1-1-2013

Introducing Transferability and the Upmds Usability Framework in a Multiple-Device System

Yunchen Huang

Introducing transferability and the UPMDS usability framework in a multiple-device system

By

Yunchen Huang

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Industrial and Systems Engineering
in the Department of Industrial and Systems Engineering

Mississippi State, Mississippi

May 2013

Introducing transferability and the UPMDS usability framework in a multiple-device system

By

Yunchen Huang

Approved:

Lesley Strawderman Assistant Professor of Industrial and Systems Engineering (Director of Dissertation)	Kari Babski-Reeves Associate Professor and Graduate Coordinator of Industrial and Systems Engineering (Committee Member)
Mingzhou Jin Committee Participant of Industrial and Systems Engineering (Committee Member)	Edward Swan II Professor of Computer Science and Engineering (Committee Member)

Sarah A. Rajala
Dean of the Bagley College of Engineering

Name: Yunchen Huang

Date of Degree: May 10, 2013

Institution: Mississippi State University

Major Field: Industrial and Systems Engineering

Major Professor: Dr. Lesley Strawderman

Title of Study: Introducing transferability and the UPMDS usability framework in a multiple-device system

Pages in Study: 222

Candidate for Degree of Doctor of Philosophy

This research introduces the concept of transferability into the usability construct and creates the Usability Paradigm for Multiple Device System (UPMDS) to conceptualize and quantify the usability in multiple device scenarios. This study fills the literature gap that no effective method exists in measuring transferability and in quantifying usability in a multiple device context. This study also answers the research questions regarding the impact of task complexity, user experience, and device order on the total usability of the system.

Study one follows a systematic approach to develop, validate, and apply a new questionnaire tailored specifically to measure the transferability within a multiple device system. The System Transferability Questionnaire (STQ) is obtained after validation with 15 question items. In a software usability study, the STQ demonstrated excellent internal reliability and validity. Results show that the STQ is effective in capturing four factors regarding transferability, which are transfer experience (TE), overall experience (OE), consistency perception (CP) and functionality perception (FP). Validation results show good convergent, discriminant, criterion and nomonlogical validity.

Study two adopts a systematic tool to consolidate usability constructs into a total usability score. The study utilizes principal component analysis (PCA) to determine the weight of the four usability components (satisfaction, transferability, effectiveness, and efficiency), which is used when obtaining the total usability score. Results show slightly different weights for the four components. This quantitative tool can be applied in different usability context in which multiple devices are involved. Usability specialists are encouraged to adjust the tool based on different usability scenarios.

Study three investigates the impact of task complexity, user experience, and device order on the total system usability. Results show that the total usability score is not affected by task complexity, user experience or device order. However, lower physical task complexity leads to longer performance time and lower errors from the users. High experienced users have significantly lower errors made in tasks. The machine order also has divergent results. When the mini-lathe machine was used first, users had better transferability results but poorer performance outcomes as compared to when the drill press was used first.

DEDICATION

I dedicate this dissertation to my parents and my wife, who always had faith in me and support me all through my study. Through this dissertation, I hope to inspire others to earn an advanced Master's or Doctorate degree, especially in the fields of science, technology, engineering and mathematics.

ACKNOWLEDGEMENTS

I would like to express my utmost appreciation to my advisor, Dr. Lesley Strawderman, for her time, patience and advice. She has provided me with valuable guidance for this dissertation. She has always been supportive and work with me as a model of a true professor, teacher, and advisor. I would also like to thank my committee members, Dr. Kari Babski-Reeves, Dr. Mingzhou Jin, and Dr. Edward Swan for their commitment in helping to formulate ideas and providing guidance essential to the completion of this dissertation.

In addition I would like to thank all of the faculty and staff of the Department of Industrial and Systems Engineering for their guidance, help and support during my doctoral study. I would also like to thank my colleges and friends of the Human Systems lab for all the support, help, motivation, laughter and company.

Last but not least, I would like to thank my mother and father for their continuous support and faith in me. I would also like to thank my wife who would always be here for me, support me, and accompany me through my doctoral study. I would like to dedicate this dissertation to my beloved parents, who supported and cared for me throughout my life. I know I could have not done all this work without their boundless support and unending love.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Motivation

To design an environment that would promote better human use has always been the objective of human factors practitioners. This need has driven the development of usability research as a way of analyzing, evaluating, and designing the products, devices, interfaces, and tools around us. Traditional usability research defined usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-11, 1998, p. 2). While this widely used definition clearly defines the context of use: "a product" "in a specified context of use", the persona: "specified users", and the usability construct: "effectiveness, efficiency and satisfaction", it still limits the context of human use to a single product. As many studies have pointed out that the context of use decide the usability constructs (e.g. Shackel, 1991; Maguire, 2001), the traditional definition of usability may limit the application and accuracy of usability evaluation in many circumstances.

With fast developing technology, user interactions with products are changing rapidly. Traditional single user and single product interaction is slowly becoming obsolete. Instead, users tend to engage more in multi-media and multi-device interaction. With cloud computing technology, numerous software, applications, and services can be

available on different devices or products (e.g. laptop, mobile phone, PDA, TV, gaming console, etc.). Medical doctors often operate multiple medical devices to diagnose and treat patients. In manufacturing facility, workers have to monitor multiple machines simultaneously. In assembling lines, workers have to use different tools or machines to finish a part. In an office, staffs have to use multiple computer software programs to accomplish a task. The context of multiple devices use is almost everywhere in our life. In these situations, traditional usability construct is not enough to characterize the quality of use of these devices. Information regarding the interrelationship of two or more devices needs to be captured to better represent the usability construct.

Not only is it important to conceptualize the usability framework for a multiple device system, it is also critical to identify an effective measurement of the usability in this construct. Three major challenges remain in the measurement of usability: defining appropriate usability framework, whether to use subjective or objective measurements, and how to adjust the framework according to specific contexts. Up till now, a wide range of usability models have been established to obtain a universal construct of usability attributes (Bevan, 1995; Macleod, 1994; Macleod and Rengger, 1993; Sears, 1995; Seffah et al, 2006). While aimed at addressing the first and second challenges, these studies failed to address the third challenge. The existing usability measurement framework literature shared the same limitation in that they primarily focused on single interface usability. In addition, how to appropriately address both subjective and objective measures in a usability study remains a challenge for many usability studies (Hornbak, 2006).

There is still a lack of understanding regarding how to appropriately measure usability when multiple interfaces are involved. To overcome this challenge, a new usability framework will be introduced that incorporates users' performance measures, single-device satisfaction and the transferability between devices.

UPMDS Usability Framework

The Usability Paradigm for Multiple Device Systems (UPMDS) was first introduced by Huang and Strawderman (2011). It was revised and used as the usability model guiding the evaluating and measuring of the usability in this dissertation. In this framework, usability is composed of a subjective component and an objective component. The subjective component is further decomposed into single device satisfaction and multi-device transferability. The objective component is further decomposed into effectiveness and efficiency (Figure 1.1).

The UPMDS framework is appropriate for evaluating the system usability for multi-device system. The multi-device system is defined as the system in which users have to interact with multiple devices to complete a goal.

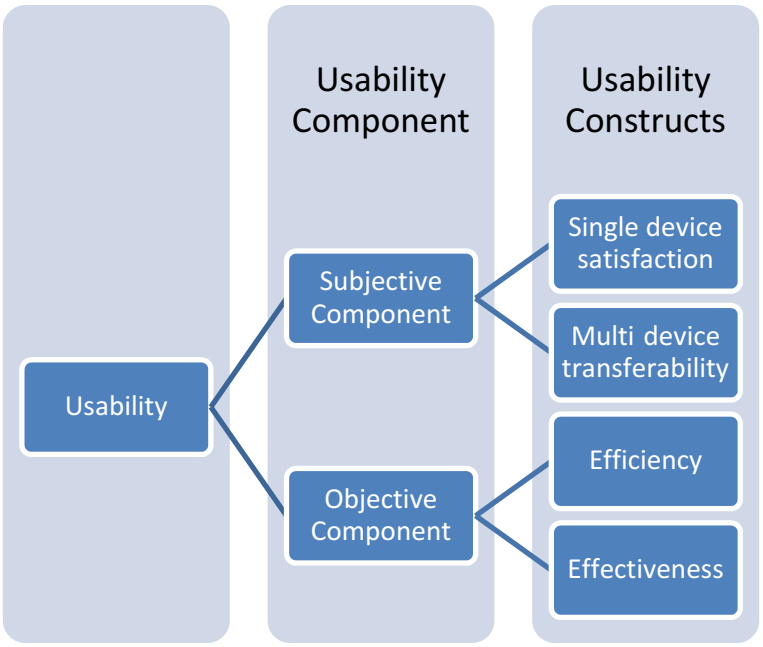


Figure 1.1 UPMDS Attributes Break Down and Corresponding Aspects

Objective measures of effectiveness and efficiency can be obtained from task completion time and errors. User satisfaction can be measured using standard questionnaires such as the System Usability Scale (SUS) or Post Study System Usability Questionnaire (PSSUQ). As transferability is another key aspect of this framework, the System Transferability Questionnaire (STQ) will be developed in this dissertation to measure this variable.

The subjective component of usability consists of users' subjective perception on the usability of each single device and subjective perception on the transferability between the devices. Transferability is a device attribute which is defined as the extent to which users can effectively transfer their knowledge of using the previous device to the learning and using of the current device. It comes from the notion of transfer of learning which describes users' attentive learning processes, but is different from transfer of

learning in that transferability describes a device's design features rather than the learning process. Transferability is a device characteristics that represents the traditional usability attributes such as learnability, retention, and consistency.

The objective component of the usability characterizes the extent to which users' performance is affected by transferring learning between devices. This subset has two usability aspects: effectiveness and efficiency. Efficiency characterizes how fast and easy users can change from using one device to using another device. It is measured from the task completion time. Effectiveness characterizes the extent to which users can successfully adopt the knowledge gained from a previous device and transfer it to a new device. It is measured by error rates or task completion.

The UPMDS framework serves as the guiding theoretical basis for this dissertation. All chapters will be based on this framework and adopt this framework for evaluating and measuring usability.

Dissertation Objective

The overall objective of the dissertation was to investigate, validate, and adjust the newly proposed Usability Paradigm for Multiple Device Systems (UPMDS). This study is also aimed at adopting this framework to measure system usability in real world applications, and apply it to solve research questions in usability and human factors areas. This study adds to the theoretically body of knowledge of current usability evaluation. The UPMDS framework can also be practically developed into an adjustable usability/transferability evaluation tool so that usability practitioners can customize and input the usability specifications; such as completion time, errors, single-device usability and transferability; to compute an output of the total system usability score.

Dissertation Structure

The overall research question of this study is: *Would UPMDS be a valid framework to characterize and measure usability in a multiple device system and can it be applied and help usability researchers in answering usability research questions?*

To effectively answer this question and the associated research objectives, three distinct studies were conducted to address the above research question. The overall research structure of the studies is illustrated in Figure 1.2.

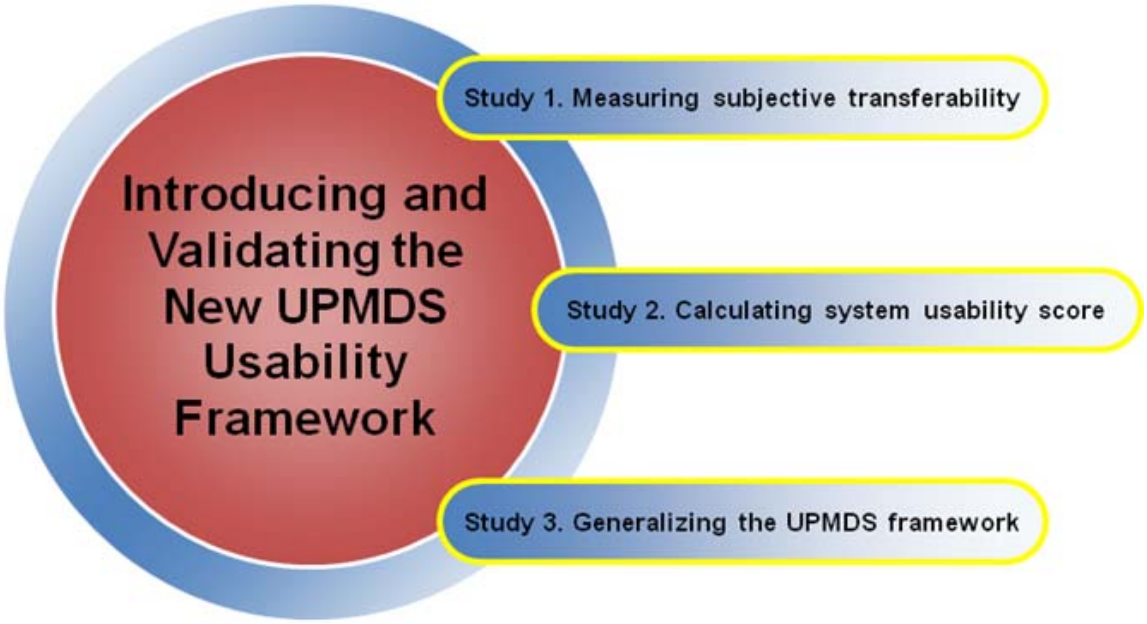


Figure 1.2 Dissertation Research Scope

Study 1was aimed at identifying an effective subjective measurement tool to characterize the transferability between devices. This filled the literature gap in measurement of subjective transferability. The System Transferability Questionnaire (STQ) was developed for the evaluation of transferability between devices. A software

usability study was conducted to test reliability and validity of STQ using factor analysis. STQ was modified according to the result of factor analysis. A complete questionnaire items were compiled as the STQ. The overall research question of Study 1 is: *Can we develop a System Transferability Questionnaire that can serve as a reliable and valid tool to effectively capture users' perception regarding the various aspects of transferability in a real world scenario?*

Study 2 adopted theoretical approaches to calculate a total usability score. The UPMDS framework is the guiding framework for calculating the total usability score. Both subjective component (transferability and satisfaction) and objective component (effectiveness and efficiency) were consolidated to obtain a single system usability score. The overall research question of study 2 is: *Can we properly identify the weight and effect different measures have in explaining the overall system usability? How to consolidate all the measures into a single score?*

Study 3 tested the reliability of STQ when applied in a machine usability scenario. More importantly, this study applied the UPMDS framework in a real world usability scenario. The framework was utilized to help answer research questions in usability area. A machine usability study was conducted to address the main research question of this study is: *What are the effects of task complexity, user experience, and task order on the total usability of the multiple device system? Is there interaction effect of task complexity and user experience?*

Theoretical and Empirical Implications of the Proposed Work

The new usability framework UPMDS is introduced to characterize usability construct in multiple-device systems. The literature gap that traditional usability tools

only measure single device usability is filled. A comprehensive and universal model for usability is still not possible, but the new usability framework would be more widely applicable in people's everyday life. When people use a multiple device system, knowledge and learning gained from the previous device may greatly affect their performance in the following devices. A cognitive mapping will happen from the previous device to the current device. Users' satisfaction on each device is no longer the only subjective measure of interest. A smooth and satisfactory transfer between devices would be the new focus of usability specialists.

This dissertation also introduces the study of transfer of learning to the area of usability. When users are transferring between multiple devices, they are in the process of transfer of learning. Users' initial interaction with the previous device will help them create a mental model of the device. When they switch to a new device, the attributes similarity or relational similarity between the two devices may trigger an analogical mapping from the previous device, which causes the effect of transfer of learning. As a traditional study that rooted in behavioral and cognitive psychology, transfer of learning is a theoretical approach. Currently, there is no consistent and comprehensive way to measure transfer of learning. Application of transfer of learning on usability studies opens a door for the measurement of transfer.

The subjective and objective measures of usability have been a debating topic in usability studies. It is recognized that both measures are necessary in usability studies because they may lead to different conclusions regarding the usability of an interface. Studies also suggested that these measures capture different aspects of user performance (Bommer et al., 1995; Yeh and Wickens, 1988). A major challenge, as put forward by

Hornbak (2006), is to "develop subjective measures for aspects of quality-in-use that are currently mainly measured by objective measures, and vice versa, and evaluate their relation." This dissertation will help in investigating the role subjective and objective measures play in evaluating usability.

With the help of this dissertation, usability researchers will now be able to assess the usability of multiple-device systems instead of single interfaces. This will benefit user groups from all areas. In manufacturing, this usability framework can examine the usability between different machines in a manufacturing cell. Workers in cellular manufacturing can improve their performance and lower the errors when switching between machines. Product designers can use our usability framework to evaluate transferability between the previous product and upgraded product, therefore improve consumer use and satisfaction. Service systems and healthcare systems can improve the transferability between devices to reduce errors and boost customer satisfaction.

References

Bevan, N. (1995). Measuring usability as quality of use, *Software Quality Journal 4*,115–130.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., & Mackenzie, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personnel Psychology, 48,* 587–605.

Hornbeck, K (2006). Current practice in measuring usability: challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*, 79-102.

Huang, Y. & Strawderman, L. (2011). Introducing a New Usability Framework for Analyzing Usability in a Multiple-device System. *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting, 55 (1),* 1696-1700.

ISO/IEC. 9241. (1998). Ergonomic requirements for office work with visual display terminals (VDT)s. ISO/IEC 9241-14: 1998 (E).

Macleod, M. (1994). Usability: practical methods for testing and Improvement, *Proceedings of the Norwegian Computer Society Software Conference*, Sandvika, Norway.

Macleod M and Rengger R (1993) The development of DRUM: a software tool for video-assisted usability evaluation. In: JL Alty et al. (eds) People and Computers VIII (pp. 293-309).

Maguire, M. C. (2001). Context of use within usability activities. *International Journal of Human-Computer Studies, 55*, 453-483.

Sears, A. (1995). AIDE: A step toward metric-based interface development tools, Proceedings of the ACM Symposium on User Interface Software and Technology (pp. 101–110). New York: ACM Press.

Seffah, A., Donyaee, M., Kline, R., & Padda, H. (2006). Usability Measurement and Metrics: A Consolidated Model, *Software Quality Journal, 14*, 159-178.

Shackel, B. (1991). Usability-Context, framework, definition, design and evaluation. In Shackel B. & Richardson S. (eds.), Human Factors for Informatics Usability. Cambridge: Cambridge University Press, pp. 21–38.

Yeh, Y, & Wickens, C. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30,* 111-120.

CHAPTER II

INTRODUCING THE SYSTEM TRANSFERABILITY QUESTIONNAIRE (STQ)

Introduction

Technology is rapidly evolving, and users' interactions are incorporating more multi-media and cross-dimensional experience. Not only is traditional service being replaced by electronic services, but a lot of services are accessible through multiple devices (e.g. mobile phones, PDAs, tablet computers, gaming consoles, etc.) with the help of cloud computing. New product upgrades continues to come into the market and replace old ones. In all these contexts, users have to interact with multiple devices to achieve their goals. Traditional usability tools become insufficient to evaluate users' experience when they transfer between using different devices. The Usability Paradigm for Multiple Device System (UPMDS) introduced in this dissertation aims at addressing the gap in measuring transferability between devices and incorporating it into the new usability framework.

As an important aspect of the UPMDS framework, transferability needs to be appropriately measured first. Currently literature on transfer of learning focuses on the measurement of the transfer process. However, as a system attribute, transferability should be measuring how easy the multiple-device system is to afford users to transfer between devices. Traditional usability literature has developed a lot of questionnaires such as the Software Usability Measurement Inventory (SUMI), Questionnaire for User

Interaction Satisfaction (QUIS), and Post-Study System Usability Questionnaire (PSSUQ) to assess the subjective perception on device attributes. However, there are still critics that most of these questionnaires are too generic (Konradt et al., 2003). In addition, it is generally confirmed that the questionnaire need to be tailored based on the context of use (e.g. van Veenendaal, 1998). Therefore, several questionnaires were developed such as Website Analysis and Measurement Inventory (WAMI) (Kirakowski & Cierlik, 1998) for website usability and Measuring Usability of Multi Media Systems (MUMMS) for the evaluation of multimedia products. This enlightened the objective of this chapter, which is filling the literature gap by developing the System Transferability Questionnaire (STQ) to assess the transferability in the context of transferring between devices. A validation study was conducted based on two software of a desktop computer to test the validity of the questionnaire.

Background and Literature Review

Transfer of Learning

The concept of transfer of learning was first introduced by Thorndike and Woodworth (1901). According to the authors, transfer of learning occurred from one context to another context that share similar characteristics. Their study implied that the amount transferred is dependent on the amount of similarity shared between the learning task and the transferred task. In a more recent study Haskell (2000) defined transfer of learning as our use of past learning when learning something new and the application of that learning to both similar and new situations. The research on transfer has emerged in numerous domains. Three major focuses were: taxonomy-oriented research that conceptualized the transfer in different situations, application-driven research that applied

transfer in specific domains, and psychologically-oriented research that studied transfer in a cognitive perspective.

Taxonomy research received much focus at the early stage of transfer research. From the effect of transfer, it can be divided into positive transfer and negative transfer. From the situation of transfer, it can be divided into specific transfer and general transfer, or near and far transfer (Haskell, 2000). From the human processing perspective, transfer can be divided into High-road and low-road transfer (Mayer & Wittrock, 1996; Salomon & Perkins, 1989).

