Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts (ISO Standard No. 9241-11:2018(en)). Retrieved July 13, 2020, from Iso website: https://www.iso.org/standard/63500.html

Isaković, M., Sedlar, U., Volk, M., & Bešter, J. (2016). Usability pitfalls of diabetes mHealth apps for the elderly. *Journal of Diabetes Research*, *2016*. https://doi.org/10.1155/2016/1604609

Isaković, Maša, Sedlar, U., Volk, M., & Bešter, J. (2016). Usability Pitfalls of Diabetes mHealth Apps for the Elderly. *Journal of Diabetes Research*, *2016*. https://doi.org/10.1155/2016/1604609

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/ida-2002-6504

Jeon, E., & Park, H.-A. (2015). Factors Affecting Acceptance of Smartphone Application for Management of Obesity. *Healthcare Informatics Research*, *21*(2), 74–82. https://doi.org/10.4258/hir.2015.21.2.74

Jiang, H., Starkman, J., Liu, M., & Huang, M. C. (2018). Food Nutrition Visualization on Google Glass: Design Tradeoff and Field Evaluation. *IEEE Consumer Electronics Magazine*, *7*(3), 21–31. https://doi.org/10.1109/MCE.2018.2797740

Joe, J., & Demiris, G. (2013, October). Older adults and mobile phones for health: A review. *Journal of Biomedical Informatics*, Vol. 46, pp. 947–954. https://doi.org/10.1016/j.jbi.2013.06.008

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475. Retrieved from https://www.jstor.org/stable/pdf/3132137.pdf

Kahneman, D. (2011). *Thinking, Fast and Slow*. Retrieved from https://books.google.com/books?id=SHvzzuCnuv8C

Kaminska, O., & Foulsham, T. (2013). Understanding Sources of Social Desirability Bias in Different Modes : Evidence from Eye-tracking Olena Kaminska, ISER Working Paper Series. *ISER Working Paper Series*, 1–11. Retrieved from https://ideas.repec.org/p/ese/iserwp/2013-04.html

Kankanhalli, A., Shin, J., & Oh, H. (2019). Mobile-Based Interventions for Dietary Behavior Change and Health Outcomes: Scoping Review. *JMIR MHealth and UHealth*, *7*(1), e11312. https://doi.org/10.2196/mhealth.11312

Karlson, A. K., Robertson, G., Robbins, D. C., Czerwinski, M., & Smith, G. (2006). *FaThumb: A Facet-based Interface for Mobile Search*. Retrieved from http://www.osha.gov/pls/imis/sic_manual.html

Kaufman, D. R., Starren, J., Patel, V. L., Morin, P. C., Hilliman, C., Pevzner, J., … Shea, S. (2003). A cognitive framework for understanding barriers to the productive use of a diabetes home telemedicine system. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2003*, 356–360. Retrieved from /pmc/articles/PMC1480118/?report=abstract

Kawano, Y., & Yanai, K. (2014). Food image recognition with deep convolutional features. *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 589–593. https://doi.org/10.1145/2638728.2641339

Keeney, R. L. (2008). Personal Decisions Are the Leading Cause of Death. *OPERATIONS RESEARCH*, *56*(6). https://doi.org/10.1287/opre.1080.0588

Kelly, J. T., Reidlinger, D. P., Hoffmann, T. C., & Campbell, K. L. (2016). Telehealth methods to deliver dietary interventions in adults with chronic disease: A systematic review and meta-analysis1,2. *American Journal of Clinical Nutrition*, *104*(6), 1693–1702. https://doi.org/10.3945/ajcn.116.136333

Klein, H. A., & Meininger, A. R. (2004). Self management of medication and diabetes: Cognitive control. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *34*(6), 718–725. https://doi.org/10.1109/TSMCA.2004.836791

Kleinen, A., Scherp, A., & Staab, S. (2014). Interactive faceted search and exploration of open social media data on a touchscreen mobile phone. *Multimedia Tools and Applications*, *71*(1), 39–60. https://doi.org/10.1007/s11042-013-1366-3

Kleinschmidt, S., Peters, C., & Leimeister, J. M. (2016). ICT-enabled service innovation in human-centered service systems: A systematic literature review. *2016 International Conference on Information Systems, ICIS 2016*. https://doi.org/10.2139/ssrn.3159136