The application-driven research has widely applied transfer of learning in many areas such as aviation, industry and education. Two major focuses of research are the factors impacting on transfer and the measurement of transfer. However, no consistent results were obtained in these two areas. Regarding on the factors impacting on transfer, numerous factors were identified such as learners' cognitive ability, motivation, personality, training design and environmental factors (Burke & Hutchines, 2007). However the amount of impact each factor has on transfer is dependent on the specific transfer situations and varied in different tasks. Few researches came up with a validated model that explained the mechanism underline the transfer of learning. Regarding on the measurement of transfer, early research was oriented to collecting learners' performance data (Ellis, 1965; Povenmire & Roscoe, 1973). Other studies tended to collect subjective data from the learners and use it as a way to measure transfer (Tziner et al., 1991). It is unknown as to which measurement method is better or whether one or more measurement methods should be used.

From psychologically-oriented research point of view, transfer of learning was studied as mental representations. Metaphor, analogy and mental schema were studied instead of the identical elements. Researchers concluded that transfer occurred if initial learning and transfer situation create identical or they overlapped representations (Anderson, 1995; Sternberg & Frensch, 1993). Anderson also redefined the identical elements as the units of declarative and procedural knowledge in the ACT theory (Andersen, 1983a,b; Singley & Anderson, 1989).

Measurement of transfer

Early research on measuring transfer mainly focused on trainees' performance increase in the situation of aviation and education. Two quantitative measurement of transfer were developed: percentage of transfer (Ellis, 1965) and transfer effectiveness ratio (Povenmire & Roscoe, 1973):

In this equation, control represents time, trials, or errors required by a control group to reach a performance criterion. Transfer represents the corresponding measure for an experimental transfer group having received training on a prior or interpolated task. Transfer group time in training program represents time, trials, or errors by an experimental transfer group during prior or interpolated practice on another task.

These two equations provide a good measure of the transfer. However, the definition of the performance criterion is vague. It could be interpreted differently by various individuals and in various situations. In addition, most performance measurements were not as simple as time, trial or error in a lot of transfer situations. As a result, although the equations gave an exact way to calculator the transfer, the interpretation of the results was actually harder.

The focus on quantitative measures of transfer continued through the transfer research in 1980s. Baldwin & Ford (1988) did a comprehensive literature review on transfer of training. Most of the referenced studies measured transfer using the learning outcome and training results. Knowledge retention and skill test was used specifically for the trained domain. These approaches were quantifiable, relevant to the specific trained domain and easy to interpret. However, Baldwin & Ford (1988) questioned the robustness of this approach based on the fact that this approach collapsed the effect of training with the effect of transferring. They suggested research explicitly examine the direct effects of training-design on training outcomes and then examine the effect on conditions of transfer.

The validity of using single-source data to access transfer outcome was a major concern of Baldwin & Ford (1988). This concern was further addressed as "a lack of attention to define the multidimensional nature of transfer" by Ford & Weissbein (1997), which was an updated literature review following the study of Baldwin & Ford (1988). Among the literatures cited in this updated review, many used multiple measurements which included self-reported degree of transfer, behavioral generalization, performance strategy use, supervisory or peer judgment, increased accuracy of performance, etc. These measurements could be divided into two categories: qualitative subjective measures and quantitative objective measures. The subjective measures complement the deficiency of objective measures in that it clearly identified the extent to which trainees has transferred their learning. But the concern over subjective measures was as well obvious. Ford & Weissbein (1997) believed that one's perceptions of transfer may be affected by social desirability, cognitive dissonance, and memory distortions. This may

potentially impact the validity of the measurement. Tziner et al. (1991) did a transfer study and found contradiction in the self-report result and supervisory ratings. These findings imply the need to use multiple criteria for an accurate and valid measurement of transfer.

Most recent research has continued in the direction of using multi-resource feedback and multi-dimensional measurements. Burke & Hutchines (2007) stated that future empirical research should directly access transfer as the criterion variable instead of individual-level variables such as transfer intentions and motivational aspects.

Another qualitative method to measure transferability is heuristic evaluation, an analysis method widely used in measuring interface usability. It involves evaluators inspecting user interfaces using recognized usability principals (Nielsen 1994). It has the advantages of low cost, easy to conduct and quick output (Nielsen & Phillips, 1993; Nielsen, 1993). In addition, heuristic evaluation can be used on incomplete interface prototypes, which can help identify usability problems in the early stage of interface design.

However, heuristic evaluation is far from perfect. Although Nielsen (1994) found that five or six usability experts could identify most of the usability problems through heuristics evaluation, many researchers hold the opposite opinion. Jeffries & Desurvire (1992) states that heuristic evaluation finds a "distressing" number of minor problems that brings about many false alarms. Since end users was not used in the evaluation, results could still be biased by the preconceptions of the evaluators (Nielsen & Molich, 1990; Kantner & Rosenbaum, 1997; Muller, Matheson, Page, & Gallup, 1995)

Therefore, the question remains whether heuristic evaluation is accurate to predict user performance, interface usability and user satisfaction. Various studies have been done to compare heuristic method with other methods. Nielsen & Phillips (1993) compared three methods as evaluating usability. They found that heuristics evaluation is highly variable and user testing is still better in estimating user performances.

A validated and robust measurement of transferability is critical to the understanding of transfer mechanism. Not only does it save a lot of repeated work in comparing different studies that use various scales to measure transfer, but it also provides consistent measurements that promote the study of transfer mechanism. A guideline could be developed to direct future design of human-computer interface so that people's transfer of learning would be facilitated. However, current literature has inconsistent opinions even on the measurement of transfer, let alone transferability.

Models of Transfer

Since the new usability attribute "transferability" was introduced in the study defined as the extent to which users can effectively transfer their knowledge of using the previous device to the learning and using of the current device, it is important to identify the models of transfer to bridge to user performance and usability domain.

Although there are studies that focus on users' learning and performance when interacting with information technology systems (Card, Moran & Newell, 1983; Olson & Olson, 1990; Payne & Green, 1986; Polson, 1987, 1988; Zaharias & Poylymenakou, 2009; Lee, Rhee & Dunham, 2009), and researchers have proposed the "transfer of design" concept (Lewis & Rieman, 1994), few studies combines the concept of transfer of learning with the evaluation of the usability and user performance of real consumer

products. In the following sections conceptual models will be summarized that address transfer of learning. In addition existing empirical studies of transfer of learning in both traditional training and new product design will be examined.

Gick and Holyoak (1980) put forward the concept of analogous thinking in complex problem-solving tasks in early transfer research. They conducted an empirical study in which participants were provided a military story and then asked to solve a medical problem that was analogous to the military problem. The results showed that participants can generate an analogous solution even with partial mapping from the base problem to the target problem. They also noted that one of the key blocks to successful use of analogous mapping would be the failure of retrieving the analogies from memory and noticing its pertinence to the target problem.

While Gick and Holyoak's study (1980) was solely based on problem solving, Dahl and Moreau (2002) applied the study of analogical thinking to product design. They used three empirical studies to examine the influence of analogical thinking on the idea-generation stage of the new product. They found three factors that influence the originality of the product design: the extent of analogical transfer, the types of analogies used and the presence of external primes.

A product should be designed not only from a designer's viewpoint, but also with consideration of users' performance as well as perceptions. Frese et al. (1991) did a study on transfer using word processing software as a platform. Participants were divided into two groups. The error-training group received training that would easily lead to user errors and require user to recover from errors by themselves. The error-avoidant-training group received training that was designed to reduce the chances to make errors.

Whenever an error occurred in error-avoidant-training group, the experimenter would correct the errors immediately. Frese et al. (1991) found that the error-training group was superior to the error-avoidant group in transfer of learning and that the error-training group exhibited better organized mental models. The study proved the validity of using computer based software as a transfer platform and the superiority of error-training design.

A consumer-oriented design philosophy is essential in product design. Chandra and Kamrani (2003) studied the knowledge management approach that focused on implementing a consumer-focused design philosophy to support decision making in the automotive industry. Their approach was successful in improving product quality. Hsieh and Chen (2005) found that both user interaction and user knowledge management are critical in creating superior new product designs. Therefore, a smooth transfer of learning is critical to ensure a better user interaction and user knowledge management.

Usability

Usability has been a key research topic in the area of human factors and human-computer interaction (HCI). There are various existing definitions of usability. One of the earliest definitions of usability was made by Bennett (1979) "the quality of interaction which takes place" (Bennett, 1979, p. 8). Nielsen (1993) defined usability using five key attributes: efficiency, learnability, memorability, errors, and satisfaction. Schneiderman (1992) provided a similar definition that decomposed usability into the speed of performance, time to learn, retention, rate of errors and satisfaction. Recently, a widely accepted definition of usability was given by ISO 9241-11 (ISO/IEC, 1998, p. 2), according to which usability is "the extent to which a product can be used by specified

users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." Usability studies have been applied in areas such as graphical user interface (GUI) design, product design, manufacturing, health care, and service systems to improve user satisfaction and performance.

Measuring Usability

One of the biggest challenges to current usability study is the measurement of usability (Hornbak, 2006). As a broad concept that characterizes interface attributes, usability cannot be directly measured. There are three categories of methods to obtain usability measurements: usability inspection, usability testing, and usability inquiry (Avouris, 2001).

Usability inspection involves having usability experts examine a user interface. It aims at identifying usability problems and the severity of those problems, usually early in the development circle. Three major methods are used: heuristic evaluation, cognitive walkthrough, and pluralistic walkthrough. This type of method is easy to conduct, with low cost and can identify most of the severe usability problems. However, the end users are not involved in the process, which make it less reliable.

Usability testing aims at evaluate a user interface by testing it on real users. A usage context and scenario will be preset before the users start. Users will be tested based on different usability criteria. Users' performances are measured based on the observation of individual users performing specific tasks with the device (e.g., completion time and number of errors). The most widely employed usability testing methods are hallway testing, remote usability testing, and field studies. General techniques involve think-aloud protocol, co-discovery, performance measurement, and eye-tracking. Usability testing

allows usability researchers to control the factor they want to test in a laboratory. Real users are involved in identifying potential usability problems. But it is also costly to carry out.

Usability inquiry involves communication between the users and the evaluators in the evaluation, either through observation, verbal questioning or written questioning. Evaluators are able to obtain users perceptions towards the interaction experience through the communication with the users. Most commonly used methods involve contextual inquiry, field observation, questionnaires, interviews, focus groups, and logging actual use.

There are still arguments whether subjective measurements or objective measurements or both should be adopted in measuring usability and how to find an appropriate usability framework that categorizes different usability attributes and measures them. This will be further elaborated in Chapter III. To further assist the purpose of this chapter, those most commonly employed usability questionnaires will be discussed in detail in the following sections.

Usability Questionnaire

There are many existing usability questionnaires. System Usability Scale (SUS) was developed by Brooke (1996) as a quick and easy way to collect a user's subjective perception about a product. This questionnaire consists of 10 questions, all aiming at addressing one dimension usability. Users are asked to rate each question with a five-point scale from "Strongly disagree" to "Strongly agree". This questionnaire can be adapted by replacing the word "system" with the current device name. This usability questionnaire is widely applicable, easy to administrate and provide a numerical score

output which is easy to interpret. Studies have validated and supported the use of SUS in many usability evaluation scenarios (e.g. Bangor et al., 2008; Kirakowski, 1994).

IBM developed several usability questionnaires among which the After-Scenario Questionnaire (ASQ), the Post Study System Usability Questionnaire (PSSUQ) and the Computer System Usability Questionnaire (CSUQ) were most frequently used. The ASQ is a three-item scenario-based questionnaire that IBM usability evaluators used to assess participant satisfaction after the completion of a scenario. PSSUQ is a 19-item instrument for assessing user satisfaction with system usability. PSSUQ is administrated after the scenario. CSUQ is modified from the PSSUQ and focus more on the computer system usability. It also has 19 questions, except that the wording of the statements does not refer to a usability testing situation. All three questionnaires demonstrated a decent reliability level with alpha greater than 0.89 (Lewis, 1995).

Other widely used questionnaires include the Questionnaire for User Interface Satisfaction (QUIS), the Software Usability Measurement Inventory (SUMI) and the Web Site Analysis and Measurement Inventory (WAMMI). A summary of available usability questionnaires (Table 2.1) is given by Bangor et al. (2008).

Table 2.1 Summary of Existing Usability Questionnaires (Bangor et al., 2008)

Survey Name	Abbreviation	Developer	Survey Length	Interface Measured	Reliability
After Scenario Questionnaire	ASQ	IBM	3	Any	0.93[a]
Computer System Usability Questionnaire	CSUQ	IBM	19	Computer based	0.95[b]
Post-study System Usability Questionnaire	PSSUQ	IBM	19	Computer based	0.96[b]
Software Usability Measurement Inventory	SUMI[c]	HFRG	50	Software	0.89[d]
System Usability Scale	SUS	DEC	10	Any	0.85[e]
Usefulness Satisfaction and Ease of Use	USE	Lund	30	Any	Unreported[f]
Web Site Analysis and Measurement Inventory	WAMI	HFRG	20	Web based	0.96[g]

Note:[a]Lewis (1995). [b]Lewis (2002). [c]Kirakowski and Corbett (1993). [d]Igbaria and Nachman (1991). [e]Kirakowski (1994). [f]Lund (2001). [g]Kirakowski, Claridge, and Whitehand (1998).

Most of the existing usability questionnaires mentioned above are aimed at accessing usability of single device. There is no known validated questionnaires that can successfully collect user's subjective perception regarding the transferability between multiple devices, which leave a research gap for this dissertation to address.

Study Objective

The objective of this study was to identify an effective approach to obtain reliable subjective measurements of system transferability. The new System Transferability Questionnaire (STQ) is introduced. A computer software study was conducted to test the validity and reliability of the STQ. Specific modifications were made to the survey questions based on validation results. The correlation of STQ scores with users' performance data and existing usability questionnaire scores was investigated.

The overall research question for this study is: *Can we develop a System Transferability Questionnaire that can serve as a reliable and valid tool to effectively capture users' perception regarding the various aspects of transferability in a real world scenario?*

The following specific research questions were raised and aimed to be addressed in this study:

- *Can we create a questionnaire that can effectively capture users' perception regarding the transferability between devices?*
- *What aspects/facet of the transferability does this questionnaire help to explain?*
- *Would this questionnaire be reliable and valid to be used in a real-world scenario?*

Methodology

In this section, the method of developing the STQ is first provided. In addition, a validation study was designed. The methodology of conducting the validation study and testing the reliability and validity of the questionnaire is also presented.

Questionnaire Construction

It's critical to establish the questionnaire construct and context of use before the development of questionnaire items (Netemeyer, et al., 2003). In this study, STQ is developed to appropriately measure users' subjective transferability when using multiple devices. This questionnaire will be designed to fit into the UPMDS framework developed in previous study (Huang & Strawderman, 2011). Therefore the STQ will represent a construct similar to the traditional usability construct. Effectiveness, efficiency, and satisfaction were adopted from the usability definition (ISO/IEC, 1998, p. 2) and were selected as the construct for STQ.

The context of use of STQ is different from traditional usability questionnaires. The STQ is to be used in multiple devices systems in which users have to interact with more than one device to achieve a goal. The targeting device could be any devices involving a user interface, ranging from mobile devices, computer software to hand tools, to machines. Previous studies found that two key aspects are indicative of the transfer performance between devices: transparency between two devices (Huang et al., 2012), and the learning effect after the task change (Huang et al., 2012; Strawderman & Huang, 2012). These two factors were also assessed in STQ questionnaires.

The original STQ questionnaire (Table 2.2) items were developed by a usability specialist with five years of research experience in human factors and usability. Both the

"transferability construct" and "context of use" were taken into consideration when creating the questionnaire.

To prevent response bias caused by users being automated in selecting scores without thinking about each statement, four questions (Q6, Q11, Q13, Q16) were altered to represent a negative opinion.

There are several scaling methods for questionnaires such as Likert scale, visual analog measures, and binary answers. A 7-point Likert scale was chosen as the scale system for STQ for the following reasons: first, Likert scales are the most widely used scale for current usability questionnaires (e.g. Brooke, 1996; Lewis, 2002; Lin, 1997), and it is proven to have excellent reliability and validity. Second, statistically, Likert scale provides a numerical scale that can differentiate users' perception with a 5-point or 7-point scale. Third, a systematic Likert scale makes it easy to compare the scores within or across questionnaires, which will assist in exploring the questionnaire and test the validity of the questionnaire. At last, chapter 3 will utilize the STQ score to create a single score for the UPMDS framework, which is easy to accomplish with the Likert scoring system.

Table 2.2 STQ Questionnaire Items.

Item	Content
1	Overall, I am satisfied with how easy it is to use the second software package after using the first software package.
2	It is simple to use the second software package after using the first software package.
3	I can quickly complete the task when using the second software package after using the first software package.
4	I can correctly complete all tasks when using the second software package after using the first software package.
5	I felt comfortable using both software packages and transferring between them.
6	I felt frustrated using the second software package after using the first software package.
7	I can quickly learn how to use the second software package after I changed from using the first software package to the second software package.
8	Using the first software package helped me learn to use the second software package faster.
9	The visual display and layout are generally consistent between the two software.
10	I felt more efficient using second software package after using the first software package.
11	The process of transferring to use the second software package after using the first software package is frustrating and makes me lost.
12	The second software package presents information that is consistent to the first software package.
13	I will easily confuse some functions in the second software package with the functions in the first software package.
14	Overall, I enjoy the experience of using both software packages
15	Overall, I am satisfied with using both software packages.
16	Overall, I'm frustrated and confused with using both software packages.

Participants

Participants were recruited from the university student population to participate in the validation experiment. Participant exclusion criteria were used for screening purpose. An online demographic survey (Appendix A) was given to the interested participants asking about their experience in using the designated software as well as their age, gender, etc. Participants who exhibited more than moderate frequency (around once per week) of using the study software (Adobe Photoshop and Adobe Acrobat) were excluded from the study. This online survey also served as the scheduling tool for qualified participants.

Altogether fifty-four participants qualified for and participated in the experiment. Literature has stated that the sample size should be larger than the number of questionnaire items (DeVillis, 1991; Kirakowski, 2000). The sample size is more than three times of the size of questionnaire items (16). The sample consisted of 20 females and 34 males, ranging from 19 to 43 years of age (M=23.04, SD=3.63). Participants were compensated either with $10/hour or with extra credit for a specific undergraduate level course.

Apparatus

Two sets of computer based software, Adobe Photoshop CS 5 and Adobe Acrobat Pro X, served as the experiment software. A desktop computer equipped with the Windows 7 operating system was used as the experiment platform. Both pieces of software were selected because they are commonly used in office environments and users often have to interact with both of them to complete a goal. These two sets of software can simulate a multi-device system which can often be encountered in daily work.

Camtasia Studio 7 screen capture software was used to record participants' screen activity during the data collection session. An audio recorder was used to record what the participants said during the experiment. This was used to obtain the think-aloud protocol from participants.

Variable Definition

The variables collected in the study included completion time per step (CTPS, calculated as the time between the start of each task to the end of each task, divided by the standard number of steps, recorded by analyzing video footage), error steps (calculated as the number of extra error steps beyond the standard number of steps for each task) and usability difficulties (calculated as the number of difficulties encountered when using the software, collected by analyzing verbal think aloud data). In addition, participants' perceived transferability between devices was collected using the 7-point Likert scale System Transferability Questionnaire developed in this study (STQ, Appendix B). Participants' perceived usability regarding each device was collected using Post-Study System Usability Questionnaire (PSSUQ, Appendix C). Participants' perceived overall satisfaction was collected using a single item questionnaire (Appendix D).

Procedure

Participants were scheduled to come to the Human Systems laboratory for the experiment after the online pre-screening survey. A brief introduction was given to participants regarding the objective of the study, what they need to do in the study, potential fatigue or discomfort, and compensation methods. Participants were informed

that they could leave at any time without penalty if they feel uncomfortable. An informed consent was provided to participant with all the above information included. The experimenter was available to answer any questions participants may have had. Consented participants signed the informed consent before starting the experiment. Four key points were repeated and stressed to make sure every participant understood them clearly:

1. This study is targeting the usability of the two software programs. Usability of any hardware or assisting software (e.g. keyboard, mouse, operating system, Camtasia, audio recording, etc.) is not of interest in this study.
2. This study is to test the usability of the software. It's not a test of users. So please relax and express your opinion regarding the usage of the software. Don't feel embarrassed just because you cannot figure out how to do the task. It's not your fault. It's our (the software's) fault.
3. Remember to use think aloud protocol when doing the tasks. You can talk out aloud what you are thinking and explain your method of attempting to complete the task, or illuminate any difficulties you encountered in the process.
4. An experimenter will sit beside you while you are completing all the tasks. He might remind you to use think aloud protocol. You are encouraged to ask questions, but the experimenter may not answer them.

After signing the consent form, participants were randomly assigned to start either from Adobe Acrobat Pro X or Adobe Photoshop CS 5. The order was counterbalanced. Before the experiment started, each participant watched a training video on the desktop computer regarding the use of the designated software. This was to help build base knowledge in the participant. The training involved six basic tasks. Each task required that the operator complete a series of operations to achieve one objective. The training video showed the screen activity of how to complete the task. Each training task lasted for around 45 seconds. An example of the training task (Appendix E) showed in training video would be:

> Using the "Image" menu, rotate the image 180 degrees.

After the training video, participants were allowed to ask any questions they have regarding the software. When no further questions were raised, the experimenter started the Camtasia screen capturing tool and audio recorder. A six-card task pack (six tasks total, one task on each card, shuffled before each participants) was provided to the participants. The tasks were similar to what the participants were showed in the training video. However, during the experiment, the task required that the participant complete an entire set of operations from start to stop. An example task would be (Appendix F):

> Open the file "Layer.psd". Add a new layer named "edit layer" with red color, dissolve mode, and 80% opacity. Save the image as its original name and close the image.

Participants were instructed to close the file after completing each task. The experimental task typically required 1-3 minutes to complete. Participants were also reminded that the thinking aloud protocol will be used in the experiment, which meant that they were asked to explain their approach of completing the task, state any difficulty

or problems they encountered while using the experimental software. The experimenter helped to ensure the think aloud protocol by reminding participants to "keep talking". Upon completion of the tasks using the first software, participants took a 5 minute break. The experimenter provided a PSSUQ for participant to fill out.

After the break, participants were directed to either Adobe Photoshop or Adobe Acrobat 7.0, whichever was not used in previous tasks. Again, participants watched a training video first. The experimenter was available to answer any questions after the training video. Then the experimenter started Camtasia screen capturing tool and audio recorder and provide the participant the task cards with six tasks in total. Participants were reminded to use the thinking aloud protocol during the experiment. Upon completion of the tasks, participants filled out questionnaires STQ, PSSUQ, and a single item questionnaire. After completion of the questionnaires, participants were compensated with $10 (for cash compensation participants only) and briefed about the experiment.

Data Analysis

Descriptive statistics about participants' task completion time, errors, number of usability difficulties, perceived usability and perceived transferability were presented.

Factor analysis (FA) is conducted on all the question items in the STQ to identify appropriate factors. FA is widely used as a statistical procedure to discover groups of related question items by examining the correlations among questionnaire items (DeVillis, 1991; Lewis, 2002; Netemeyer et al., 2003). A factor analysis is conducted in this study to identify the number of factors or latent variables that are representative of the underlying construct of usability.

A scree plot is used together with eigenvalue procedure to determined appropriate number of factors. Varimax-rotated patterns were used to identify questions items that corresponded to each of the factors. Questions that have low loadings or cross-loadings are removed.

The STQ is tested for its reliability and validity. Cronbach's Alpha is used to test for the internal reliability of the questionnaire. Three types of validity are tested: construct validity, criterion validity, and nomological validity. Test for the construct validity includes testing for the convergent validity and discriminant validity. The convergent validity tests whether the evidence from different sources gathered in different ways all indicated the same or similar meaning of a construct. The convergent validity is tested by calculating the average variance extracted (AVE). If the calculated AVE is greater than 0.5, the convergent validity is evident. Discriminant validity tests whether the construct can significantly differentiate with other constructs that it should theoretically be different from. Discriminant validity can be established by comparing the average shared variance (ASV) between each pair of construct against the minimum of the AVEs of these two construct (Fornell and Larker, 1981; Hair et al., 2010). If the average shared variance is lower than the minimum of their AVEs, then discriminant validity is proved.

Then criterion validity is tested to see if the outcome of STQ can match up with other measures or outcomes (the criteria) already held to be valid. Criterion validity can be tested using regression analysis. The overall satisfaction obtained using a single question survey serves as the dependent variable and the STQ factors serve as the independent variables.

Nomological validity tests whether the measures can correlate with the theoretically related constructs. Pearson correlation analysis is performed between STQ and related variables such as completion time per step, errors, usability difficulties, PSSUQ, and single item questionnaire score to investigate the nomological validity.

Results

Factor Analysis

To obtain a detailed insight of the factor structure of the questionnaire and refine question items, exploratory factor analysis was conducted using SAS 9.2 statistical software. The scree plot is showed in Figure 2.1. The plot indicates that the curve turns to a flat slope when the number of factors is greater than four. This effect is even more obvious when the number of factors is greater than six. This indicates that either four or six factors should be retained (Cattell, 1966). However, using the Kaiser-Guttman criterion (factors with eigenvalue greater than 1 are retained and factors with eigenvalue less than 1 are excluded), four factors should be retained (Kaiser, 1960). To make a decision on how many factors should be retained, the total variance explained by these factors was examined. Table 2.3 shows that with four factors, 76.44% of total variance is explained. With six factors, 85.09% of total variance is explained. As factors five and six each only help to explain less or equal to 5% of total variance, they are not significantly meaningful to explain the total construct. Therefore, four factors are selected as the number of factors on which to run the factor analysis.

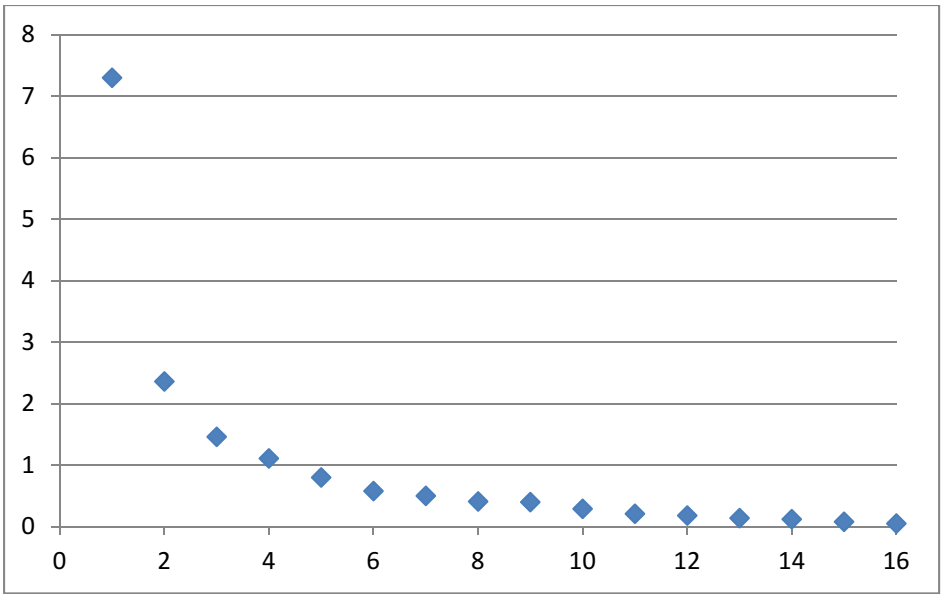


Figure 2.1 Scree Plot of Eigenvalues

Table 2.3 Eigenvalues and percentage of variance explained by each factor

Factors	Eigenvalue	Proportion%	Cumulative%
1	7.3	45.62	45.62
2	2.36	14.76	60.39
3	1.46	9.1	69.48
4	1.11	6.95	76.44
5	0.8	5.01	81.45
6	0.58	3.63	85.09
7	0.5	3.11	88.2
8	0.41	2.59	90.78
9	0.4	2.47	93.26
10	0.29	1.88	90.78
11	0.21	1.33	96.41
12	0.18	1.13	97.54
13	0.14	0.85	98.39
14	0.12	0.76	99.15
15	0.08	0.52	99.67
16	0.05	0.33	100.00

The varimax-rotated procedure is used to rotate the factor pattern with four factor groups. Results are shown in Table 2.4. According to the table, factor one includes the largest number of items with eight items (Q1, Q2, Q3, Q4, Q6, Q7, Q10, Q11), factor two has four items (Q5, Q14, Q15, Q16), factor three has three items (Q8, Q9, Q12), and factor four only has one item (Q13). All questions are significantly loaded on one of the factors (factor loadings greater than 0.5). No cross loadings greater than 0.50 is identified.

Table 2.4 Varimax-rotated factor pattern for the factor analysis using four factors

Item	Factor 1	Factor 2	Factor 3	Factor 4
Q3	**0.9**	0.05	0.17	0.09
Q1	**0.89**	0.19	0.03	-0.03
Q2	**0.89**	0.08	0.25	0.05
Q4	**0.81**	0.12	0.07	0.07
Q7	**0.81**	0.32	0.16	0.11
Q10	**0.77**	0.08	0.32	-0.08
Q6	**0.76**	0.47	-0.17	0.07
Q11	**0.71**	0.45	0	0.27
Q15	0.23	**0.83**	0.25	-0.04
Q14	0.28	**0.81**	0.16	-0.1
Q5	0.23	**0.78**	0.24	-0.24
Q16	-0.06	**0.77**	0.13	0.35
Q8	0.08	0.26	**0.78**	0.09
Q12	0.4	0.06	**0.72**	-0.04
Q9	0.03	0.19	**0.71**	0.06
Q13	0.18	-0.05	0.09	**0.92**

Note: Bold number in the table highlights factor loadings that exceeded .50

In the next step, the entire sixteen question items are sorted according to the varimax-rotated factor patterns of the factor analysis and are further explored to make meaningful explanation for the four factors. Four specific factors are identified. Factor

one explains the transfer experience from the users (TE). Factor two is the overall experience from the user regarding the use of both devices (OE). Factor three explains users' perception towards the consistency between two devices (CP). Factor four explains users' perception towards functionality of the devices (FP). After close examination and expert evaluation of the content of questionnaire items, question 8 is removed because its content does not fit into either of the factor groups.

As the question structure has changed, the factor analysis procedure is repeated without question 8. The results are slightly improved compared to the previous one, as expected. Table 2.5 shows that four factors are retained which explained 78.09% of total variance, which is higher than previous results (76.44%). All question items are significantly loaded on one of the factors. In addition, all of the question loadings except for one (Q12) are higher compared to the ones before question eight was removed (Table 2.6). This showed a positive improvement of the factor construct when question 8 is removed.

Table 2.5 Eigenvalues and percentage of variance explained after removing Q8

Factors	Eigenvalue	Proportion%	Cumulative%
1	7.15	47.64	47.64
2	2.25	15.03	62.67
3	1.2	8.01	70.68
4	1.11	7.41	78.09
5	0.75	5	83.08
6	0.56	3.74	86.82
7	0.47	3.13	89.96
8	0.4	2.66	92.62
9	0.29	1.96	94.58
10	0.22	1.44	96.01
11	0.18	1.2	97.22
12	0.15	1.01	98.23
13	0.17	0.85	99.07
14	0.08	0.55	99.63
15	0.06	0.37	100.00

Table 2.6 Varimax-rotated factor pattern with four factors after removing Q 8.

Item	Factor 1	Factor 2	Factor 3	Factor 4
Q3	**0.91**	-0.05	0.16	0.09
Q1	**0.89**	0.19	0.05	-0.02
Q2	**0.89**	0.09	0.27	0.06
Q7	**0.81**	0.33	0.12	0.11
Q4	**0.81**	0.12	0.1	0.08
Q10	**0.78**	0.11	0.26	-0.09
Q6	**0.76**	0.46	-0.22	0.05
Q11	**0.72**	0.45	-0.1	0.25
Q15	0.22	**0.85**	0.22	-0.03
Q14	0.28	**0.82**	0.15	-0.09
Q5	0.23	**0.79**	0.2	-0.24
Q16	-0.04	**0.78**	0	0.33
Q9	0.01	0.24	**0.8**	0.11
Q12	0.41	0.12	**0.69**	-0.02
Q13	0.17	-0.04	0.09	**0.93**

Note: Bold number in the table highlights factor loadings that exceeded .50

Table 2.7 shows the final summary of the factor structure of STQ as well as the average scores of each factor. FP exhibits the highest score (M=5.2, SD=1.56, out of 7), followed by OE (M=5.11, SD=1.72) and TE (M=4.28, SD=1.96). CP has the lowest score (M=3.69, SD=1.61) indicating user's frustration with the consistency between two devices.

Table 2.7 Factor Arrangement and Average Scores

Factor Group	Factor Name	Survey Items	Average Score
1	Transfer Experience (TE)	Q1, Q2, Q3, Q4, Q6, Q7, Q10, Q11	4.28
2	Overall Experience (OE)	Q5, Q14, Q15, Q16	5.11
3	Consistency Perception (CP)	Q9, Q12	3.69
4	Functionality Perception (FP)	Q13	5.20

Test for Reliability

The reliability of a measure is the extent to which it is free from random error. To estimate the reliability of the questionnaire, Cronbach's coefficient alpha (Cronbach, 1951) is used. Cronbach's coefficient alpha is a widely used statistic to test internal reliability in questionnaire development and validation process. Cronbach's alpha estimates how closely related a set of items are as a group. The coefficient can be calculated by:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right) \quad (2.1)$$

Where:

K= number of items,
$\sigma_{Y_i}^2$= variance of item Y_i, and
σ_X^2= variance of total questionnaire scores

Table 2.8 shows the Cronbach's alpha values for the overall and each factor groups in the questionnaire. The overall Cronbach's α for the entire STQ is 0.91 which is much higher than the normally acceptable level 0.70. Cronbach's α for TE exhibits the highest value at 0.95, followed by OE (α=0.87), and CP (α=0.68). Since FP only has one question item, the Cronbach's alpha is not applicable for this factor group. All the rest of groups presented medium to high internal reliability which is at acceptable levels (Table 2.8).

Table 2.8 Cronbach's Alpha Values for Each Factor Group and All Items.

Factor Group	Factor Characteristics	Cronbach's α
1	Transfer Experience (TE)	0.95
2	Overall Experience (OE)	0.87
3	Consistency Perception (CP)	0.68
4	Functionality Perception (FP)	N/A
Overall		0.91

Test for Validity

Validity of a measure is the extent to which it measures what it is supposed to measure, as compared to reliability (the extent of consistency). Three types of validity are usually tested: construct validity, criterion validity, and nomological validity. Construct validity refers to the extent that the questionnaire construct do actually measure what they are supposed to measure. Construct validity can be evidenced when both convergent validity and discriminant validity are proved. Criterion validity refers to the extent to which the factors measured can be proved with other measures or outcomes already held to be valid. Nomological validity is a type of validity in which a measure should

correlates positively in the theoretically predicted way with measures of different but related constructs (Yang, 2005).

Descriptive Statistics

Before testing the validity of the STQ, descriptive statistics are provided for the variables measured in this study (Table 2.9). The STQ scores are calculated with reverse questions transformed back to a normal scale. The results show that participants reported less than one usability difficulty in each task, but on average made more than 13 error steps in completing the tasks. On average, it took around 9.63 seconds for participants to complete one step.

Table 2.9 Descriptive Statistics of Study Variables

Variables	Mean	SD	Plot
Average CTPS (s)	9.63	2.65	
Average Errors/Task	13.44	6.63	
Average UX Difficulties/Task	0.21	0.21	
STQ Scores (1-7)	4.16	1.11	
PSSUQ Scores (1-7)	4.51	1.23	
Single Item Q Score (1-7)	5.33	0.95	

Among the three questionnaire scores, the single items score that asked about participants' overall satisfaction scores the highest (M=5.33, SD=0.95), followed by PSSUQ scores (M=4.51, SD=1.23) and STQ scores (M=4.16, SD=1.11). Since STQ is the primary questionnaire that is investigated, a breakdown of the descriptive statistics of the score of each question item from STQ is provided in Table 2.10.

Table 2.10 Descriptive Statistics of the Question Items form STQ by Factor Groups

Factor Group	Question #	Mean	SD	Maximum	Minimum
TE	Q1	4.39	2.05	7	1
	Q2	4.26	2.01	7	1
	Q3	4.13	1.72	7	1
	Q4	3.94	1.95	7	1
	Q6	4.61	2.20	7	1
	Q7	4.34	1.64	7	1
	Q10	3.63	1.85	7	1
	Q11	4.94	1.84	7	1
OE	Q5	4.70	1.21	7	2
	Q14	4.91	1.36	7	1
	Q15	4.96	1.20	7	2
	Q16	5.85	1.09	7	2
CP	Q9	3.80	1.68	7	1
	Q12	3.59	1.55	7	1
FP	Q13	5.20	1.56	7	1
Not Included	Q8	3.24	1.65	7	1

Among the 16 question items, Q16 (M=5.85) has the highest average scores, followed by Q13 (M=5.20), and Q 15 (M=4.96). Q8 (M=3.24) exhibits the lowest scores. Within each question, the maximum scores are all 7. The minimum scores are all 1 except for three questions (Q5, Q15, and Q16), which scores 2 as the minimum score.

Convergent Validity

Convergent validity is one type of construct validity. Convergent validity can be evidenced when the measures from different items (questions) from the same construct (factor groups) indicate same or similar meanings (converge). To test the convergent validity, the average variance extracted (AVE) is calculated for each construct. It is believed that convergent validity is proved when AVE is greater than 0.5 (Fornell and Larcker, 1981). AVE can be calculated by the following equation:

$$AVE = \frac{\sum SL^2}{\sum SL^2 + \sum Err} \quad (2.2)$$

Where:

SL^2= the standard loadings square
Err= indicator measurement error

Results show that group one (TE) exhibits an AVE of 0.68. Group two (OE) has an AVE of 0.66. Group three (CP) has an AVE of 0.56. At last, group four (FP) has an AVE of 0.86. All four groups exhibit AVEs greater than 0.5, suggesting that the convergent validity of the four-group construct of the questionnaire was evidenced (Table 2.11).

Table 2.11 AVE Values for Each Factor Group.

Factor Group	Factor Characteristics	AVE
1	Transfer Experience (TE)	0.68
2	Overall Experience (OE)	0.66
3	Consistency Perception (CP)	0.56
4	Functionality Perception (FP)	0.86

Discriminant Validity

Discriminant validity is another type of construct validity. As opposed to convergent validity, discriminant validity implies that the measures from conceptually different constructs are truly uncorrelated with (discriminant from) each other. Discriminant validity can be established by comparing the average shared variance (ASV) between each pair of construct against the minimum of the AVEs of these two construct (Fornell and Larker, 1981; Hair et al., 2010). If the average shared variance is

lower than the minimum of their AVEs, then discriminant validity is proved. ASV can be calculated by the following equation:

$$ASV = \frac{\sum COV(i,j)}{n-1} \quad (2.3)$$

Where:
$COV(i,j)$= the covariance of each possible combination of questions in-between two factor constructs.
n = the total number of factor constructs.

ASV is calculated for each factor construct (Table 2.12). The ASV value is compared with the AVE results obtained above. The ASV values of TE, OE, and FP are all lower than their AVE values. Only CP shows a little higher ASV value than AVE values. This indicates that the CP group may not be sufficiently discriminated from the rest of factors. The rest of ASV results support the discriminant validity of the factor construct. In addition, in the factor analysis, no significant cross-loadings are identified in any of the factor constructs. This indicates that each question item can be clearly discriminated from questions in other factors. This also supports the discriminant validity of the factor construct.

Table 2.12 ASV and AVE Values for Each Factor Group.

Factor Group	Factor Characteristics	AVE	ASV
1	Transfer Experience (TE)	0.68	0.65
2	Overall Experience (OE)	0.66	0.27
3	Consistency Perception (CP)	0.56	0.61
4	Functionality Perception (FP)	0.86	0.16

Criterion Validity

Criterion validity (in this case concurrent validity) refers to the extent to which the factors measured can be used to indicate a pre-specified criterion. Criterion validity is often tested by examining the correlation between measures of various factors and the specific criterion (Lewis, 1995; Netemeyer et al., 2003). Researchers also use regression analysis to examine the predictive ability of the different measures (Yang et al., 2005).

In this study, a regression analysis was performed to test the criterion validity. The single item questionnaire score is used as the dependent variable and serve as the criterion. The mean score of the four derived factor constructs of STQ is used as independent variables. The overall regression model is significant ($F(4,53)= 18.73$, $p<0.0001$, $R^2=0.60$). The results indicate that the criterion validity is evidenced. Regression results are shown in Table 2.13.

Table 2.13 Regression Analysis Results

Variable	DF	Parameter Estimate	t-Value	p-Value
Intercept	1	1.35	2.62	0.01
TE	1	-0.02	-0.30	0.76
OE	1	0.69	7.06	<0.0001
CP	1	0.06	0.85	0.40
FP	1	0.06	1.06	0.29

Nomological Validity

Nomological validity tests whether the measures can correlate with the theoretically related constructs. Following the definition to test the nomological validity of STQ, a Pearson correlation analysis between STQ and related variables such as

completion time per step, errors, usability difficulties, PSSUQ, and single item questionnaire score was performed.

The first set of Pearson correlation coefficient is shown in Table 2.14. All variables, except for STQ and single item questionnaire, were averaged over the whole experiment. The results show that the STQ is significantly and highly correlated with both the PSSUQ (r=0.63, $p<0.0001$) and the single item questionnaire (r=0.53, $p<0.0001$). However, STQ is not significantly correlated with completion time per step, errors, or usability difficulty. This finding supports the nomological validity of the STQ. In addition, PSSUQ is significantly correlated with all measures. PSSUQ is positively and highly correlated with the STQ and the single item questionnaire (r=0.61, $p<0.0001$), but has a low negative correlation with completion time per step (r=-0.27, $p=0.05$), errors (r=-0.33, $p=0.02$), and usability difficulties (r=0.26, $p=0.06$). Additionally, completion time per step and errors are positively correlated with each other (r=0.53, $p<0.0001$).