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., … Diabetes Prevention Program Research Group. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine*, *346*(6), 393–403. https://doi.org/10.1056/NEJMoa012512

Kurniawan, S., & Zaphiris, P. (2005, December 12). *Research-derived Web Design Guidelines for Older People*. 129–135. https://doi.org/10.1145/1090785.1090810

Kwon, M.-W., Mun, K., Lee, J. K., McLeod, D. M., & D'Angelo, J. (2017). Is mobile health all peer pressure? The influence of mass media exposure on the motivation to use mobile health apps. *Convergence*, *23*(6), 565–586. https://doi.org/10.1177/1354856516641065

Laing, B. Y., Mangione, C. M., Tseng, C. H., Leng, M., Vaisberg, E., Mahida, M., … Bell, D. S. (2014). Effectiveness of a smartphone application for weight loss compared with usual care in overweight primary care patients. *Annals of Internal Medicine*, *161*(10 Suppl), S5–S12. https://doi.org/10.7326/M13-3005

Lee, D. Y., & Lehto, M. R. (2013). User acceptance of YouTube for procedural learning: An extension of the Technology Acceptance Model. *Computers and Education*, *61*(1), 193–208. https://doi.org/10.1016/j.compedu.2012.10.001

Lee, E., Han, S., & Jo, S. H. (2017). Consumer choice of on-demand mHealth app services: Context and contents values using structural equation modeling. *International Journal of Medical Informatics*, *97*(97), 229–238. https://doi.org/10.1016/j.ijmedinf.2016.10.016

Lee, W., Chae, Y. M., Kim, S., Ho, S. H., & Choi, I. (2010). Evaluation of a mobile phone-based diet game for weight control. *Journal of Telemedicine and Telecare*, *16*(5), 270–275. https://doi.org/10.1258/jtt.2010.090913

Lehto, M. R., Nah, F. F.-H., & Yi, J. S. (2012). Decision-Making Models, Decision Support, and Problem Solving. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 192–242). Retrieved from http://doi.wiley.com/10.1002/9781118131350.ch7

Lehto, T., Oinas-Kukkonen, H., & Drozd, F. (2012). Factors affecting perceived persuasiveness of a behavior change support system. *International Conference on Information Systems, ICIS 2012*, *3*, 1926–1939. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.959.1225&rep=rep1&type=pdf

Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Liang, X., & Yan, Z. (2019). A survey on game theoretical methods in Human-Machine Networks. *Future Generation Computer Systems*, *92*, 674–693. https://doi.org/10.1016/j.future.2017.10.051

Lin, T., & Imamiya, A. (2006). Evaluating usability based on multimodal information: An empirical study. *ICMI'06: 8th International Conference on Multimodal Interfaces, Conference Proceeding*, 364–371. https://doi.org/10.1145/1180995.1181063

Liu, C., Zhu, Q., Holroyd, K. A., & Seng, E. K. (2011). Status and trends of mobile-health applications for iOS devices: A developer's perspective. *Journal of Systems and Software*, *84*(11), 2022–2033. https://doi.org/10.1016/j.jss.2011.06.049

Liu, Chang, Cao, Y., Luo, Y., Chen, G., Vokkarane, V., & Ma, Y. (2016). Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9677*, 37–48. https://doi.org/10.1007/978-3-319-39601-9_4

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLOS ONE*, *13*(8), e0199661. https://doi.org/10.1371/journal.pone.0199661

Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability Interface*, *8*(2), 3–6. https://doi.org/10.1177/1078087402250360

Macmillan, N. A., & Creelman, C. D. (2005a). *Detection theory : a user's guide*. Lawrence Erlbaum.