Table 2.14 Pearson correlation score of STQ and other variables averaged throughout experiment

	STQ	Single Q	PSSUQ Mean	CTPS Mean	Errors Mean	UX Difficulty Mean
STQ	1					
Single Q	0.53**	1				
PSSUQ Mean	0.63**	0.61**	1			
CTPS Mean	-0.01	-0.12	-0.27*	1		
Errors Mean	-0.22	-0.04	-0.33*	0.53**	1	
UX Difficulty Mean	-0.07	-0.07	-0.26*	0.16	0.2	1

Note:*numbers indicate significant correlation at α=0.05 level.
**numbers indicate significant correlation at α=0.001 level.

The second set of Pearson correlation coefficient is shown in Table 2.15. All variables except for the STQ and the single item questionnaire are calculated as the difference between second software and first software (e.g. $PSSUQ_b$ score – $PSSUQ_a$ score). This approach is used to examine the performance and perception difference after the transfer process. The results show that the STQ difference has an excellent correlation with the PSSUQ difference (r=0.72, p<0.0001). The STQ difference is also significantly correlated with usability difference (r=-0.35, p=0.01). These results further supported the conclusion that nomological validity was evidenced. In addition, the performance measures (CTPS difference, error difference, and usability difficulty difference) are all mildly correlated with each other. Additionally, the PSSUQ difference is significantly correlated with usability difficulty difference (r=-0.52, p<0.0001).

Table 2.15 Pearson Correlation Score of STQ and Other Variable Difference

	STQ	PSSUQ Diff	CTPS Diff	Errors Diff	UX Difficulty Diff
STQ	1				
PSSUQ Diff	0.72**	1			
CTPS Diff	-0.16	-0.13	1		
Errors Diff	-0.06	-0.15	0.59**	1	
UX Difficulty Diff	-0.35*	-0.52**	0.37*	0.37*	1

Note: *numbers indicate significant correlation at α=0.05 level.
**numbers indicate significant correlation at α=0.001 level.

Discussion

Three research questions were aimed to be addressed in this study:

- Can we create a questionnaire that can effectively capture users' perception regarding the transferability between devices?
- What aspects/facet of the transferability does this questionnaire help to explain?
- Would this questionnaire be reliable and valid to be used in a real-world scenario?

The following section will address these three research questions separately.

Questionnaire Structure

The discussion of the first research question will be dependent on the results of research question two and three. Therefore, research question two is addressed first.

The system transferability questionnaire was original developed with sixteen question items. Exploratory factor analysis identifies four factor structures based on statistical procedure: Transfer experience (TE), overall experience (OE), consistency

perception (CP), and functionality perception (FP). To explore the feasibility of other factor structures, three-factor structure, five-factor structure, and six-factor structure are further examined (Appendix G).

Compared to a four-factor structure, a three-factor structure removes question 13 as it does not significantly load on any factor group. However, question 13 obtained information regarding the functionality of two different software packages, which is an important aspect. In addition, this question alone adds to around 8% of the total variance explained. Therefore, a three-factor structure is not deemed acceptable.

A Five-factor structure separates question 8 and question 12 while keeps the rest the same as the four-facture structure. Since question 8 and question 12 both stress the consistency between the two software packages, these two questions are essentially belong to the same factor group but explain slightly different facet (visual display and information presentation). Thus, these two question items are retained in the same factors, eliminating the fifth factor.

Six-factor structure result is based on the five-factor result. Question 16 is further separated from Q5, Q14, and Q15, which is not supported by question examination. In addition, cross loadings are present when more factor groups were retained. Therefore, a six-factor structure is not adopted.

With question 8 removed after further examination, a four factor group structure is finally confirmed.

Transfer Experience (TE)

Eight question items are retained in the first factor group: Q1, Q2, Q3, Q4, Q6, Q7, Q10, and Q11. This group is named transfer experience as most of the questions in this group are created to elicit users' perception regarding their experience when transferring between two devices. Specific perception includes satisfaction, easiness, efficiency, effectiveness, frustration, learning, etc. The key word in the questions in this group is "after" (e.g. I felt frustrated using the second software after using the first software), which stresses the transfer between devices. This group receives an average rating of 4.28, which is an above average score. There may be some transfer issues in the system but on average users found it acceptable.

Overall Experience (OE)

Four question items are retained in the second factor group: Q5, Q14, Q15, and Q16. All the questions in this group asked the users' perception of overall experience using both devices. Therefore, this group is named overall experience. Specific perception includes enjoyment, satisfaction, frustration, and comfortableness. As opposed to TE, this group has one common key word, "overall" (e.g. Overall, I am satisfied with using both Software) indicating that this factor was formed to elicit users' perception regarding the overall experience using both device. An average score of 5.11 is obtained in the study for OE. Users appear to be fairly satisfied with the overall experience using both devices.

Consistency Perception (CP)

This group consists of two question items: Q9 and Q12. Both of these two questions ask about users' perception on consistency between two devices. Question 9 focuses on visual display while question 12 focuses on information presentation. Both questions involved the key word: "consistent", which was used to name this group. As consistency (transparency) is found to be an important factor impacting user performance during transfer (Huang et al., 2012), this factor would serve as a key facet of the system transferability questionnaire. In the validation study, an average score of 3.69 is obtained, indicating a poor consistency between devices. Users seem to have a lot of issues regarding the consistency using two software packages. This also means that when redesigning the system, consistency issues should be the first to be addressed.

Functionality Perception (FP)

This group only incorporates only one question: Q13. However, it helps to explain the functionality of both devices. An average score of 5.20 was obtained in the study indicating a fairly good functionality of both devices.

Overall, the four groups established are meaningful. The validation study helps to make sense of these factors and the results were expected. Users gave highest ratings for OE and FP because the testing platforms (Adobe Acrobat and Adobe Photoshop) are commercial software and are available in the market. Their functionalities were well designed and constructed to meet the needs of majority of users including expert users. However, TE exhibits lower scores because users identify transfer issues in the transfer process. Although these two software platform s are developed by the same company, they were designed to address different user objectives. Adobe Acrobat was designed for

document organization and editing, thus was operated more like word-processing software. Adobe Photoshop was designed for image processing and editing, which incorporated a layout and operating style that was unfamiliar to most of the users. A lot of transfer issues and difficulties were expected when users transferring between these two software. This also explains the low score for CP, as inconsistency is one of the major reasons leading to the transfer difficulties of the users.

Factor CP has two question items and factor FP has only one question. This categorization may affect the internal reliability and validity of the STQ. Using traditional usability questionnaires (PSSUQ, CSUQ) as a guideline, three to five question items are appropriate to measure a factor within the usability construct. Additional question items will be added to the CP and FP factors in a future study in order to explore this relationship further.

Therefore, the research question is answered by the above analysis. To directly address the research question: *The STQ was developed with 15 question items that help to explain a total of four factors: transfer experience, overall experience, consistency perception, and functionality perception.*

Questionnaire Reliability and Validity

The second objective of this study is to examine the reliability, construct validity, and criterion validity of the system transferability questionnaire using the validation study.

Reliability

The overall questionnaire show an excellent internal reliability (Cronbach's α=0.91). Each factor group also had medium to high reliability. Consistency perception exhibited the lowest reliability (α=0.68). Two possible reasons were identified. First, the CP factor only consists of two question items. Any minor variation between these two questions may lead to a low Cronbach's alpha value. Second, the number of samples collected in this study is limited. The variance from samples may cause a low Cronbach's alpha value. Since it is close to the acceptable level (α=0.70), it is considered marginally acceptable. Therefore, the STQ and its factor groups meet the internal reliability standard.

Possible ways to improve the internal reliability of the STQ include adding more question items to the factor group CP and FP, testing STQ on more participants, and slightly revising the question items.

Construct Validity

The construct validity of STQ is assessed using three criteria: convergent validity, discriminant validity, and nomological validity. All three criteria indicated an evidenced validity of STQ. The convergent validity indicates that within each factor group, the questions items correlate with each other to explain the factor, which supports our decision to group them together. The discriminant validity indicates that the question items in each factor group can be sufficiently distinguish against question items in other factor groups, supporting our categorization of the four factor groups.

Nomonlogical validity tests the STQ as a whole construct with other theoretically related measures. As a questionnaire that measures one aspect of usability, STQ is hypothesized to be positively correlated with PSSUQ and single item questionnaire. The

STQ is hypothesized to be uncorrelated with performance measures such as completion time per step, errors and usability difficulties. These objective measures capture the usability within each device, which is the reason why they are highly correlated with the PSSUQ. However, the STQ measures the transferability between devices, which represents a different construct of usability. The results support both assumptions. STQ is not only positively correlated with average PSSUQ scores (representing the average experience using two devices) but is also positively correlated with the PSSUQ score difference (representing the transfer impact between devices). In addition, the STQ is also positively correlated with the single item questionnaire. This supports the statement that STQ not only helps to explain some aspects of usability, but also explains users' experience and perception toward the transfer process.

The performance measures are not significantly correlated with STQ, which is expected. This finding indicated that performance measures may represent other constructs of usability that differ from the construct measured by the STQ. This result corresponds well with the claim by a lot of literature that subjective and objective measures capture difficult constructs of the usability (Bommer et al., 1995; Yeh and Wickens, 1988). This also serves as a theoretical and empirical rationale which leads us to develop the UPMDS framework to measure usability in Chapter III. With the above analysis, the construct validity of STQ is sufficiently evidenced.

Criterion Validity

The criterion validity is evidenced with a significant regression model. The four factor groups help to measure the overall satisfaction indicating that the criterion was set up appropriately.

The above analysis all help to address the research question. To answer the research question directly: *A validation study based on a real life scenario was conducted. The STQ and its factor groups proved to be a reliable and valid tool to measure users' perception towards the transfer experience between using two devices.*

System Transferability Questionnaire (STQ)

With the above two research questions answered, we can confidently state that: The System Transferability Questionnaire (STQ) was developed as a valid tool to effectively capture users' perception regarding the transferability between devices.

Table 2.16 Reordered STQ question items based on factor groups

Factor Group	New Item#	Content
TE	Q1	Overall, I am satisfied with how easy it is to use the second software package after using the first software package.
	Q2	It is simple to use the second software package after using the first software package.
	Q3	I can quickly complete the task when using the second software package after using the first software package.
	Q4	I can correctly complete all tasks when using the second software package after using the first software package.
	Q5	I felt frustrated using the second software package after using the first software package.
	Q6	I can quickly learn how to use the second software package after I changed from using the first software package to the second software package.
	Q7	I felt more efficient using second software package after using the first software package.
	Q8	The process of transferring to use the second software package after using the first software package is frustrating and makes me lost.
OE	Q9	I felt comfortable using both software packages and transferring between them.
	Q10	Overall, I enjoy the experience of using both software packages
	Q11	Overall, I am satisfied with using both software packages.
	Q12	Overall, I'm frustrated and confused with using both software packages.
CP	Q13	The visual display and layout are generally consistent between the two software.
	Q14	The second software package presents information that is consistent to the first software package.
FP	Q15	I will easily confuse some functions in the second software package with the functions in the first software package.

A new set of STQ items are presented in Table 2.16 showed above. These new set of items are grouped and reordered according to factors.

The STQ is designed specifically to obtain users' perception towards transferability of a multiple-device system. It can be used together with other usability questionnaires to gain a more comprehensive view of the system usability. To better utilize the questionnaire, it should be administrated by a usability specialist. It should be provided to the users after they have used both devices. Users should be informed that this is the test on the devices instead of a test on them. Users are also allowed to provide extra comments regarding any items or select "N/A" for items that are not applicable to their experience. When used to measure different platform, key words should be adjusted according to the specific platform (e.g. software, machines, devices, tools, etc.). When calculating the scores, inverse question items should be altered back in scale and an average score will be calculated as the overall STQ score. STQ includes four sub-factors: transfer experience, overall experience, consistency perception, and functionality perception. Scoring of those sub-factors will help us understand the details lying below the overall transferability score. The reordered STQ and administration details are provided in Appendix H.

Conclusion

In this study, a system transferability questionnaire is developed with 15 question items and four factor groups. This questionnaire tool is validated in a software usability study and proves to be effective in measuring the system transferability and users' perception towards the transfer process. It fills the literature gap that no subjective tool can be used to assess the transferability within a multiple device system. To a wider

perspective, it can be generalized and used in any multiple-device system in which the transferability between devices needs to be measured. Specific scenarios include:

1. Two devices that are distinct regarding the interface, but both have to be used to achieve a specific goal.
2. Two of the same product with one being the previous version and the other upgraded version.
3. The same online service that can be accessed through different devices.

This study also has several limitations. First, the expertise of users may affect the transferability score. This study only excluded high expertise participants. But the effect of expertise is still unknown. Second, this study is based on the computer software. A wider selection of application platforms would be helpful to prove the generalization of the STQ. Third, the sensitivity of the STQ was not tested due to the experiment design. Fourth, more question items will be added to the factor CP and FP to improve these two factor groups and the overall STQ. At last, the STQ was developed as a subjective measure of the transferability. In future work, objective measures of transferability, such as change of completion time and change of errors, should be explored and incorporated.

References

Anderson, J. R. (1983 a). The architecture of cognition. Harvard University Press, Cambridge.

Anderson, J. R. (1993 b). Rules of the mind. Erlbaum, Hillsdale.

Anderson, J. R. (1995). Learning and Memory. New York: John Wiley & Sons, Inc

Avouris, N. M. (2001). An introduction to software usability. In Proceeding of 8th Panhellenic Conference on Informatics, Workshop on Software Usability, Nicosia, 514-522.

Baldwin, T.T. and Ford, J.K., (1988). Transfer of training: A review and directions for future research. *Personal Psychology 41 (1)*, 63.

Bangor,A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24(6):*574–594.

Bennett, J. L. (1979). The commercial impact of usability in interactive systems. In B. Shackel (Ed.), Man/computer communication: Infotech state of the art report (Vol. 2, pp. 1-17). Maidenhead: Infotech International.

Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), Usability Evaluation in Industry. London: Taylor & Francis.

Cattell, R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Card, S., Moran, T. and Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum.

Chandra, C., Kamrani, A.K., 2003. Knowledge management for consumer-focused product design. *Journal of Intelligent Manufacturing 14 (6),* 557e580.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Dahl,D.W.,&Moreau, P. (2002). The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research, 39*, 47–60.

DeVillis, R. F. (1991). Scale development: Theory and applications. Newbury Park, CA: Sage.

Ellis, H. C. (1965). The Transfer of Learning. New York: MacMillan

Ford, J. K., & Weissbein, D.A. (1997). Transfer of training: An update review and analysis. *Performance Improvement Quarterly, 10,* 22-41.

Fornell, C. and David G. L. (1981). "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research, 18(1),* 39-50.

Frese, M., Brodbeck, F., Heinbokel, T., Mooser, C., Schleiffenbaum, E., & Thiemann, P. (1991). Errors in training computer skills: On the positive function of errors. Human-Computer Interaction, 6, 77–93.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Hair, J., Black, W., Babin, B., and Anderson, R. (2010). Multivariate data analysis (7th ed.): Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Hsieh, L. F., & Chen, S. K. (2005). Incorporating voice of the consumer: Does it really work? *Industrial Management & Data Systems, 105,* 769–785.

Haskell, R. E. (2000). Transfer of Learning: Cognition, Instruction, and Reasoning. Academic Press.

Hornbeck, K (2006). Current practice in measuring usability: challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*, 79-102.

Huang, Y. & Strawderman, L. (2011). Introducing a New Usability Framework for Analyzing Usability in a Multiple-device System. *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting, 55 (1),* 1696-1700.

Huang, Y., Strawderman, L., & Murray, D. (2012). Investigating the Impact of Task Change Type and Transparency on Transfer of Learning. *International Journal of Human-Computer Interaction, 28*, 61-71.

ISO/IEC. 9241. (1998). Ergonomic requirements for office work with visual display terminals (VDT)s. ISO/IEC 9241-14: 1998 (E),.

Jeffries, R. & Desurvire, H. (1992). Usability testing vs. heuristic evaluation: Was there a contest? SIGCHI Bulletin, 24,4, 39-41

Lewis, J. R. (2002). Psychometric evaluation of the pssuq using data from five years of usability studies. International Journal of Human-Computer Interaction, 14(3-4), 463-488.

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. Behaviour & Information Technology, 16(4/5), 267-278.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational Psychology Measurement, 20,* 141– 151.

Kantner, L., & Rosenbaum, S. (1997). Usability studies of WWW sites: heuristic evaluation vs. laboratory testing. In: Proceedings of ACM Special Interest Group for Documentation (SIGDOC). Snowbird, UT, pp. 153-160.

Kirakowski, J. (1994). The use of questionnaire methods for usability assessment (unpublished manuscript). http://sumi.ucc.ie/sumipapp.html

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of websites. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.* Chicago, 424–428.

Kirakowski, J. (2000). Questionnaires in usability engineering: A list of frequently askedquestions [HTML]. Retrieved 11/26, 2003, from the World Wide Web: http://www.ucc.ie/hfrg/resources/qfaq1.html

Konradt, U., Wandke, H., Balazs, B., & Christophersen, T. (2003). Usability in online shops: Scale construction, validation and the influence on the buyers' intention and decision. *Behavior & Information Technology, 22(3)*, 165-174.

Lee, D., Rhee, Y., & Dunham, R. B. (2009). The roll of organizational and individual characteristics in technology acceptance. *International Journal of Human–Computer Interaction, 25,* 623–646.

Lewis, C., & Rieman, J. (1994). Task-Centred User Interface Design: A Practical Introduction. Available at http://dcti.iscte.pt/cgm/web/TCUID_PI.pdf

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of Educational Psychology (pp.47–62). New York: MacMillan Library Reference Usa.

Mullens, M. A., & Armacost, R. L. (1995). A two stage approach to concept selection using the analytic hierarchy process. 2(3), 199-208.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). Scaling procedures: Issues and applications. Thousand Oaks, CA: Sage Publications, Inc.

Nielsen, J. (1993). Usability engineering. Cambridge, MA: Academic Press.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In CHI 194 Conference Proceedings, (pp. 152-158). New York: ACM Press.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. Proc. ACM INTERCHI'93 Conf, 214-221.

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. Proc. ACM CHI'90 Conf., 249 - 256.

Olson, J. R., & Olson, G. M. (1990). The growth of cognitive modeling in human-computer interaction since GOMS. *Human-Computer Interaction, 5,* 221–265.

Payne, S. J. & Green, T. R. G. (1986) Task-Action Grammars: a model of the mental representation of task languages. Human-Computer Interaction, 2, 93-133.

Polson, P. G. (1987) A quantitative theory of human-computer interaction. In J. M. Carroll (Ed.) Interfacing thought: Cognitive Aspects of Human- Computer Interaction. Cambridge, MA: Bradford Books/MIT Press.

Polson, P. G. (1988) Transfer and Retention. In R. Guindon (Ed.). Cognitive Science and its application for human-computer interaction. Hillsdale, NJ: Erlbaum.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. Educational Psychologist, 24, 113–142.

Schneiderman, B. (1992). Designing the User Interface: Strategies for Effective Human-Computer Interaction (2nd ed.), Reading, MA: Addison-Wesley.

Singley, M. K., & Andersen, J. R. (1989). The transfer of cognitive skill. Cambridge, MA: Harvard University Press.

Sternberg, R. J., & Frensch, P. A. (1993). Mechanisms of transfer. In D. K. Detterman & R. J. Sternberg (Eds.), Transfer on Trial: Intelligence, Cognition, and Instruction (pp. 25–38). Stamford, CT: Ablex Publishing Corp.

Thorndike, E. L. & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 9*, 374-382

Tziner, A., Haccoun, R. R., & Kadish, A. (1991). Personal and situational characteristics influencing the effectiveness of transfer of training improvement strategies. *Journal of Occupational Psychology, 64*, 167–177.

van Veenendaal, E. (1998). Questionnaire based usability testing. In Proceeding of European Software Quality Week, Brussels.

Yang, Z., Cai, S., Zhou, Z., and Zhou, N. (2005). Development and validation of an instrument to measure user perceived service quality of information presenting Web portals. Information and Management, 42(4), 575-589.

Zaharias, P., & Poylymenakou, A. (2009). Developing a usability evaluation method for e-learning applications: beyond functional usability. *International Journal of Human-Computer Interaction, 25,* 75–98.

CHAPTER III

TRANSFERABILITY, SATISFACTION, AND USER PERFORMANCE, A TOTAL SYSTEM USABILTY SCORE FOR MULTIPLE-DEVICE SYSTEMS

Introduction

The UPMDS framework is introduced in Chapter I to conceptualize the usability model. As a key objective of the UPMDS framework, it should be capable of providing an effective evaluation tool to measure the overall system usability. Usability practitioners should be able to customize the tool and input different usability measures such as completion time, errors, usability difficulties, subjective satisfaction and transferability and obtain an overall system usability score. Although various usability evaluation tools utilize different approaches such as heuristics analysis, think aloud protocol, performance measures and usability questionnaires, a single usability score is useful in that it not only provide an easy method to interpret and benchmark outcome for the usability practitioners and product designers, but also allows for further data analysis such as regression analysis and hypothesis testing to explore the impact of other causal factors on system usability.

Another reason for a single score usability evaluation outcome is the divergent opinion regarding objective and subjective results. It is believed that both subjective and objective measures need to be collected to evaluate usability. When subjective and objective measures of usability agree, it's easy to choose one for usability evaluation.

When they disagree, the choice between subjective and objective measures may depend on the situation of tasks or the objective of the measurement (Lewis, 1995). A single score evaluation tool would be able to combine both subjective and objective measures. Proper weighting mechanism would utilize the characteristics of the data set (using principal component analysis) to determine the weight of different measures. Compared to Lewis (1995)'s traditional method, this would be more quantitative and mathematically grounded

In addition, chapter II identifies STQ as an effective tool to measure subjective transferability. STQ is highly correlated with PSSUQ and single item questionnaire, but has low correlation with objective measures such as completion time per step (CTPS), errors, and usability difficulties. It is possible that objective approaches were measuring some constructs of usability that were different from what subjective approaches were measuring. A method to consolidate both measurement approaches is critical to provide a valid measure to evaluate overall system usability.

Background and Literature Review

Usability Frameworks

Three major challenges remain in the current usability literature. First, how to find an appropriate usability framework that categorizes different usability attributes and measures them. Second, whether subjective measurements or objective measurements or both should be adopted in usability studies. Third, how to adjust the usability framework to measure usability in various application contexts such as mobile devices, home technology, and multiple-device systems.

Ever since Nielsen (1993) identified usability attributes as efficiency, learnability, memorability, errors, and satisfaction, many studies have been trying to construct a comprehensive, yet universally applicable model of usability. The Metrics for Usability Standards in Computing (MUSiC; Bevan, 1995) was a model developed for software usability evaluation. It provided measures for user performance, such as task effectiveness, temporal efficiency, and length of productive period. With the addition of Software Usability Measurement Inventory (SUMI; Kirakowski and Corbett, 1993), this model could also provide measures of global user satisfaction as well as usability attributes such as effectiveness, efficiency, helpfulness, control, and learnability.

John and Kieras (1996) used the Goals, Operators, Methods, and Selection rules (GOMS) to model a particular task within a software system. Initially developed as a human information processing and behavior model, the GOMS model is capable of predicting task performance time based on a hierarchical structure in the GOMS framework. However, GOMS does not take into account user unpredictability, such as errors, fatigue, and learning effect. The model is a prediction for expert user performances in ideal situations. Real world evaluation is needed to validate the prediction.

In recent studies, researchers incorporated attractiveness or affection as an additional usability attribute (Sutcliffe, 2002; De Angeli et.al, 2006; Thuring & Mahlke, 2007). Users' subjective perceptions were no longer limited to satisfaction regarding the functional performance. Attractiveness and user emotions (Sutcliffe, 2002; De Angeli et.al, 2006; Thuring & Mahlke, 2007) were studied as an indicator to their preference over different interfaces. Seffah et al. (2006), instead, combined several usability models

and proposed a comprehensive model QUIM (Quality in Use Integrated Measurement) that incorporated attributes such as efficiency, effectiveness, safety, trustfulness and accessibility. A summary of the differences and commonalities of existing usability studies is outlined in Table 3.1.

Whether to use subjective or objective measures to evaluate usability has been the focus of usability studies. It is recognized that both measures are necessary because they may lead to different conclusions regarding the usability of an interface. Studies also suggested that these measures capture different aspects of user performance (Bommer et al., 1995; Yeh and Wickens, 1988). A major challenge, as put forward by Hornbak (2006, p. 92), is to "develop subjective measures for aspects of quality-in-use that are currently mainly measured by objective measures, and vice versa, and evaluate their relation."

It is important to adjust the usability framework to measure usability in various application contexts, such as mobile devices, home technology, or multiple-device systems. Traditional usability frameworks were created to measure a single product or software, making results very context specific and hard to generalize. Additional usability measures may be necessary when the context of use is changed for a specific framework. Monk (2002) and Soloway et al. (1994) studied the usability in the non-traditional context of use and require usability framework be appropriate adjusted to measure the system.

A multiple device system is common in our everyday life. Traditional usability models are not sufficient to evaluate the usability of this type of system. Denis and Karsenty (2003) put forward a conceptual framework of inter-usability. They defined the inter-usability as "the ease with which users transfer what they have learned from

previous uses of a service when they access the service on a new device" (Denis and Karsenty, 2003, p.381). They believed that knowledge continuity and task continuity were important and ergonomic design principles including consistency, transparency, and dialogue adaptability should be followed to ensure a good inter-usability across multiple user interfaces. Denis and Karsenty's study (2003) first created the notion of inter-usability to evaluate usability in multiple-device systems. However, their study was limited to the use of the same service system on different devices. In addition, a lack of theoretical support weakened the generalization of their approaches.

Table 3.1 Summary of the Difference & Overlap of the Existing Usability Models.

Author(s)	Efficiency	Learnability	Retention	Effectiveness	Satisfaction	General Usability	Aesthetics	Other
Bevan (1995)	Efficiency		Productive period	Effectiveness				
Constantine & Lockwood (1999)	Efficiency in use	Learnability	Rememberability	Reliability in use	User satisfaction			
De Angeli et al. (2006)						General Usability	Aesthetics	Interaction style
Denis & Karsenty (2003)								Inter-usability
ISO 9214-11 (1998)	Efficiency			Effectiveness	Satisfaction			
Kirakowski & Corbett (1993)	Efficiency	Learnability		Effectiveness	Satisfaction			Helpfulness, Control
Nielsen (1993)	Efficiency of use	Learnability	Memorability	Errors	Satisfaction			
Pohl et al. (2007)						General Usability		Transferability
Preece et al. (1994)	Throughput	Learnability		Throughput	Attitude			
Schneiderman (1992)	Speed of performance	Time to learn	Retention	Rate of errors	Satisfaction			
Seffah et al. (2006)	Efficiency	Learnability	Memory load	Effectiveness	Satisfaction		Attractiveness	127 usability metrics
Sutcliffe (2002)						General Usability	Attractiveness	
Thuring & Mahlke (2007)					Emotions		Aesthetics	

Studies of Single Usability Score

A lot of studies have tried to derive single-score usability metric. Shrestha et al. (2008) proposed a metric based on analytical hierarchical process for website usability evaluation. However, the results seem to vary according to the domain of website services and the application is restricted to website usability.

Babiker et al. (1991) proposed a single metric for usability in hypertext systems based on objective performance measures. They used three objective measures: user performance time, key stroke time and error rate and found correlation between their metric and subjective measures. However, this metric is still restricted to the hypertext system usability analysis.

McGee (2004) proposed a Master Usability Scaling (MUS) that utilize magnitude estimation for the analysis of usability. MUS was based on a subjective usability measurement Usability Magnitude Estimation (UME) (McGee, 2003) to standardize ratios of participants' subjective assessment ratings on tasks to derive a single score for task usability. The author derived this tool to be robust and universally applicable to a variety of tasks and products. However, the use of single source data that only represent subjective perception of the users may not be truly representative of the usability of an entire system.

Many usability questionnaires are also utilized to provide a single score for the analysis although they may not be designed for that purpose. SUS (Brook, 1996) was designed as a "quick and dirty" tool that assesses only one subset which is the usability. The score of 10 questions can be averaged to obtain an overall score of usability. CSUQ (Lewis, 1992a) and PSSUQ (Lewis, 1992b) were designed as a 19-item questionnaire to

evaluate computer system usability. Although study shows that CSUQ and PSSUQ measure three subsets of usability, System Usefulness, Information Quality, and Interface Quality (Lewis, 1995), they can usually be used to obtain a single score. Similar situation exists for QUIS (Chin et al., 1988) and SUMI (Kirakowski and Corbett, 1993). These tools for assessing usability with single score are beneficial in that they are quick, easy to administrate and easy to interpret. However, the reliance on only subjective data may lead to concerns about the reliability and validity of the construct of usability.

Usability Aspects

Efficiency is a widely accepted usability aspects (e.g. ISO 9214-11, 1998; Nielsen, 1993; Schneiderman,1992). Whether users could efficiently complete the task on a device directly indicate the usability of the device and affect users' experience with the device. Task completion time is often used to represent the efficiency dimension and proven to be a reliable objective measure.

Effectiveness is also widely used by many researchers. Most common ways to measure effectiveness is using errors or error opportunity. However, there are concerns whether errors should be included in a usability model (ANSI, 2001). And due to the variation in definitions of errors and highly subjectivity in detection of error, the results are often questionable. Other methods to represent effectiveness include task completion and usability difficulty. Task completion records whether or not participant complete a task. But as a binary variable, the information it provided is very limited. Usability difficulty can be extracted from user think aloud transcript. It provides information regarding how many usability difficulty users encounters when interacting with the interfaces.

Satisfaction represents users' perception toward the device. Many usability questionnaires were created to obtain this information (e.g. SUS, PSSUQ, CSUQ, SUMI). Many of these questionnaires were designed to measure more than one aspects of usability. Compared to thses usability questionnaires, SUS was designed to represent only one factor: system usability, which makes it easy and representative of users' subjective perception.

Transferability refers to the extent to which user can easily transfer between using multiple devices and adopt knowledge from previous device in using the new device. Transferability can be measured subjectively using STQ developed in Chapter II.

Other usability aspects could include learnability or memorability (Abran et al., 2003), and attractiveness and esthetics (Seffah et al., 2006; De Angeli et al., 2006; Sutcliffe, 2002). The usability aspects adopted should be dependent on the context of use and also the objective of the usability evaluation.

Standardized Usability Score

One of the biggest challenges of combining various measures into one single score is that these measures usually have different scale, thus these variables have different variance. Sauro and Kindlund (2005) proposed a method to standardize usability measures to a single score so that the scale issue is mitigated. Their approach was based on the usability aspects defined by ISO (ISO/IEC, 1998). They investigated task times, task completion, error counts and satisfaction scores to represent efficiency, effectiveness and satisfaction. Through principal component analysis they find similar loadings for all four measures. Therefore they use same weight for all standardized measures. This

approach employed various sources of data. They standardized the various variables using normal standardization as show in the equation below:

$$Z = \frac{X-\mu}{\sigma} \tag{3.1}$$

The z-equivalent was calculated for each variable. This standardization procedure ensures that the new variable has standard deviation of one. This approach of standardization is effective when comparing the usability between several choices. However, in order to calculate the specification $\bar{X}$, large amount of empirical data were needed. In a single experiment, the standardized variable would have a mean of zero, which provides no meaningful information regarding the usability of a device. In addition, for a multiple device system, it may not be sufficient to access the usability and transferability within the system and between the devices.

There are other ways to standardized variables such as scaling using the ideal value (SIV) and simple linearization (SL) (Yoon and Hwang, 1995). For SIV method, a maximum criterion value H_j is set as the ideal value for maximizing criterion (L_j was set as the ideal value for minimizing criterion). The scaling process is then represented by equation 2 below.

$$r_{ij} = \begin{cases} \frac{f_{ij}}{H_j} & formaximzing\ criterion \\ \frac{f_{ij}}{L_j} & forminimizing\ criterion \end{cases} \tag{3.2}$$

The SL method scale the variable into the range determined by the variable itself. Equation 3 shows the approach of SL method:

$$r_{ij} = \begin{cases} \frac{f_{ij} - L_j}{R_j} & formaximzing\ criterion \\ \frac{H_j - f_{ij}}{R_j} & forminimizing\ criterion \end{cases} \quad (3.3)$$

Where:
$R_j = H_j - L_j$
H_j = the highest value
L_j = the lowest value
The SL approach is advantageous because it can help to scale the variables into having the same standard deviation, which makes combining variables easier.

Study Objective

The existing literature presents two major gaps: the lack in theoretical approaches to combine subjective results and objective results, and the insufficient usability framework to characterize multiple-device system. With an established subjective transferability questionnaire to collect users' subjective perception regarding transferability and the newly proposed UPMDS usability framework, the overall research question of the chapter is: *Can we properly identify the weight and effect different measures have in explaining the overall system usability? How to consolidate all the measures into a single score?*

To address this research question, two objectives are established for this chapter. Since many subjective and objective usability measures (e.g. usability questionnaire, completion time, errors, transferability, usability difficulties, etc.) were identified to characterize different facets of overall usability construct, the first objective is to find out the role of these variables in explaining the overall usability.

A single score of system usability is helpful for reporting the usability, making decision regarding redesign, and make further statistical analysis (ANOVA, regression analysis, hypothesis testing) regarding the casual effect of design features on usability.

Therefore the second objective of this chapter is to adopt theoretical approaches to combine the single-device usability, users' performance data (completion time, error rates), and transferability to an overall system usability score.

This usability evaluation tool would be developed to not only provide an overall system usability score, but also be able to inform the usability practitioners which aspect of usability factors plays a more important role in overall system usability. Usability practitioners should be able to customize this tool based on the type of device system they are going to evaluate (single/multiple devices), the performance and perception measures they've recorded (completion time, errors, questionnaires, think aloud protocols, etc.) and obtain reasonable outcome regarding the overall system usability.

Methodology

In this section, the methodology of conducting the experiment and data collection as well as data analysis is presented. This study is based on the same experiment and the same participants with the study one. Therefore, most of the methodology of this study is the same with Chapter II (please refer to section 2.4 of Chapter II for details). However, this study utilizes slightly different data sets of variables and adopted different data analysis approach. These differences are described in this section.

Variable Definition

The variables that are used in the study to construct single usability score include objective measures and subjective measures. Objective measures include completion time (CTPS, calculated as the time between the start of each task to the end of each task, divided by the standard number of steps, recorded by analyzing video footage), error

steps (calculated as the number of extra error steps beyond the standard number of steps) and usability difficulties (calculated as the number of difficulties encountered when using the software, collected by analyzing verbal think aloud data). Subjective measures include participants' perceived transferability between devices that was collected using 7-point Likert scale System Transferability Questionnaire developed in this study (STQ, Appendix B). Participants' perceived usability regarding each device was collected using System Usability Scale (SUS, Appendix I). SUS was used instead of PSSUQ because it is design to be a one-dimensional questionnaire measuring usability instead of several factors measured by PSSUQ. Participants' perceived overall satisfaction was collected using a single item questionnaire (Appendix D).

Data Analysis

Principal component analysis (PCA, Jolliffe, 2002) was conducted for all variables. PCA is a mathematical procedure that uses orthogonal transformation to convert an original set of variables into a smaller set of uncorrelated variables that explain most of the variability in the original set of variables. PCA was developed to reduce the dimensionality of the original data set and is now widely used in exploratory data analysis. PCA is found to be effective in summarizing behavioral data in the social sciences (Dunteman, 1989; Jolliffe, 2002) and in usability studies (e.g. Calisir and Calisir, 2004; Sauro and Kindlund, 2005). As a comparison, the factor analysis used in Chapter II was aimed at investigating the underlying structure of the data with many variables. PCA, on the other hand, aimed at using a set of linearly uncorrelated principal components to simplify the huge data set, which will be helpful for explaining the data set or for further data analysis.

PCA was used in this study to uncover the contributions of different subjective and objective measures to system usability and remove variables that do not significantly explain the variability of the system usability. The weighting value obtain in PCA will be used to further construct a consolidated usability score.

The number of principals is decided using the Kaiser criterion (Kaiser, 1960, factors with eigenvalue greater than 1 are retained and factors with eigenvalue less than 1 are excluded) and scree plot rules (Cattell, 1966). The principal component loadings (eigenvectors) obtained for each variable would be used to decide the weight for these variables. Since different variables are obtained in different scale and approaches, standardization must be conducted before we can appropriately combine them. After standardization, these variables would have the same variance. The variables were standardized using simple linearization method. The overall system usability score can be obtained by a weighted average of the standardized scores of all the applicable variables. The weight of all factors would sum up to one. The simple linearization scale the value into [0,1]. In addition, the weighting from principal loadings ranges from 0 to 1. Therefore, theoretically, the final consolidated usability score will range from 0 to 1. Total usability score that is greater than 0.5 is considered decent and acceptable.

The variables used to obtain the overall usability score included the average subjective usability scores (from Systems Usability Survey, SUS), the average completion time per step (CTPS), the average usability difficulty (number of usability difficulties, calculated as the usability problems encountered in each software), and the system transferability questionnaire score (STQ).

Once the overall score of system usability was obtained, Pearson correlation was calculated between the system usability score and participants' one question survey to test whether this tool actually measures the overall system usability which it is what it is supposed to measure.

Results

Descriptive Statistics

Descriptive statistics are provided for all variables collected in the study (Table 3.2). These results are collected based on different scales. They have different variance. Therefore, they will need to be standardized before being combined.

Table 3.2 Descriptive Statistics for All Variables.

	Variables	Mean	SD	Max	Min
Objective Measures	Average Completion Time/Step (s)	9.63	2.65	15.92	4.36
	Average Errors/Task	13.44	6.63	37.08	2.75
	Average Usability Difficulties	0.21	0.21	0.92	0
	Task Completion (%)	91.1	9.42	100	66.67
Subjective Measures	STQ Scores (1-7)	4.16	1.11	6.31	1.94
	Single Item Score (1-7)	5.33	0.95	7	3
	Average SUS score (0-100)	64.56	11.91	97.5	42.5

Principal Component Analysis

To simplify factors and identify the weight for subjective and objective measures, principal component analysis is conducted using SAS 9.2. Five raw variables (three objective variables: average completion time, average errors, and average usability difficulties and two subjective variables: STQ scores and Average SUS score) were used.

The average completion time per step is used as a key measure for efficiency aspect of usability. The average errors per task is introduced as a measure for effectiveness. The average usability difficulty was developed as an objective measure that is measured subjectively. It provided information regarding effectiveness and user satisfaction. The STQ score is included as a subjective measure of the system transferability. At last, the average SUS score was included as a subjective measure of user satisfaction.

The scree plot (Figure 3.1) is created first. Based on the plot, two or three principal components would be appropriate. The eigenvalues of each principal component (Table 3.3) were further analyzed. The eigenvalue of principal component one and principal component two are greater than one, indicating retaining two principal components. The first two principal components help to explain a total of 66.52% of the total variance, which is marginally acceptable. Ideally, a cumulative variance of 70%-90% would be appropriate (Jolliffe, 2002). Based on the results, two principal components are retained.

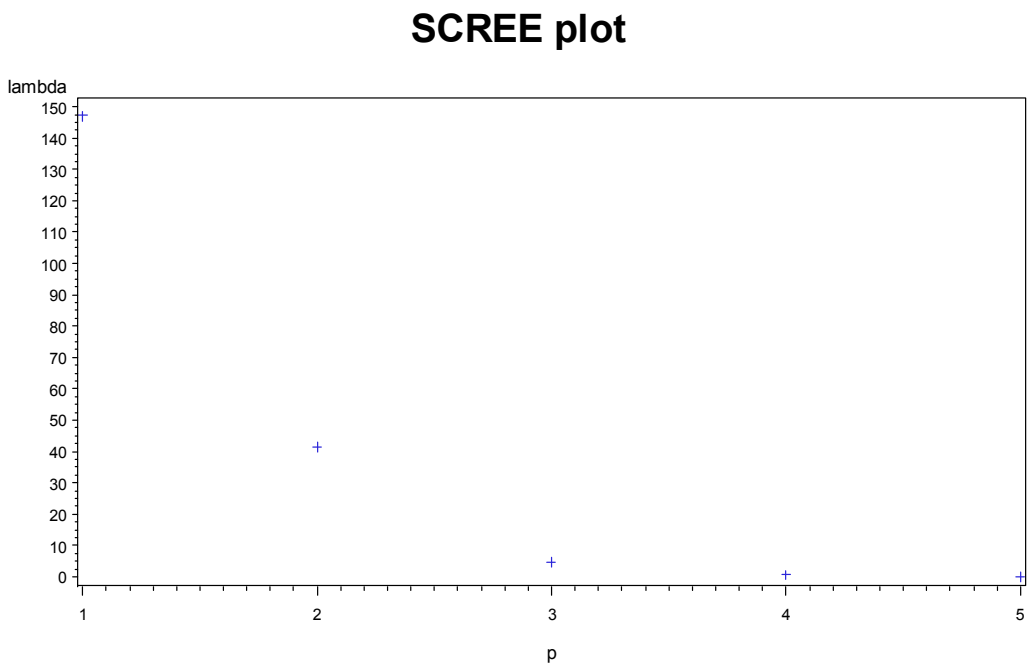


Figure 3.1 Scree Plot for the Principal Component Analysis.

Table 3.3 Eigenvalues of the Principal Components and the Variance Explained.

Principal Components	Eigenvalue	Proportion%	Cumulative%
1	2.01	40.14	40.14
2	1.32	26.38	66.52
3	0.9	18.06	84.59
4	0.46	9.18	93.77
5	0.31	6.23	100

Table 3.4 shows the eigenvectors (principal loadings) of the first two principal components. The first principal component is interpreted as the total usability because negative loadings appear before the SUS and the STQ scores and positive loadings appear for the performance measures. Principal component two is interpreted as the effort users have to exert to complete the tasks because it has higher loadings on CTPS and

STQ scores, which indicates the time spent on completing the task and the disturbance of transferring between devices. The Usability difficulty had very low loadings in both principal components (<0.3). This indicates that the variable of usability difficulty does not provide enough information to help explain the entire construct of usability. Therefore, the usability difficulty variable is removed.

Table 3.4 Eigenvectors (principal loadings) of the first two principal components.