Macmillan, N. A., & Creelman, C. D. (2005b). Detection theory: A user's guide, 2nd ed. In *Detection theory: A user's guide, 2nd ed.* Retrieved from https://psycnet.apa.org/record/2004-19022-000

Maglio, P. P. (2015, June 1). Smart service systems, human-centered service systems, and the mission of service science. *Service Science*, Vol. 7, pp. ii–iii. https://doi.org/10.1287/serv.2015.0100

Malik, V. S., Willett, W. C., & Hu, F. B. (2016, August 9). The revised nutrition facts label: A step forward and more room for improvement. *JAMA - Journal of the American Medical Association*, Vol. 316, pp. 583–584. https://doi.org/10.1001/jama.2016.8005

Malloy-Weir, L., & Cooper, M. (2017). Health literacy, literacy, numeracy and nutrition label understanding and use: a scoping review of the literature. *Journal of Human Nutrition and Dietetics*, *30*(3), 309–325. https://doi.org/10.1111/jhn.12428

Mancia, G., De Backer, G., Dominiczak, A., Cifkova, R., Fagard, R., Germano, G., … Zanchetti, A. (2007, June 1). 2007 Guidelines for the management of arterial hypertension. *European Heart Journal*, Vol. 28, pp. 1462–1536. https://doi.org/10.1093/eurheartj/ehm236

Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, *14*(1), 81–95. https://doi.org/10.1007/s10209-014-0348-1

Marteau, T. M., Ogilvie, D., Roland, M., Suhrcke, M., & Kelly, M. P. (2011, January 29). Judging nudging: Can nudging improve population health? *BMJ*, Vol. 342, pp. 263–265. https://doi.org/10.1136/bmj.d228

Martin, D., Vicente, O., Vicente, S., Ballesteros, J., & Maynar, M. (2014, March 15). *I Will Prescribe You an App*. 58:1-58:8. Retrieved from http://dl.acm.org/citation.cfm?id=2685617.2685675

Matthew-Maich, N., Harris, L., Ploeg, J., Markle-Reid, M., Valaitis, R., Ibrahim, S., … Isaacs, S. (2016). Designing, Implementing, and Evaluating Mobile Health Technologies for Managing Chronic Conditions in Older Adults: A Scoping Review. *JMIR MHealth and UHealth*, *4*(2), e29. https://doi.org/10.2196/mhealth.5127

McColl-Kennedy, J. R., Vargo, S. L., Dagger, T. S., Sweeney, J. C., & van Kasteren, Y. (2012). Health Care Customer Value Cocreation Practice Styles. *Journal of Service Research*, *15*(4), 370–389. https://doi.org/10.1177/1094670512442806

McCormack, B., Roberts, T., Meyer, J., Morgan, D., & Boscart, V. (2012). Appreciating the ‘person’ in long-term care. *International Journal of Older People Nursing*, *7*(4), 284–294. https://doi.org/10.1111/j.1748-3743.2012.00342.x

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*(6), 882–888. https://doi.org/10.1037/0022-006X.51.6.882

McLaughlin, A. C., Whitlock, L. A., Lester, K. L., & McGraw, A. E. (2017). Older adults’ self-reported barriers to adherence to dietary guidelines and strategies to overcome them. *Journal of Health Psychology*, *22*(3), 356–363. https://doi.org/10.1177/1359105315603472

Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., & French, D. P. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: The CALO-RE taxonomy. *Psychology & Health*, *26*(11), 1479–1498. https://doi.org/10.1080/08870446.2010.540664

Mika, S. (2011). Challenges for nutrition recommender systems. *CEUR Workshop Proceedings*, *786*, 25–33. Retrieved from www.nhs.uk/livewell/healthy-

Mitzner, T. L., Boron, J. B., Fausset, C. B., Adams, A. E., Charness, N., Czaja, S. J., … Sharit, J. (2010). Older adults talk technology: Technology usage and attitudes. *Computers in Human Behavior*, *26*(6), 1710–1721. https://doi.org/10.1016/j.chb.2010.06.020

Mokdad, A. H., Marks, J. S., Stroup, D. F., & Gerberding, J. L. (2004, March 10). Actual Causes of Death in the United States, 2000. *Journal of the American Medical Association*, Vol. 291, pp. 1238–1245. https://doi.org/10.1001/jama.291.10.1238

Morey, S. A., Barg-Walkow, L. H., & Rogers, W. A. (2017). Managing Heart Failure On the Go: Usability Issues with mHealth Apps for Older Adults. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1–5. https://doi.org/10.1177/1541931213601496

Murillo-Munoz, M. F., Vazquez-Briseno, M., Cota, C. X. N., & Nieto-Hipolito, J. I. (2018). A framework for design and development of persuasive mobile systems. *2018 28th International Conference on Electronics, Communications and Computers, CONIELECOMP 2018*, *2018-January*, 59–66. https://doi.org/10.1109/CONIELECOMP.2018.8327176

Nielsen, J. (1994, May 7). *Usability Inspection Methods*. 413–414. https://doi.org/10.1145/259963.260531

Nielsen, J. (2004). Usability engineering. In *Computer Science Handbook, Second Edition* (pp. 45-1-45–21). https://doi.org/10.1201/b16768-38

Nielsen, J. (2013). Minimize Cognitive Load to Maximize Usability. Retrieved July 21, 2020, from Nielsen Norman Group website: https://www.nngroup.com/articles/minimize-cognitive-load/

Nolte, S., Elsworth, G. R., & Osborne, R. H. (2013). Absence of social desirability bias in the evaluation of chronic disease self-management interventions. *Health and Quality of Life Outcomes*, *11*(1), 114. https://doi.org/10.1186/1477-7525-11-114

Norman, D. A., & Draper, S. W. (1986). User Centered System Design: New Perspectives on Human-Computer Interaction. In *L. Erlbaum Associates Inc.*

Nurgalieva, L., Jara Laconich, J. J., Baez, M., Casati, F., & Marchese, M. (2019). A Systematic Literature Review of Research-Derived Touchscreen Design Guidelines for Older Adults. *IEEE Access*, *7*, 22035–22058. https://doi.org/10.1109/ACCESS.2019.2898467

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, *24*(1), 485–500. https://doi.org/10.17705/1cais.02428

Olson, K. E., O'Brien, M. A., Rogers, W. A., & Charness, N. (2011). Diffusion of technology: Frequency of use for younger and older adults. *Ageing International*, *36*(1), 123–145. https://doi.org/10.1007/s12126-010-9077-9

Parker, S. J., Jessel, S., Richardson, J. E., & Reid, M. C. (2013). Older adults are mobile too!Identifying the barriers and facilitators to older adults' use of mHealth for pain management. *BMC Geriatrics*, *13*(1), 1–8. https://doi.org/10.1186/1471-2318-13-43

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). https://doi.org/10.1016/b978-0-12-590241-0.50006-x

Perry, C., Chhatralia, K., Damesick, D., Hobden, S., & Volpe, L. J. Della. (2015). *Behavioural insights in health care Nudging to reduce inefficiency and waste*. Retrieved from https://pdfs.semanticscholar.org/56a7/5974e5d3630b490e45124aa7361531987975.pdf

Pew Reserach Center. (2018). *Mobile Fact Sheet*. Retrieved from Pew Research Center website: https://www.pewinternet.org/fact-sheet/mobile/

Poushter, J. (2016). Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. In *Pew Research Center* (Vol. 22). Retrieved from www.pewresearch.org.

Powers, D. M. W. (2003). Recall and Precision versus the Bookmaker. *International Conference on Cognitive Science*, (July 2003), 529–534. Retrieved from https://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/27159/Powers Recall.pdf?sequence=1&isAllowed=y

Powers, D. M. W. (2011). Evaluation: From Precision, Recall And F-measure To ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. Retrieved from http://dspace.flinders.edu.au/dspace/http://www.bioinfo.in/contents.php?id=51

Proctor, R. W., & Zandt, T. Van. (2018). *Human Factors in Simple and Complex Systems*. Retrieved from https://books.google.com/books?id=mBFFDwAAQBAJ

Putzer, Gavin; Jaramillo, J. (2015). Premature Mortality Costs Associated with Lifestyle Factors among US Citizens. *Review of Public Administration and Management*, *03*(01), 1. https://doi.org/10.4172/2315-7844.1000177

Putzer, Gavin; Jaramillo, J. (2017). Trends in Behavioral Risk Factors Resulting in Premature Death in US from 2000-2015. *International Journal of Research in Business Studies and Management*, *4*(4), 8–12. https://doi.org/10.22259/ijrbsm.0404002

Qu, Q. X., Zhang, L., Chao, W. Y., & Duffy, V. (2017). User experience design based on eye-tracking technology: A case study on smartphone apps. *Advances in Intelligent Systems and Computing*, *481*, 303–315. https://doi.org/10.1007/978-3-319-41627-4_27

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Ray, J. J. (1988). Lie scales and the elderly. *Personality and Individual Differences*, *9*(2), 417–418. https://doi.org/10.1016/0191-8869(88)90106-7