Variables	**Principal Component 1**	**Principal Component 2**
SUS	-0.54	0.42
STQ	-0.48	0.53
CTPS	0.40	0.55
Errors	0.51	0.39
UX Difficulty	0.24	0.29

PCA is conducted again with variable usability difficulty removed. The scree plot (Figure 3.2) again shows that two principal components are appropriate. The analysis of eigenvalues (Table 3.5) shows that first two principal components had eigenvalues greater than one. In addition, the first two principal components help to explain a total of 80.70% of cumulative variance, which is acceptable (Jolliffe, 2002). Therefore, two principal components are retained. This four-factor construct showes an improvement from the five-factor construct. The cumulative variance explained improves for both the first and the second principal components.

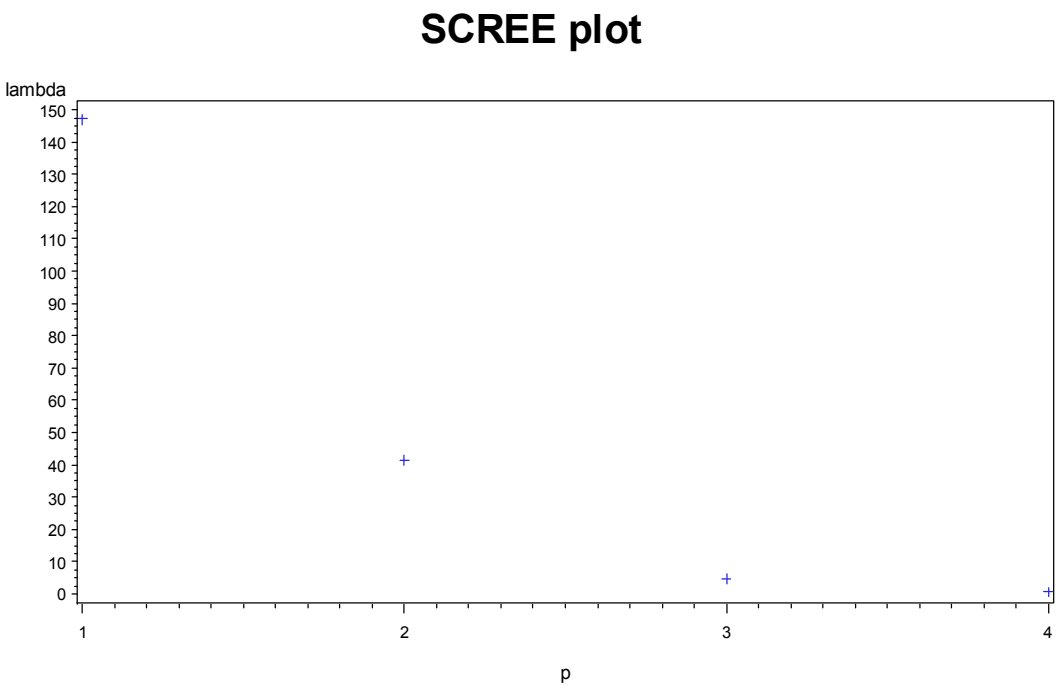


Figure 3.2 Scree Plot for the Principal Component Analysis (UX Difficulty removed)

Table 3.5 Eigenvalues of the Principal Components and the Variance

Principal Components	Eigenvalue	Proportion%	Cumulative%
1	1.95	48.73	48.73
2	1.28	31.96	80.7
3	0.46	11.51	92.21
4	0.31	7.79	100

Note: Factor "UX Difficulty" was removed

Table 3.6 shows the eigenvectors (principal loadings) of the first two principal components. The first principal component is interpreted as the overall usability. It has positive principal loadings for the SUS and the STQ because for these two questionnaires, the higher values indicate higher usability. The negative loadings for CTPS and Errors indicate that the usability was lower when users exhibited higher

completion time or higher errors. Principal component two helps to explain the effort in transfer process. However, it does not help to explain the usability of the system. Therefore, it is not used to determine the weighting of variables.

Table 3.6 Eigenvectors (principal loadings) of the first two principal components

	Principal Component 1	Principal Component 2
SUS	0.57	0.37
STQ	0.52	0.50
CTPS	-0.39	0.62
Errors	-0.50	0.46

Note: Factor "UX Difficulty" was removed

Variable Weightings

The weighting for each variable was decided based on the principal loadings of the PCA. The results showed that subjective and objective measures all have similar absolute principal loadings. The negative loadings indicate that CTPS and Errors are inversely correlated with SUS and STQ score, meaning that they follow a minimization criterion. When determining the factor weights, the absolute values of all loadings were used.

All principal loadings were first rounded to the closest decimals. The exact weighting is obtained as: SUS 0.6, STQ 0.5, CTPS 0.4, and Errors 0.5. Then, for standardization purpose, the weighting is adjusted so that the weights sum to 1. The final weighting is obtained as SUS 0.3, STQ 0.25, CTPS 0.2, and Errors 0.25 (Table 3.7).

Table 3.7 Procedure of Obtaining Standardized Weighting of the Variables.

	Principal Loadings	Principal Loadings (rounded)	Standardized Weightings
SUS	0.57	0.60	0.30
STQ	0.52	0.50	0.25
CTPS	-0.39	0.40	0.20
Errors	-0.50	0.50	0.25

Variable Standardization

All variables have to be standardized before they can be consolidated into an overall usability score. A simple linearization (SL) procedure was used to standardize all variables. STQ scores and SUS scores are standardized using equation 4 because they are a maximizing criterion.

$$r_{ij} = \frac{f_{ij} - L_j}{R_j} \tag{3.4}$$

Completion time per step and errors are standardized using equation 5 because they are a minimizing criterion.

$$r_{ij} = \frac{H_j - f_{ij}}{R_j} \tag{3.5}$$

Descriptive statistics of the variables after standardization are provided in Table 3.8. All variables range from zero to one. It also shows that all variables have similar standard deviation. The F- test for equality of variance (Table 3.9) shows the same results. Only STQ and average errors have marginally significant different variance. All the rest of variables show no significant difference regarding on standard deviation. Therefore, all variables are considered appropriate to combine.

Table 3.8 Descriptive statistics for variables after standardization

	Variables	Mean	SD	Max	Min
Objective Measures	Average CPTS	0.54	0.23	1	0
	Average Errors	0.69	0.19	1	0
Subjective Measures	STQ Scores	0.51	0.25	1	0
	SUS Scores	0.40	0.22	1	0

Table 3.9 p-values for the F- test for equality of variance

	SUS	STQ	CTPS	Error
SUS	1			
STQ	0.26	1		
CTPS	0.67	0.48	1	
Error	0.41	0.05	0.21	1

Total Usability Score

A total usability score (TUS) is calculated as the weighted average of the standardized value from the four variables. Equation 6 demonstrates the calculation:

$$TUS = 0.3 * SUS + 0.25 * STQ + 0.2 * CTPS + 0.25 * Err \quad (3.6)$$

The conversion of raw scores is provided in Appendix J. The total usability score (M=0.53, SD=0.16) ranges from 0.21 to 0.95. Theoretically, the total usability score would range from 0 (the worst usability) to 1 (the best usability). The results indicate that the multiple device system studied exhibits a medium overall usability.

To test if this construct really represent users' opinion about system usability, A pearson correlation is performed using the total usability score and one-item questionnaire score. These two variables are significantly correlated with a medium

correlation (r=0.49, p=0.0002). This shows that the usability construct developed in this study helps to explain users' opinion regarding on the system usability. Although the entire construct of overall system usability is still unknown. We can claim that the construct created in this study is capable of contributing to explain a large portion of the usability aspects.

Discussion

Variable Selection

Four variables (STQ scores, SUS scores, errors, and CTPS) are used to represent the four factors (transferability, satisfaction, effectiveness, and efficiency) respectively in UPMDS framework (Figure 3.3).

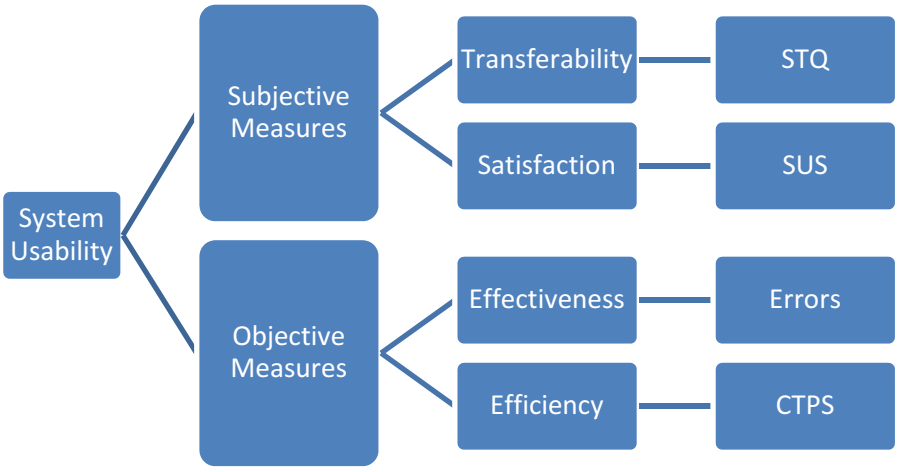


Figure 3.3 Usability Break Down and Corresponding Measures.

Task completion is used in many cases as a measurement for effectiveness. However, it is largely limited by the task completed. In this study, participants were able

to complete most of the tasks. Even if participants failed to complete the task in the correct way, they often managed to complete the task in a different or incorrect approach. This was encouraged because participants need time to explore the software, transfer their learning and evaluate the use of that software. Therefore, the task completion is not appropriate as a measurement factor in this study. Generally, task completion would be more suitable in situation of a more rigid usability testing scenario. In an open-ended testing scenario, errors would be more effective as a measurement for effectiveness.

The number of usability difficulties is explored in this study as a measurement for satisfaction and effectiveness. However, it is not included in the final model as it loaded weakly in the PCA analysis. The correlation analysis shows that it is weakly correlated with errors and CTPS. Although the number of usability difficulties was obtained subjectively, it will still be regarded as an objective measure. Its effect in predicting system usability has yet to be proven. But it can still serve as a valuable tool to elicit users concern regarding the effectiveness of the device.

Principal Components

Two principal components are obtained in the PCA analysis. The first principal component can be easily interpreted as the system usability as it exhibited positive loadings on SUS and STQ and negative loadings on CTPS and errors. The higher SUS scores and STQ scores are and the lower performance time per step and errors are, the higher system usability is achieved. In addition, the first principal component helps to explain around half of the total variance. Therefore, we can use it to decide the weighting of each variable. The second principal component indicated similar loadings on each of the variable. It is more difficult to interpret. Possible interpretation would be users'

familiarization with the device. However, with limited data, the result has not been supported.

Variables Weight

The PCA analysis obtains similar weighting for the four factors (SUS 0.3, STQ 0.25, CTPS -0.2, and Errors -0.25). This result corresponds well with the literature (Nunnally, 1978; Sauro & Kindlund, 2005). This shows that all four factors are important and indicative of the system usability. When evaluating the system usability, all four factors should be taken into consideration. This result also provides empirical support the UPMDS framework we developed in Chapter I, which introduces transferability as an equally important factor with the satisfaction, effectiveness, and efficiency.

Total Usability Score

A consolidated score is obtained using the approach introduced in this study. This consolidated score is aimed to represent the total system usability. This score will range from 0 to 1 with 0 representing the worst usability and 1 representing the best usability. This score could serve as a quick usability tool and give usability specialist a quick indication of the current usability of the system. If needed, the four usability sub-factors can be evaluated to identify potential usability problems. Priority of redesign should be focused on sub-factors that have the worst sub-scores. This usability evaluation tool has the following advantages.

First, this tool is based on the UPMDS framework. Compared with traditional usability evaluation tool, this tool captures a new construct of usability, which is the

transferability. This will be helpful in characterizing the usability issues user experience when transferring between using different devices.

Second, this tool provides a quantitative approach to evaluate total usability. Comparing with traditional usability evaluation approach that only collect performance measures or only use questionnaire, this tool involves both subjective and objective measures. The final score are indicative of the effectiveness, efficiency, user satisfaction, and the transferability, with user satisfaction a little higher weight the efficiency a little lower weight.

Third, this tool is quick, easy to administrate, widely applicable and adjustable. As long as the usability evaluator has data regarding different measures of the usability constructs, these data can be summated to a total usability score. This tool can be applied in not only multiple devices, but also other usability context. Usability evaluators just need to adjust the construct and measures and assign weightings (this was not done in this study, but future studies can examine the feasibility of using this approach).

Conclusion

This study utilizes an empirical software experiment to support the UPMDS framework. Four usability sub-factors are identified with similar weightings. A consolidated usability score iswas created for the software devices. The software system has usability slightly better than average, the biggest concern is in satisfaction and the best aspect is effectiveness.

To answer the research questions, we can successfully identify the weight and effect different measures have in explaining the overall system usability and we have created an approach to consolidate all the measures into one single score.

The study also has some limitation. First, larger data are needed to adopt better standardization procedures. With limited data, the standardized data may not truly represent the construct of user performance and perception. Second, more application context need to be tested using this approach. Slightly changes may be necessary if the measurement, usability context, or user group varies.

Future study should be able to generalize this approach to enable usability practitioner to customize the number of variables they want in this usability framework. They should be able to input the data of different variables and obtain the corresponding weight for each variable to create an overall usability score.

References

Abran, A., Khelifi, A., Suryn, W. & Seffah, A. (2003). Usability Meanings and Interpretations in ISO Standards. Software Quality Journal, 11(4), 325-338.

ANSI (2001). Common industry format for usability test reports (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.

Babiker, E.M., Fujihara, H., & Boyle, C.D.B. (1991). A metric for hypertext usability. ACM Systems Documentation 91, 95-104.

Bevan, N. (1995). Measuring usability as quality of use, Software Quality Journal 4,115–130.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M., & Mackenzie, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personnel Psychology, 48,* 587–605.

Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), Usability Evaluation in Industry. London: Taylor & Francis.

Calisir, F. and Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with Enterprise Resource Planning (ERP) systems, Computers in Human Behavior, 20(4), 505–515.

Cattell, R. B. (1966). The Scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In Proceeding of ACM CHI'88, Washington, DC, 213-218.

Constantine, L.L. & Lockwood, L.A.D. (1999). Software for Use: A Practical Guide to the Models and Methods of Usage-Centred Design, New York: Addison-Wesley.

De Angeli, A., Sutcliffe, A., & Hartmann, J. (2006). Interaction, usability and aesthetics: what influences users' preferences? 2006 ACM Press, 271-280.

Denis, C., & Karsenty, L. (2004). Inter-usability of multi-device systems – a conceptual framework, In: Seffah, A., Javahery, H. (Eds.), Multiple User Interfaces: Cross-Platform Applications and Context-Aware Interfaces (pp. 381–383). John Wiley & Sons, Ltd, West Sussex, England

Dunteman, George H. (1989) Principal Components Analysis. In Sage University Papers Series Quantitative Applications in the Social Sciences ; No. 07-069 Newbury Park Sage Publications, Inc.

Hornbeck, K (2006). Current practice in measuring usability: challenges to usability studies and research. *International Journal of Human-Computer Studies, 64,* 79-102.

John, B.E. & Kieras, D. E. (1996). Using GOMS for user interface design and evaluation: Which technique? ACM Transactions on Computer-Human Interaction 3: 287–319.

Jolliffe, Ian T.(2002). Principal Component Analysis. Secaucus, NJ, USA: Springer-Verlag.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational Psychology Measurement, 20, 141– 151.

Kirakowski, J. & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory, British Journal of Educational Technology 24: 210–212.

Lewis, J. R. (1992a). Psychometric evaluation of the computer system usability questionnaire: The CSUQ (Tech. Report 54.723), Boca Raton, FL: International Business Machines Corporation.

Lewis, J. R. (1992b). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In Proceedings of the Human Factors Society 36th Annual Meeting (pp. 1259-1263). Santa Monica, CA: Human Factors Society.

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.

McGee, M (2004). Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In Proceeding CHI 2004, 335 - 342.

Monk, A. F. (2002). Fun, communication and dependability: Extending the concept of usability. In Proceedings of HCI 2002, London; 3-14.

Nielsen, J. (1993). Usability engineering. Cambridge, MA: Academic Press.

Pohl, M., Rester, M., & Wiltner, S. (2007). Usability and transferability of a visualization methodology for medical data. in A. Holzinger (Ed.): USAB 2007, LNCS 4799, (pp. 171–184).

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). Human Computer Interaction, Wokingham, UK: Addison-Wesley.

Sauro, J. & Kindlund, E. A Method to Standardize Usability Metrics into a Single Score, Proc. CHI 2005, ACM Press (2005), 401-409.

Schneiderman, B. (1992). Designing the User Interface: Strategies for Effective Human-Computer Interaction (2nd ed.), Reading, MA: Addison-Wesley.

Seffah, A., Donyaee, M., Kline, R., & Padda, H. (2006). Usability Measurement and Metrics: A Consolidated Model, Software Quality Journal, 14, 159-178.

Shrestha, S., Helm, S.A., & Chaparro, B.S. (2008). Using the analytic hierarchical process to create a single usability score for website interfaces, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 52, 16, 1122-1126

Soloway, E., Guzdial, M., & Hay, K. E. (1994). Learner-centered design: The challenge for HCI in the 21st century. Interactions, 1(2), 36–47.

Sutcliffe, A. G. (2002). Assessing the reliability of heuristic evaluation for website attractiveness and usability. In Proceedings HICSS-35 Hawaii International Conference on System Sciences (Hawaii). IEEE Computer Society Press, Los Alamitos, CA, 1838-1847.

Thuring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction, International Journal of Psychology, 42, 253-264.

Yeh, Y, & Wickens, C. (1988). Dissociation of performance and subjective measures of workload. Human Factors, 30, 111-120.

Yoon, K.P; Hwang, C.-L. 1995: Multiple Attribute Decision Making, An Introduction. London: Sage Publications

CHAPTER IV

INVESTIGATING THE EFFECT OF TASK COMPLEXITY MACHINE ORDER AND USER EXPERIENCE ON SYSTEM USABILITY USING THE UPMDS FRAMEWORK

Introduction

Usability practitioners often adopt different approaches to evaluate the usability of various devices, trying to find out the factors that impact system transferability and how to modify those factors to improve system usability and user experiences. The factors identified typically fall into three categories: interface related factors (graphical user interface design, labeling), task related factors (task hierarchy, task complexity), and user characteristics (user experiences and training design). Interface factors have received a lot of focus and have become the key research area of usability studies. However, the latter two factors are overlooked in many usability studies. Task complexity was found to have moderating effects on user performance (Chae & Kim, 2004) and may lead to different user control and processing (Strawderman & Huang, 2012), causing additional usability problems. User experience may promote or limit user's interaction with the device, thus is one of the most important user characteristics to take into account when evaluating device usability. Therefore, it's critical to investigate the effect of task complexity and user experience on the usability outcome.

Previous chapters have laid down a theoretical and empirical basis for this study. The measurement of system usability will follow the UPMDS framework established in Chapter I. The calculation of usability score will follow the approach developed in Chapter III. To be a robust usability tool, this framework should be practically applicable to most usability scenarios and be able to contribute to both practical applications and research studies. A different application area, machine usability, will be used in this study, not only to serve for the main objective of this study, but also to test the validity and generalization of the UPMDS framework.

Literature Review

Task Analysis

Usability practitioners have been trying to design a user friendly interface by focusing on the interface design as well as taking into account the user characteristics. Central to achieving a friendly user interface, it is important to first understand what users want to achieve. What are the user's goals when they interact with the interface? What are their tasks? According to Hollnagel (2006), a task is defined as one or more functions or activities that must be carried out to achieve a specific goal. Task analysis methods came into place during the early 20th century to formally structure the physical tasks performed by the workers. Task analysis digs into the details of the task and tells us how things are being done or should be done. With the development of information systems and the increasingly dominant cognitive tasks, task analysis evolved into a method that aims at facilitating the design of complex human-computer interface.

Traditional task analysis started with sequential task analysis and was later dominated by Hierarchical Task Analysis (HTA). HTA was developed by Annett and

Duncan (1967) to evaluate the skills required in complex non-repetitive operator tasks. HTA breaks tasks into subtasks and operations or actions and represent task components in a hierarchical chart. HTA is aimed at analyzing and representing the behavioral aspects of complex tasks such as planning, diagnosis and decision making (Annett and Stanton, 2000). HTA is widely used by usability practitioners because it provides a model to evaluate the goals, tasks, subtasks, operations, and plans that are critical to users' activities. HTA is effective for decomposing complex tasks; however, the cognitive processes required of the user is not considered in the analysis. Although Annett and Stanton (2000) suggest that HTA can progress by embracing contextual analysis, there is still little research that can provide a systematic way for dealing with the social and physical context in which cognitive activities are prevalent.

The GOMS model (Goals, Operators, Methods, and Selection rules) is another established method for characterizing complex tasks (Card et al., 1983). GOMS models tasks in terms of a set of Goals (the objective that users intend to accomplish), a set of Operators (perceptual, motor or cognitive acts to achieve the goal), a set of Methods for achieving the goals (procedures that accomplished the goals), and a set of Selection rules (how users choose a certain method over the other competing methods) . The GOMS model provides a way to quantitatively predict user performance in an interactive system. It focuses on the keystroke level of a task which makes the results easily impacted by contextual factors. In addition, user factors such as errors, fatigue, learning effects, expertise were not fully account for in this model.

As noted by Barnard and May (2000), with the development of modern technology, "tasks have become more intricate, knowledge-intensive, and subject to

increasingly integrated forms of technological support, traditional forms of task decomposition appear to have an overly restricted scope" (Barnard and May, 2000:147). This necessitates the emergence of Cognitive Task Analysis (CTA) which focuses on more abstract, high-level cognitive functions. CTA is defined as the extension of traditional task analysis techniques to yield information about the knowledge, through processes, and goal structures that underlie observable task performance (Schraagen et al., 2000). Compared to HTA, CTA aimed at understanding modern task environment that require a lot of cognitive activity from the user, such as decision-making, problem-solving, memory, attention and judgment. However, cognitive task analysis does not always capture other non-cognitive attributes necessary for completing the tasks such as physical capabilities, access to resources, etc.

Tasks analysis incorporate models that focus on the microscopic parts of a task as well as models that focus on the high level of tasks like decision making and information need. It is not only effective for analyzing single task but also helpful in investigating tasks within a multiple device system. In usability research, task analyses is mainly used as a method to obtain information about the interface, capture user requirements, model and simulate user performance, and identify errors (Diaper and Stanton, 2008; Hackos and Redish, 1998). Few studies have utilized task analysis for usability evaluation of multiple device system. Task analysis method needs to be combined with other methods and techniques to effectively evaluate a multiple device system. Therefore, HTA will be used in this study to evaluate and understand the task structure.

Task Complexity

Wood (1986) defined task complexity into three types: component complexity, the number of different components associated with the task, coordinative complexity, the level of interaction between the components, and dynamic complexity, the degree to which the relationship between task related input cues and product changes over time. Total task complexity is further defined as a combination of the three objective complexity sources.

Prior research has found that high task complexity would increase the load on information processing, decision making and demand more cognitive resources from the users (Bystrom and Jarvelin, 1995;Klemz and Gruca 2003; Speier 2003). It is believed that complex tasks will lead to extensive use of cognitive resources which will cause people's attention to be diluted (Kahneman, 1973) or lead to a compromise of task performance for saving effort (Todd and Benbasat, 1999). Other research (Shiffrin and Schneider, 1977; Strawderman and Huang, 2012) found that simple tasks tend to lead to automated human cognitive processing, which require little or no cognitive effort from the user. This state of automated processing makes users slow in adapting to new tasks and vulnerable to task change and transfer effect. It is still unknown whether the resource depletion theory or the automaticity theory will dominate regarding the effect of task complexity in a multiple device system.

Wood (1986) defined three types of task complexity: component complexity, coordinative complexity and dynamic complexity. While the first two types of complexity can be measured and quantified, the dynamic complexity is high subjective and may vary according to different context. Campbell (1988) summarized the

characteristics of complex tasks as multiple paths, multiple end states, conflicting interdependence, and uncertainty or probabilistic linkages. This categorization is more intangible and difficult to quantify. Frese (1987) proposed that task complexity is determined by the number of decisions that have to be made and by the relations among these decisions. It is true that the number of decision points within a task structure represent the level of cognitive complexity of that task. In the tasks where physical steps also play an important role, the number of physical steps in a task structure should also be considered as one type of task complexity.

Therefore, using Frese (1987)'s definition, two task complexity will be adopted in this study. Cognitive task complexity is defined as the number of cognitive decision point in a task structure to help complete the task. Physical task complexity is the number of physical steps or processes needed to complete the task.

User Experience

Besides different levels of task complexity, user characteristics may have an important interaction effect on the usability of devices. Most existing literature of transfer of learning found a positive relationship between individual cognitive ability, motivation, self-efficacy and the transfer performances. However, limited number of studies examined the impact of previous experience on users' performance and perception towards the transferability of multiple device (e.g. Shanteau 1992; Ye and Salvendy 1994). Users' performance and perceptions can be affected by their mental models of the device, which is formed during their previous experience with the device. It is expected that less experienced users will be more sensitive to surface features, which refer to visual presentation and surface attributes. In contrast, experienced users tend to utilize

their knowledge of the underlying structure of the device and identify what they are able to do and how to proceed (Holyoak & Koh, 1987). With elaborated mental or cognitive model of the devices, experienced users may be more adapted to transfer between devices.

There are other opinions that high experience in a domain specific knowledge actually interferes with the transfer of learning to a novel situation. In a multiple-device system, this interference may cause poor transferability between devices and thus impact the whole system transferability.

Study Objective & Hypotheses

This chapter has two major objectives. The first objective is to test the reliability of the STQ. This questionnaire was applied to a different application setting and the study will test whether this tool could be generalized to other usability applications and successfully help in usability research endeavors. Therefore, the first research question is: *Will STQ be reliable when applied in a machine usability study?*

The second objective of this chapter is to investigate the impact of task complexity as well as user experience on the system usability. A laboratory study was designed to simulate different task complexity in a manufacturing environment. This chapter characterizes task complexity according to its physical complexity and cognitive complexity, by analyzing the task using hierarchical task analysis (HTA). Task complexity involves a larger spectrum of task characteristics including the physical and cognitive demand imposed on the end user. Higher cognitive task complexity would require more mental resources from the user. In a transfer situation, this would cause less disturbance to the user compared to lower cognitive complexity tasks in which situation

users are often automated performing the tasks. Higher physical complexity tasks would cause more disturbances due to its number of physical operations. Each task would be categorized in to one of the four levels: high cognitive high physical, high cognitive low physical, low cognitive high physical and low cognitive low physical.

The second research question is: *What are the effects of physical and cognitive complexity on the system usability? Do physical and cognitive complexity have interaction effect on the system usability?* To answer this research question, hypotheses one through three were created:

- *H1: Higher task complexity (cognitive complexity and physical complexity) would lead to lower overall usability of the system.*
- *H2: Lower cognitive task complexity would lead to lower transferability of the system, but higher satisfaction and better performance measures.*
- *H3: Lower physical task complexity would lead to higher satisfaction and better performance measures, but no change on transferability.*

This study is also interested in identifying the effect of individual differences on the evaluation of the system usability. Users' experience may modulate the performance and perception in a multiple-device situation.

The next research question is: *Would user experience affect system usability or users' perceptions towards the device?* Hypotheses four and five are based on the third research question:

- *H4: Experienced users would exhibit a higher overall system usability score.*

- *H5: Experienced users would exhibit better transferability, satisfaction and performance measures as compared to inexperienced users.*

It is of interest to investigate the interaction effect between task complexity and individual difference on the overall systems usability. The result of this question may be used to guide task design towards accommodation for different user experience groups. Thus, the research question is: *Is there interaction effect between user experience and task physical/cognitive complexity, or machine order on system usability?* Hypotheses six and seven are based on the fourth research question:

- *H6: Experienced users would exhibit higher overall usability score of the system when doing high complexity tasks. For low complexity tasks, experienced users would exhibit the same overall usability score as inexperienced users.*
- *H7: Inexperienced users will encounter greater impact by machine order. This effect would not impact experienced users.*

Methodology

Experimental Design

A between subjects design was used to test for the effect of cognitive task complexity (2 levels), physical task complexity (2 levels), user experience (3 levels), and machine order (2 levels) on the total system usability. All independent variables are between subjects variables. All two way interaction effects were examined. Factorial ANOVAs were conducted with total usability score as the dependent variable and cognitive complexity, physical complexity, and user experience as the independent variable. Repeated measures ANOVAs were conducted with completion time per step

and errors as the dependent variables and cognitive complexity, physical complexity, and user experience as the independent variables.

Variable Definition

Dependent Variables

Total system usability was used as the main dependent variable. It was calculated based on the UPMDS framework and scoring approach developed in Chapter III. Equation 1 demonstrates the calculation:

$$TUS = 0.3 * SUS + 0.25 * STQ + 0.2 * CTPS + 0.25 * Err \quad (4.1)$$

To better understand the effect of independent variables on the framework, the four sub-factors (completion time per step, errors, satisfaction, and transferability) within the UPMDS framework were also used as dependent variables. Completion time per step was calculated as the total completion time divided by the total number of physical steps necessary to complete the task. Errors per step was calculated as the total number of errors divided by the total number of physical steps necessary to complete the task. A user satisfaction score was obtained from SUS. A transferability score was obtained from the STQ questionnaire.

Participants' errors per step were further decomposed into different error types. Two classification schemes were adopted: Rasmussen's SRK model (Rasmussen, 1986) and a modified C/O/S/M model based on the model of Meister and Rabideau (1965). The SRK model categorizes errors into skill-based, rule-based, and knowledge based errors. The C/O/S/M model categorizes errors into commission errors, omission errors, sequence errors, and mistakes. These categories of errors were also used as dependent variables.

Independent Variables

A total of four independent variables were examined. Two types of task complexity, cognitive task complexity and physical task complexity were used as first two independent variables. Both type of task complexity involve two levels: high and low. Each task was designed and analyzed using hierarchical task analysis (HTA). The number of decision points (cognitive) and physical steps (physical) were recorded. At last, tasks were categorized in to one of the two levels (high and low) of cognitive complexity based on the number of cognitive decision points, and one of the two levels (high and low) of physical complexity based on the physical steps.

Participants' previous experience in using the experiment device was the third independent variable. Participants' experience was captured using an online demographic survey with a scoring system (Appendix K). Participants' experience score range from 0 to 10 (median=4) with a mean of 3.45 and a standard deviation of 3.07. Participants were divided in to three experience group based on the median and mean score of experience. Participants with experience score from 0-2 were categorized as inexperienced users, participants with experience score from 3-4 were categorized as medium experience users, and participants with experience score greater than 5 were categorized as experienced users.

The machine order was used as the last independent variable. There were two levels of order. The first level is using drill press first and the second level is using mini-lathe machine first.

Participants

Altogether forty-two participants were recruited from the university student population to participate in the experiment. One participant's data was incomplete due to a technical failure. Therefore, the participant was removed from the analysis, yielding a sample of 41(15 females and 26 males). Participants' age ranged from 18 to 53 years of age (M=23.88, SD=6.3). Participants were divided into three experience groups (16 in high experience group, 8 in medium experience group, and 17 in low experience group) according to their experience with the experiment machines. Participants were compensated with $10/hour for their participation, rounded to the nearest half hour.

Apparatus

Two machines: a mini-lathe and a drill press were selected as the experiment platforms for this study. They were selected for three primary reasons. First, both devices are commonly used in many manufacturing settings and tasks using these two machines are representative of typical manufacturing tasks, Second, the UPMDS framework and th STQ needed to be tested using a different platform to prove they are universally applicable. Machine platforms were selected because they are very different from software platforms. Third, both types of task complexity can be represented by operating tasks using these two machines.

The two machines used in this study are shown in Figure 4.1. A lathe machine (7" x 10" Precision Mini-lathe, produced by Central Machinery) and a drill press (3/8" drill press, produced by Shopmate) served as the study platform for this study. Two cameras were installed to capture user performance during the experiment from two angles (overhead and perpendicular) (Figure 4.2). These two cameras were synchronized and

controlled using EZWatch security camera system. Video data files were stored in a pass code enabled desktop computer.

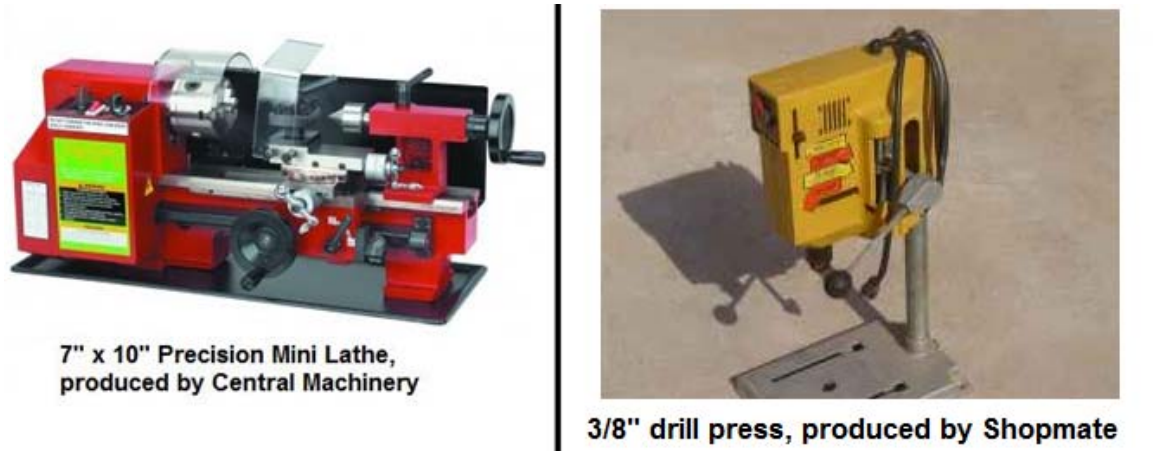


Figure 4.1 Two machines used in the study

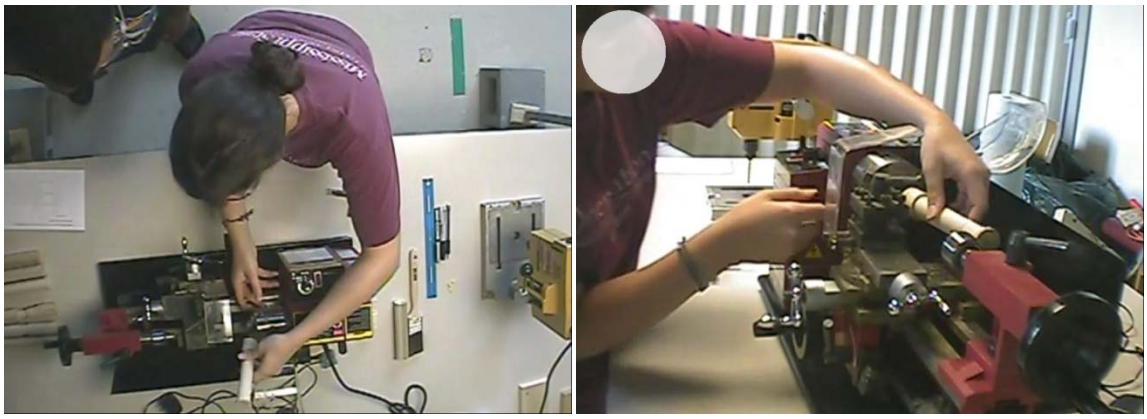


Figure 4.2 Two Camera Angles

Procedure

Participants were scheduled to come to the Human Systems Engineering Laboratory after the online screening and demographic survey. Before coming to the

laboratory, participants were informed to avoid wearing loose clothing, pull back long hair, and avoid wearing any watches or jewelry for the purpose of safety.

When participants came to the laboratory, they were first directed to the work table. An experimenter measured the elbow height of the participant and table height to ensure that the participants' work was within a comfortable range (elbow height within 2-10 inches above table height). Participants with a lower elbow height were compensated by standing on a large wood platform. For participants with higher elbow height, the table was raised.

A brief introduction was given to participants regarding the objective of the study, what they need to do in the study, potential fatigue or discomfort, safety precautions and compensation methods. Participants were informed that they can leave at any time without penalty if they feel uncomfortable. An informed consent was provided to each participant with all the above information included. The experimenter was available to answer any questions the participants had. Consented participants signed the informed consent before starting the experiment.

Each participant was randomly directed to start either from drill press or mini-lathe machine and exposure to machine was counterbalanced. One of the four task complexity combinations (high cognitive/high physical, high cognitive/low physical, low cognitive/high physical, and low cognitive/low physical) were selected prior to the participant's arrival. A training session was conducted for each participant and each machine. This was to help build base knowledge of the machine for each participant and as a safety precaution. The training involved a complete set of tasks covering the typical operations needed to complete the task. After the training, participants were given a test

run in which they were allowed to perform a trial task under the help of experimenter. Participants were allowed to ask any questions they had regarding the machines and tasks.

When no further questions were raised by participants, they were directed to prepare for starting the experiment. The experimenter started the camera capture software. Participants were informed that a think aloud protocol would be used which means they would be asked to state their thoughts of how to do the task, and any problems encountered while doing the experiment. Three tasks (Appendix L) were provided to participants in the form of card. Each card involved one task (task description on the top and sketch of finished products on the bottom). The order that tasks were presented was randomized for each participant. Due to the noise of machines, an experimenter took notes of the think aloud protocol from the participants instead of using audio recording.

Upon completion of the tasks using the first machine, participants took a five minute break. The experimenter administrated a paper-based SUS questionnaire for the participant to fill out. After the break, participants were directed to either the drill press or the mini-lathe machine, whichever was not used in previous tasks. Again, participants received a training session before staring the tasks. The experimenter answered any questions after the training session. Participants were reminded to use the think aloud protocol during the experiment. Participants then began the experiment with another card set of tasks. Upon completion of the tasks, participants completed the STQ, SUS, and the single item questionnaire. After completion of the questionnaires, participants were

compensated with $10/h based on their participation time and briefed about the experiment. An experiment protocol document is available in Appendix M.

Data analysis

All data analysis was conducted using SAS 9.2 statistical software. All results were considered significant at α=0.05 level. A factorial ANOVA was conducted with total usability score the dependent variables and levels of physical task complexity, levels of cognitive task complexity, machine order, and level of user experience as the independent variables. Potential interaction effects between user experience and task complexity were examined. Tukey's pair-wise comparisons were conducted to investigate the difference between different main levels and interaction levels of the independent variables. A repeated measure ANOVA was conducted with performance time per step, errors per step, and different breakdown of error types as the dependent variables and levels of physical task complexity, levels of cognitive task complexity, machine order and level of user experience as the independent variables.

Results

Descriptive Statistics

Descriptive statistics are provided for the dependent variables. Table 4.1 shows the raw statistics of the four factors in UPMDS. For calculation of the total usability score, the standardized scores are also calculated for the four variables. The descriptive statistics of standardized score and total usability score are provided in Table 4.2. Results show that the effectiveness factor has the highest score (M=0.80, SD=0.21) among the four factors. The total usability scored an average of 0.68 with a standard deviation of

0.13, which demonstrates an above average score. Figure 4.3 displays a histogram of the standardized total usability score.

Table 4.2 Descriptive Raw Statistics for the Factors of UPMDS.

Variables	Factor	Mean	SD	Max	Min
Avg. Completion Time/Step (s)	Efficiency	9.54	3.97	20.66	3.78
Avg. Errors/Step	Effectiveness	0.07	0.06	0.31	0.01
STQ Scores (1-7)	Transferability	4.67	1.15	6.40	1.47
SUS Scores (0-100)	Satisfaction	69.93	14.59	96.25	30.00

Table 4.3 Descriptive Standardized Statistics for the Factors of UPMDS and Total Usability Score.

Variables	Factor	Mean	SD	Max	Min
Avg. Completion Time/Step (s)	Efficiency	0.66	0.24	1	0
Avg. Errors/Step	Effectiveness	0.8	0.21	1	0
STQ Scores (1-7)	Transferability	0.65	0.23	1	0
SUS Scores (0-100)	Satisfaction	0.6	0.22	1	0
Total Usability Score		0.68	0.13	0.96	0.31

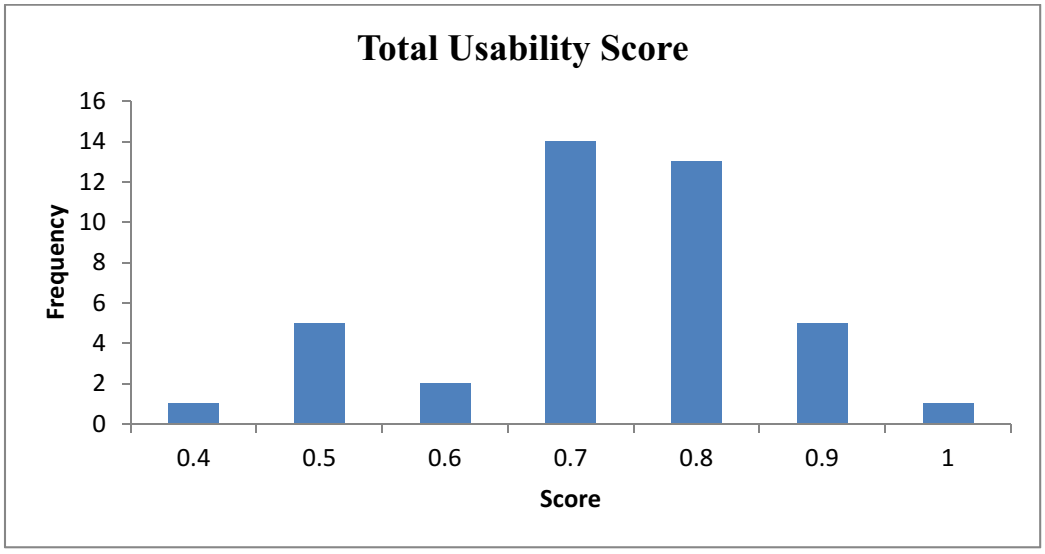


Figure 4.3 Histogram of the Standardized Total Usability Score

Participants' errors in operating the tasks are categorized using two types of error classification scheme: Rasmussen's SRK model (Rasmussen, 1986) and a modified C/O/S/M model based on the model of Meister and Rabideau (1965). The descriptive statistics of the participants' error types together with the percentage of recognized error and recovered errors are provided in Table 4.3.