Reddy, S. S. K. (2000, October 1). Health outcomes in type 2 diabetes. *International Journal of Clinical Practice, Supplement*, pp. 46–53. Retrieved from https://europepmc.org/article/med/11965832

Reynolds, W. M. (1982). Development of reliable and valid short forms of the marlowe-crowne social desirability scale. *Journal of Clinical Psychology*, *38*(1), 119–125. https://doi.org/10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Robinson, E., Thomas, J., Aveyard, P., & Higgs, S. (2014). What everyone else is eating: a systematic review and meta-analysis of the effect of informational eating norms on eating behavior. *Journal of the Academy of Nutrition and Dietetics*, *114*(3), 414–429. https://doi.org/10.1016/j.jand.2013.11.009

Rodeschini, G. (2011, December 1). Gerotechnology: A new kind of care for aging? An analysis of the relationship between older people and technology. *Nursing and Health Sciences*, Vol. 13, pp. 521–528. https://doi.org/10.1111/j.1442-2018.2011.00634.x

Rogers, W. A., & Fisk, A. D. (2010). Toward a psychological science of advanced technology design for older adults. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *65 B*(6), 645–653. https://doi.org/10.1093/geronb/gbq065

Rossi, S., Staffa, M., & Tamburro, A. (2018). Socially Assistive Robot for Providing Recommendations: Comparing a Humanoid Robot with a Mobile Application. *International Journal of Social Robotics*, *10*(2), 265–278. https://doi.org/10.1007/s12369-018-0469-4

Rosson, M. B., & Carroll, J. M. (2002). Usability engineering : scenario-based development of human-computer interaction. In *Interface*. Retrieved from https://books.google.com/books?hl=en&lr=&id=sRPg0IYhYFYC&oi=fnd&pg=PP2&dq=scenario+based+usability&ots=mHNo6gPFIR&sig=11DGzkEdgeuzOriRGbHnvkXRc0g#v=onepage&q=scenario based usability&f=false

Sanjari, S. S., Jahn, S., & Boztug, Y. (2017). Dual-process theory and consumer response to front-of-package nutrition label formats. *Nutrition Reviews*, *75*(11), 871–882. https://doi.org/10.1093/nutrit/nux043

Santa-Rosa, J. G., & Fernandes, H. (2012). Application and analysis of the affinities diagram on the examination of usability problems among older adults. *Work*, *41*(SUPPL.1), 328–332. https://doi.org/10.3233/WOR-2012-0177-328

Sattler, H., & Hensel-Börner, S. (2001). A Comparison of Conjoint Measurement with Self-Explicated Approaches. In *Conjoint Measurement* (pp. 121–133). https://doi.org/10.1007/978-3-662-06392-7_5

Sayago, S., & Blat, J. (2007). *A preliminary usability evaluation of strategies for seeking online information with elderly people*. Retrieved from http://www.edaverneda.org

Schnall, R., Cho, H., & Liu, J. (2018). Health information technology usability evaluation scale (Health-ITUES) for usability assessment of mobile health technology: Validation study. *Journal of Medical Internet Research*, *20*(1). https://doi.org/10.2196/mhealth.8851

Schnall, R., Rojas, M., Bakken, S., Brown, W., Carballo-Dieguez, A., Carry, M., … Travers, J. (2016). A user-centered model for designing consumer mobile health (mHealth) applications (apps). *Journal of Biomedical Informatics*, *60*, 243–251. https://doi.org/10.1016/j.jbi.2016.02.002

Selinger, E., & Whyte, K. P. (2010). Competence and Trust in Choice Architecture. *Knowledge, Technology & Policy*, *23*(3–4), 461–482. https://doi.org/10.1007/s12130-010-9127-3

Sharit, J., Hern, M. A., Czaja, S. J., HernándezHern, M. A., & Czaja, S. J. (2008). Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the Web. *ACM Transactions on Computer-Human Interaction*, *15*(1). https://doi.org/10.1145/1352782.1352785

Shegog, R., & Begley, C. E. (2017). Clinic-based mobile health decision support to enhance adult epilepsy self-management: An intervention mapping approach. *Frontiers in Public Health*, *5*(OCT), 256. https://doi.org/10.3389/fpubh.2017.00256

Shim, J.-S., Oh, K., & Kim, H. C. (2014). Dietary assessment methods in epidemiologic studies. *Epidemiology and Health*, *36*, e2014009. https://doi.org/10.4178/epih/e2014009

Sorgente, A., Pietrabissa, G., MauroManzoni, G., Re, F., Simpson, S., Perona, S., … Castelnuovo, G. (2017, June 26). Web-based interventions for weight loss or weight loss maintenance in overweight and obese people: A systematic review of systematic reviews. *Journal of Medical Internet Research*, Vol. 19, p. 229. https://doi.org/10.2196/jmir.6972

Soubelet, A., & Salthouse, T. A. (2011). Influence of social desirability on age differences in self-reports of mood and personality. *Journal of Personality*, *79*(4), 741–762. https://doi.org/10.1111/j.1467-6494.2011.00700.x

Spiegler, R. (2015). On the equilibrium effects of nudging. *Journal of Legal Studies*, *44*(2), 389–416. https://doi.org/10.1086/684291

Stanislaw, H. (1999). Calculation of signal detection theory measures. In *Behavior Research Methods, Instruments, & Computers* (Vol. 3).

Stephens, M. A. P., Rook, K. S., Franks, M. M., Khan, C., & Iida, M. (2010). Spouses Use of Social Control To Improve Diabetic Patients' Dietary Adherence. *Families, Systems and Health*, *28*(3), 199–208. https://doi.org/10.1037/a0020513

Stoica, E., & Hearst, M. A. (2004). *Nearly-automated metadata hierarchy creation*. 117–120. https://doi.org/10.3115/1613984.1614014

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *28*(2), 191–193. https://doi.org/10.1002/1097-4679(197204)28:2<191::AID-JCLP2270280220>3.0.CO;2-G

Svensson, A., & Larsson, C. (2015). A Mobile Phone App for Dietary Intake Assessment in Adolescents: An Evaluation Study. *JMIR MHealth and UHealth*, *3*(4), e93. https://doi.org/10.2196/mhealth.4804

Svensson, A., Magnusson, M., Larsson, C., Svensson, Å., Magnusson, M., & Larsson, C. (2016). Overcoming Barriers: Adolescents' Experiences Using a Mobile Phone Dietary Assessment App. *JMIR MHealth and UHealth*, *4*(3), e92. https://doi.org/10.2196/mhealth.5700

Swets, J. A. (2014). Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. In *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. https://doi.org/10.4324/9781315806167

Talati, Z., Pettigrew, S., Kelly, B., Ball, K., Dixon, H., & Shilton, T. (2016). Consumers' responses to front-of-pack labels that vary by interpretive content. *Appetite*, *101*, 205–213. https://doi.org/10.1016/j.appet.2016.03.009

Taype, G. E. E., & Calani, M. C. B. (2020). Extending persuasive system design frameworks: An exploratory study. *Advances in Intelligent Systems and Computing*, *1137 AISC*, 35–45. https://doi.org/10.1007/978-3-030-40690-5_4

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Retrieved from https://books.google.com/books?id=cYdYngEACAAJ

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Revised &). Retrieved from https://www.amazon.com/Nudge-Improving-Decisions-Health-Happiness/dp/014311526X

Trang Tran, T. N., Atas, M., Felfernig, A., & Stettinger, M. (2018). An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*, *50*(3), 501–526. https://doi.org/10.1007/s10844-017-0469-0

Tuomilehto, J., Lindström, J., Eriksson, J. G., Valle, T. T., Hämäläinen, H., Ilanne-Parikka, P., … Uusitupa, M. (2001). Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. *New England Journal of Medicine*, *344*(18), 1343–1350. https://doi.org/10.1056/NEJM200105033441801

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. Retrieved from https://science.sciencemag.org/content/185/4157/1124

United Nations, D. of E. and S. A. (2019). World Population Prospects 2019: Highlights. In *United Nations Publication*. Retrieved from https://population.un.org/wpp

Uziel, L. (2010, May). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, Vol. 5, pp. 243–262. https://doi.org/10.1177/1745691610369465

van der Mark, M., Jonasson, J., Svensson, M., Linnéb, Y., Rössner, S., & Lagerros, Y. T. (2009). Older members perform better in an internet-based behavioral weight loss program compared to younger members. *Obesity Facts*, *2*(2), 74–79. https://doi.org/10.1159/000209383

van Kleef, E., Seijdell, K., Vingerhoeds, M. H., de Wijk, R. A., & van Trijp, H. C. M. (2018). The effect of a default-based nudge on the choice of whole wheat bread. *Appetite*, *121*, 179–185. https://doi.org/10.1016/j.appet.2017.11.091

van Oostrom, S. H., Gijsen, R., Stirbu, I., Korevaar, J. C., Schellevis, F. G., Picavet, H. S. J., & Hoeymans, N. (2016). Time Trends in Prevalence of Chronic Diseases and Multimorbidity Not Only due to Aging: Data from General Practices and Health Surveys. *PLOS ONE*, *11*(8), e0160264. https://doi.org/10.1371/journal.pone.0160264

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Manage. Sci.*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). *Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead*. Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2800121

Vincent Delhomme. (n.d.). *Front-of-pack nutrition labelling in the European Union : a behavioural, legal and political analysis*. https://doi.org/10.13140/RG.2.2.17513.93287

von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior. In *Theory of Games and Economic Behavior*. https://doi.org/10.2307/2019327

Wagner, A., Tran, T., & Ladwig, G. (2011). *Browsing-Oriented Semantic Faceted Search*. https://doi.org/10.1007/978-3-642-23088-2_22

Wallace, S. E., Graham, C., & Saraceno, A. (2013). Older Adults' Use of Technology. *Perspectives on Gerontology*, *18*(2), 50–59. https://doi.org/10.1044/gero18.2.50

Wandke, H., Sengpiel, M., & Sönksen, M. (2012, October). Myths about older people's use of information and communication technology. *Gerontology*, Vol. 58, pp. 564–570. https://doi.org/10.1159/000339104

Wang, B. R., Park, J. Y., Chung, K., & Choi, I. Y. (2014). Influential Factors of Smart Health Users according to Usage Experience and Intention to Use. *Wireless Personal Communications*, *79*(4), 2671–2683. https://doi.org/10.1007/s11277-014-1769-0

Wang, S., Bolling, K., Mao, W., Reichstadt, J., Jeste, D., Kim, H.-C., & Nebeker, C. (2019). Technology to Support Aging in Place: Older Adults' Perspectives. *Healthcare*, *7*(2), 60. https://doi.org/10.3390/healthcare7020060

Weinmann, M., Schneider, C., & Brocke, J. vom. (2016). Digital Nudging. *Business and Information Systems Engineering*, *58*(6), 433–436. https://doi.org/10.1007/s12599-016-0453-1

Weiss, B. D., Mays, M. Z., Martz, W., Castro, K. M., DeWalt, D. A., Pignone, M. P., … Hale, F. A. (2005). Quick Assessment of Literacy in Primary Care: The Newest Vital Sign. *The Annals of Family Medicine*, *3*(6), 514–522. https://doi.org/10.1370/afm.405

Wellard-Cole, L., Potter, M., Jung, J. J. (Joseph) (Joseph) J., Chen, J., Kay, J., Allman-Farinelli, M., … Allman-Farinelli, M. (2018). A Tool to Measure Young Adults' Food Intake: Design and Development of an Australian Database of Foods for the Eat and Track Smartphone App. *JMIR MHealth and UHealth*, *6*(11), e12136. https://doi.org/10.2196/12136

Wenzel, M. A., Schultze-Kraft, R., Meinecke, F. C., Cardinaux, F., Kemp, T., Müller, K. R., … Blankertz, B. (2015). EEG-based usability assessment of 3D shutter glasses. *Journal of Neural Engineering*, *13*(1), 016003. https://doi.org/10.1088/1741-2560/13/1/016003

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error Rates in Users of Automatic Face Recognition Software. *PLOS ONE*, *10*(10), e0139827. https://doi.org/10.1371/journal.pone.0139827

Whitehead, L., & Seaton, P. (2016, May 1). The effectiveness of self-management mobile phone and tablet apps in long-term condition management: A systematic review. *Journal of Medical Internet Research*, Vol. 18, p. 97. https://doi.org/10.2196/jmir.4883

Whitlock, L. A., & McLaughlin, A. C. (2012). Identifying usability problems of blood glucose tracking apps for older adult users. *Proceedings of the Human Factors and Ergonomics Society*, 115–119. https://doi.org/10.1177/1071181312561001

Wildenbos, G. A., Jaspers, M. W. M., Schijven, M. P., & Dusseljee-Peute, L. W. (2019). Mobile health for older adult patients: Using an aging barriers framework to classify usability problems. *International Journal of Medical Informatics*, *124*, 68–77. https://doi.org/10.1016/j.ijmedinf.2019.01.006

Wildenbos, G. A., Peute, L., & Jaspers, M. (2018). *Aging barriers influencing mobile health usability for older adults: A literature based framework (MOLD-US)*. https://doi.org/10.1016/j.ijmedinf.2018.03.012

Wildenbos, G. A., Peute, L. W., & Jaspers, M. W. M. (2017). Influence of human factor issues on patient-centered mhealth apps' impact; where do we stand? In *Studies in Health Technology and Informatics* (Vol. 228). https://doi.org/10.3233/978-1-61499-678-1-190

Wilson, M. L. (2011). Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *3*(3), 1–143. https://doi.org/10.2200/s00371ed1v01y201111icr020

Wilson, M. L., André, P., & Schraefel, M. C. (2008). Backward highlighting: enhancing faceted search. *UIST 2008 - Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*, 235–238. https://doi.org/10.1145/1449715.1449754

World Health Organization. (2017). *"Best buys" and other recommended interventions for the prevention and control of noncommunicable diseases. Updated (2017) appendix 3 of the global action plan for the prevention and control of noncommunicable diseases 2013-2020.* Retrieved from http://www.who.int/ncds/governance/appendix3-update-discussion-paper/en/

Wu, Y., Yao, X., Vespasiani, G., Nicolucci, A., Dong, Y., Kwong, J., … Li, S. (2017). Mobile App-Based Interventions to Support Diabetes Self-Management: A Systematic Review of Randomized Controlled Trials to Identify Functions Associated with Glycemic Efficacy. *JMIR MHealth and UHealth*, *5*(3), 35. https://doi.org/10.2196/mhealth.6522

Yang, S., Chen, M., Pomerleau, D., & Sukthankar, R. (2010). Food recognition using statistics of pairwise local features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2249–2256. https://doi.org/10.1109/CVPR.2010.5539907

Yei-Yu Yeh, & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*(1), 111–120. https://doi.org/10.1177/001872088803000110

Yuan, S., Ma, W., Kanthawala, S., & Peng, W. (2015). Keep Using My Health Apps: Discover Users' Perception of Health and Fitness Apps with the UTAUT2 Model. *Telemedicine and E-Health*, *21*(9), 735–741. https://doi.org/10.1089/tmj.2014.0148

Zapata, B. C., Fernández-Alemán, J. L., Idri, A., & Toval, A. (2015). Empirical Studies on Usability of mHealth Apps: A Systematic Literature Review. *Journal of Medical Systems*, *39*(2), 1–19. https://doi.org/10.1007/s10916-014-0182-2

Zhou, J., Rau, P. L. P., & Salvendy, G. (2014). Older adults use of smart phones: An investigation of the factors influencing the acceptance of new functions. *Behaviour and Information Technology*, *33*(6), 552–560. https://doi.org/10.1080/0144929X.2013.780637

Zhu, F., Bosch, M., Khanna, N., Boushey, C. J., & Delp, E. J. (2015). Multiple hypotheses image segmentation and classification with application to dietary assessment. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 377–388. https://doi.org/10.1109/JBHI.2014.2304925

Zhu, M., Xu, C., & Wu, Y. F. B. (2013). IFME: Information filtering by multiple examples with under-sampling in a digital library environment. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 107–110. https://doi.org/10.1145/2467696.2467736

Zickuhr, K., & Madden, M. (2012). Older Adults and Internet Use. *Pew Research Center's Internet & American Life Project*, 2–23. Retrieved from http://pewinternet.org/Reports/2012/Older-

Ziefle, M., & Bay, S. (2005). How older adults meet complexity: Aging effects on the usability of different mobile phones. *Behaviour & Information Technology*, *24*(5), 375–389. https://doi.org/10.1080/01449290512331319030