Table 4.4 Descriptive Statistics of Error per Step and Percentage of Recognized and Recovered Errors.

		Mean	SD
C/O/S/M Model	Commission Error	0.0154	0.0682
	Omission Error	0.0244	0.0379
	Sequence Error	0.0148	0.0309
	Mistake	0.0159	0.0345
	Total Errors	0.0705	0.0967
SRK Model	Skill Based Error	0.0405	0.0592
	Rule Based Error	0.0239	0.0568
	Knowledge Based Error	0.0061	0.0127
	Total Errors	0.0705	0.0967
Other	% Error Recognized	8.92	21.12
	% Error Recovered	6.45	18.87

Reliability of STQ

To test the reliability of STQ in the machine usability evaluation, confirmative factor analysis is conducted for the question items in STQ. A varimax-rotated factor pattern of the factor analysis is presented in Table 4.4. All factor patterns are consistent with the findings in Chapter II except for Q12: "The second machine presents information that is consistent to the first machine". This question item was in the

consistency perception (CP) group in the software usability study in Chapter 2. However, in this study, it is categorized into Functionality group.

Cronbach's Alpha for the STQ is 0.91, which is the same as the results in Chapter II. Cronbach's Alpha for transfer experience (TE) sub-factor is 0.94 while Cronbach's Alpha for overall experience (OE) is 0.83.

Table 4.5 Varimax-Rotated Factor Pattern for the Factor Analysis of Machine Transferability Using Four Factors.

Item	Factor 1	Factor 2	Factor 3	Factor 4
Q2	**0.94**	0.13	0.05	-0.06
Q1	**0.9**	0.14	0.11	-0.13
Q3	**0.9**	0.1	0.01	-0.1
Q7	**0.8**	0.24	-0.16	-0.19
Q4	**0.77**	0.22	0.01	0.21
Q10	**0.75**	0.03	0.44	0.11
Q6	**0.7**	0.51	0.05	0.17
Q11	**0.63**	0.61	-0.1	0.22
Q5	0.04	**0.85**	-0.2	0.22
Q14	0.23	**0.76**	0.37	-0.21
Q16	0.12	**0.76**	0.17	0.09
Q15	0.35	**0.71**	0.43	-0.04
Q9	-0.03	0.19	**0.93**	-0.13
Q13	0.25	0.14	-0.11	**0.82**
Q12	0.37	-0.03	0.04	**-0.65**

Factorial ANOVA

A factorial ANOVA is conducted to examine the effect of task complexity, user experience, and machine order on the total usability score, satisfaction (SUS scores), and

transferability (STQ scores). The ANOVA results of total system usability are presented in Table 4.5. No significant results are found at α=0.05 level.

Table 4.6 AVOVA results for the total system usability score

Source	DF	F-value	P-value
Cog_Complexity	1	2.05	0.1637
Phy_Complexity	1	2.14	0.1550
Experience	2	0.80	0.4609
Machine Order	1	1.52	0.2283
Cog_Complexity*Experience	2	0.48	0.6241
Cog_Complexity*Machine Order	1	0.85	0.3654
Phy_Complexity*Cog_Complexity	1	1.19	0.2845
Machine Order*Experience	2	0.23	0.7952
Phy_Complexity*Experience	2	0.03	0.9739
Phy_Complexity*Machine Order	1	2.54	0.1229
Error	26		
Total	40		

ANOVA results for the system transferability, as measured by STQ, are provided in Table 4.6. Results show that the main effect of machine order has significant impact on the transferability of the system ($F(1, 26) = 42.94$, $p < .0001$). In addition, the main effect of cognitive complexity has marginally significant impact on the transferability of the system ($F(1, 26) = 3.97$, $p = .0570$). No significant interaction effects are identified. Tukey's post-hoc comparison shows that the order of mini-lathe first drill press second exhibited significant higher transferability than the order of drill press first and mini-lathe second. In addition, low cognitive complexity tasks shows significant higher system transferability than high cognitive complexity tasks.

Table 4.7 AVOVA results for the system transferability.

Source	DF	F-value	P-value
Cog_Complexity	**1**	**3.97**	**0.057**
Phy_Complexity	1	1.02	0.3227
Experience	2	0.29	0.7511
Machine Order	**1**	**42.94**	**<0.0001**
Cog_Complexity*Experience	2	0.91	0.4150
Cog_Complexity*Machine Order	1	1.27	0.2695
Phy_Complexity*Cog_Complexity	1	0.12	0.7291
Machine Order*Experience	2	1.69	0.2048
Phy_Complexity*Experience	2	1.37	0.2723
Phy_Complexity*Machine Order	1	0.82	0.3746
Error	26		
Total	40		

ANOVA results for user satisfaction (Table 4.7), as measured by SUS, are similar to the transferability results (Table 4.6). Results show that the main effect of machine order has significant impact on the user satisfaction ($F(1, 26) = 4.82$, $p = .0373$). In addition, the main effect of cognitive complexity has marginally significant impact on satisfaction ($F(1, 26) = 3.90$, $p = .0590$). No significant interaction effects were identified. Tukey's post-hoc comparison shows that the order of mini-lathe first drill press second exhibits significant higher satisfaction than the order of drill press first and mini-lathe second. In addition, low cognitive complexity tasks shows significant higher satisfaction than high cognitive complexity tasks.

Table 4.8 AVOVA Results for the Satisfaction.

Source	DF	F-value	P-value
Cog_Complexity	**1**	**3.9**	**0.059**
Phy_Complexity	1	0.27	0.6078
Experience	2	0.07	0.9371
Machine Order	**1**	**4.82**	**0.0373**
Cog_Complexity*Experience	2	0.76	0.4785
Cog_Complexity*Machine Order	1	0.19	0.6692
Phy_Complexity*Cog_Complexity	1	0.03	0.868
Machine Order*Experience	2	1.81	0.1831
Phy_Complexity*Experience	2	0.2	0.8185
Phy_Complexity*Machine Order	1	0.56	0.4597
Error	26		
Total	40		

Repeated Measures ANOVA

Repeated measures ANOVA is first conducted with completion time per step as the dependent variable and cognitive complexity, physical complexity, user experience, and machine order as the independent variables, with repeated measures on task order (task order refers to the order each task was presented to the participants, different from the machine order which is the order of the machine that participants used). Results are shown in Table 4.8. Results show that the interaction effect of machine order and physical complexity has significant impact on the users' completion time per step ($F(1, 201) = 5.82$, $p = .0168$). The interaction effect of machine order and cognitive complexity also has significant impact on the users' completion time per step ($F(1, 201) = 6.55$, $p = .0112$). The interaction of physical complexity and cognitive complexity also has a significant effect on the users' completion time per step ($F(1, 201) = 4.21$, $p = .0414$). In

addition, machine order and user experience has an interaction effect on the users' completion time per step ($F(2, 201) = 4.21$, $p = .0162$).

Post-hoc analysis with Tukey's adjustment indicates that when the mini-lathe was presented first and the physical complexity was low, user's completion time per step is significantly higher than the rest of combination groups (Figure 4.4). When the drill press was presented first and cognitive task complexity was low, users exhibit significantly lower completion time per step than the other combination groups (Figure 4.5). In addition, when the task had low cognitive complexity and high physical complexity, users exhibit a significantly lower completion time per step as compared to the rest of combination groups (Figure 4.6). Finally, participants with medium or low experience levels exhibit significantly higher completion time per step when the mini-lathe was presented first as compared to when the drill press was presented first. This effect is not significant for high experience participants (Figure 4.7).

Table 4.9 Repeated measures AVOVA results for the completion time per step.

Source	DF	F-value	P-value
Cog_Complexity	**1**	**17.94**	**<0.0001**
Phy_Complexity	**1**	**6.88**	**0.0094**
Experience	2	0.71	0.4949
Machine Order	**1**	**18.75**	**<0.0001**
Task Order	5	1.61	0.1587
Machine Order*Phy_Complexity	**1**	**5.82**	**0.0168**
Machine Order*Cog_Complexity	**1**	**6.55**	**0.0112**
Phy_Complexity*Cog_Complexity	**1**	**4.21**	**0.0414**
Task Order*Machine Order	5	0.93	0.3363
Task Order*Phy_Complexity	5	1.91	0.0948
Task Order*Cog_Complexity	5	0.67	0.6448
Machine Order*Experience	**2**	**4.21**	**0.0162**
Phy_Complexity*Experience	2	0.24	0.7851
Cog_Complexity*experience	2	1.44	0.2403
Task Order*experience	10	1.03	0.4234
Error	201		
Total	245		

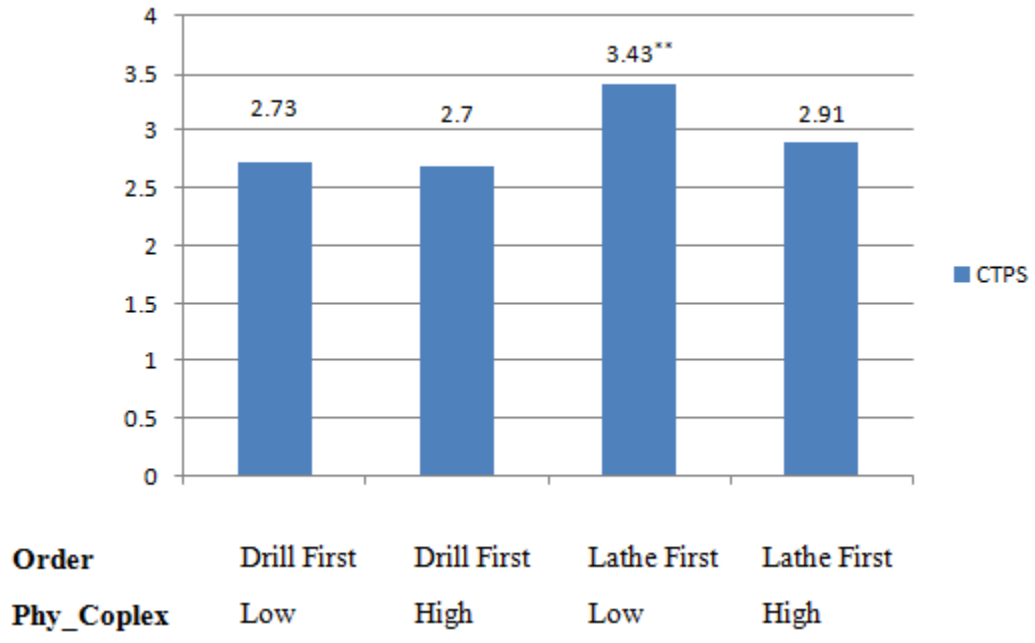


Figure 4.4 Post hoc comparison of the machine order*physical complexity effect

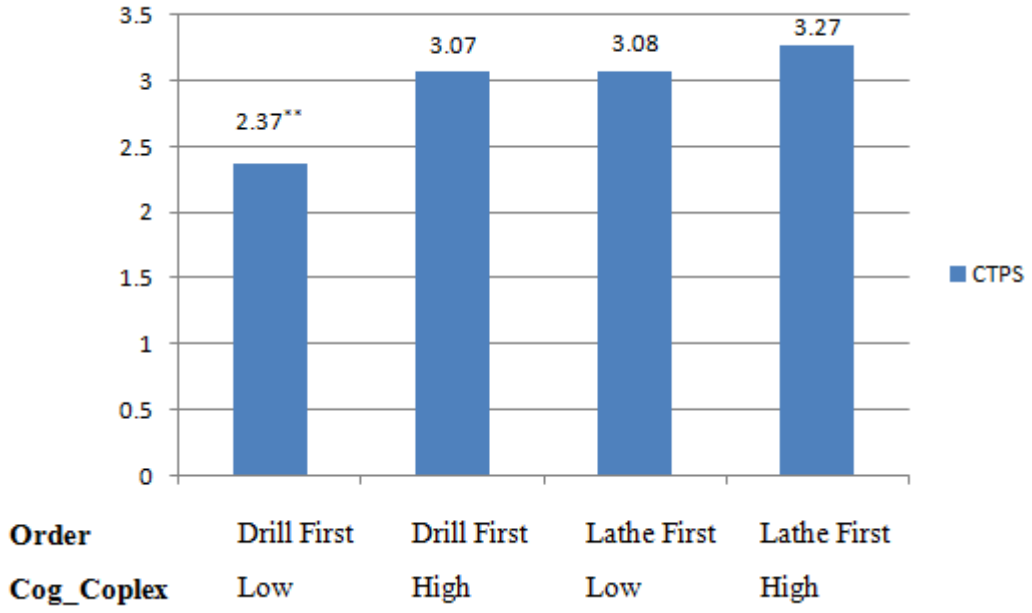


Figure 4.5 Post hoc comparison of the machine order*cognitive complexity effect

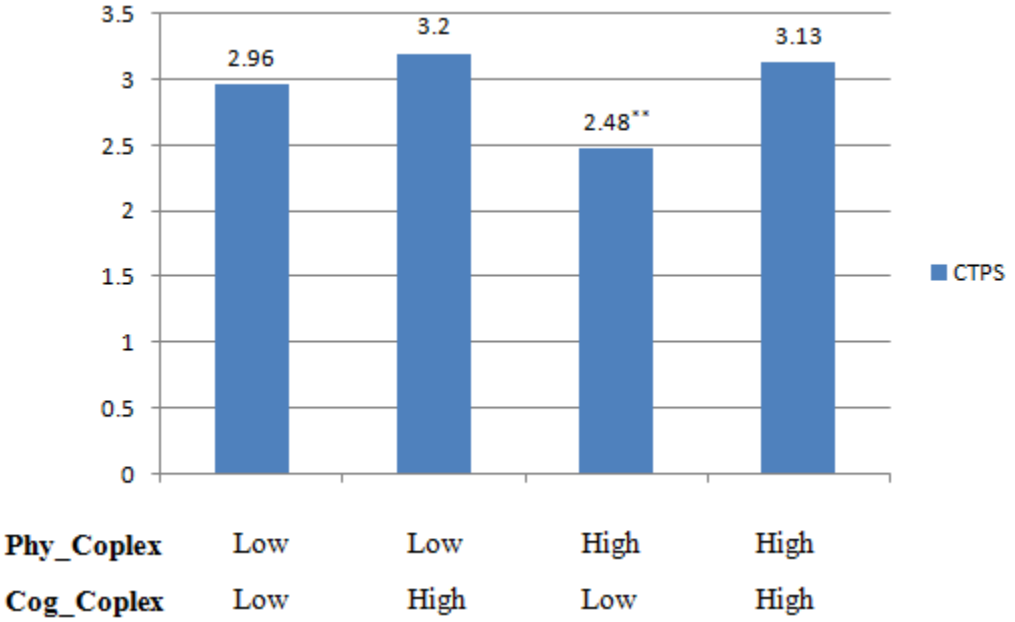


Figure 4.6 Post hoc comparison of physical complexity*cognitive complexity effect

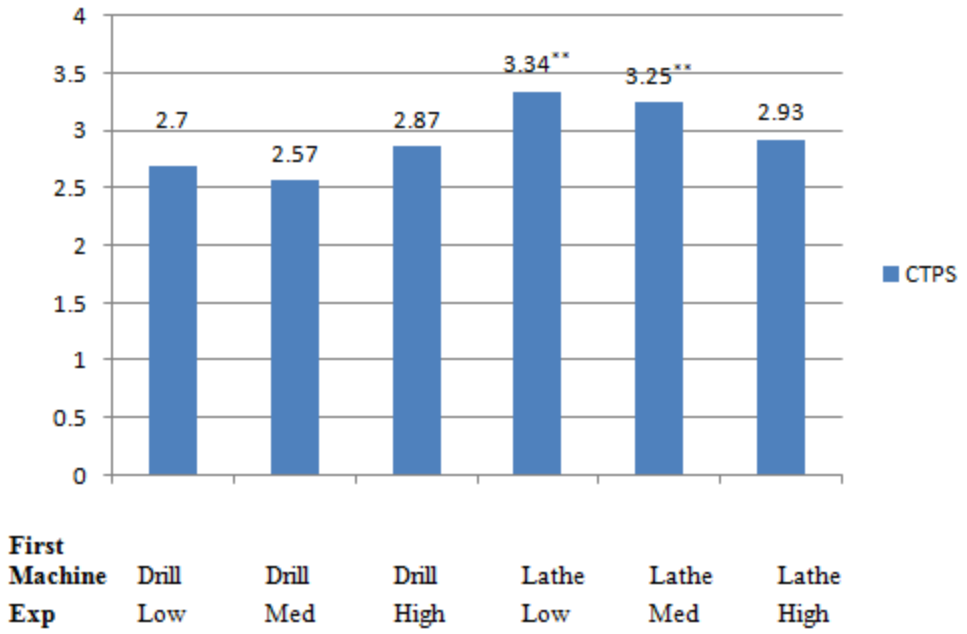


Figure 4.7 Post hoc comparison of the machine order*experience effect

Repeated measures ANOVA is also conducted with errors per step as the dependent variable and cognitive complexity, physical complexity, user experience, and machine order as the independent variable, with repeated measures on task order. Results are showed in Table 4.9. Results show that the main effect of cognitive task complexity ($F(1, 201) = 4.76$, $p = .0304$) , physical task complexity ($F(1, 201) = 8.12$, $p = .0048$), user experience ($F(2, 201) = 8.80$, $p = .0002$), and machine order ($F(1, 201) = 12.32$, $p = .0006$) has a significant impact on users' errors per step. In addition, the interaction effect of task order and machine order has significant impact on the users' errors per step ($F(5, 201) = 5.31$, $p = .0001$).

Post-hoc analysis with Tukey's adjustment indicates that low physical complexity tasks exhibited significantly lower number of errors per step compared to high physical complexity. Low cognitive complexity tasks exhibit significantly higher number of errors

per step as compared to high cognitive complexity tasks. High experience participants exhibit significantly lower errors per step as compared to medium and low experience participants. In addition, when the mini-lathe was first used, the highest number of errors per step was found. Figure 4.8 shows the LS means of the combinations. Left side is the drill press first and right side is the mini-lathe first. Task 1-6 represent task orders of drill press first scenario. Task 7-12 represent the task 1-6 in mini-lathe first scenario.

Table 4.10 Repeated measures AVOVA results for the errors per step.

Source	DF	F-value	P-value
Cog_Complexity	**1**	**4.76**	**0.0304**
Phy_Complexity	**1**	**8.12**	**0.0048**
Experience	**2**	**8.80**	**0.0002**
Machine Order	**1**	**12.32**	**0.0006**
Task Order	5	0.49	0.7829
Machine Order*Phy_Complexity	1	2.22	0.1376
Machine Order*Cog_Complexity	1	0.92	0.3384
Phy_Complexity*Cog_Complexity	1	1.13	0.2899
Task Order*Machine Order	**5**	**5.31**	**0.0001**
Task Order*Phy_Complexity	5	0.30	0.9117
Task Order*Cog_Complexity	5	0.61	0.6903
Machine Order*Experience	2	0.85	0.4286
Phy_Complexity*Experience	2	0.24	0.7863
Cog_Complexity*Experience	2	0.32	0.7248
Experience*Task Order	10	0.84	0.5923
Error	201		
Total	245		

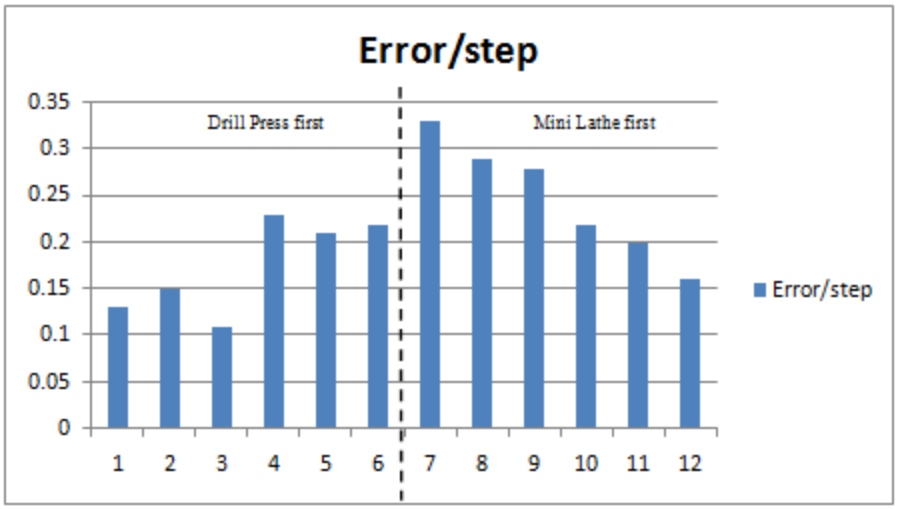


Figure 4.8 LS means for the interaction effect of task order and machine order

Note: * 1-6 represent task order 1-6 when drill press was used first. 7-12 represent the task order 1-6 when mini-lathe was used first.

Different error classifications were also examined. A modified C/O/S/M model based on the model of Meister and Rabideau (1965) is used to classify the errors. Commission errors refer to the errors that extra steps were taken by participants that were unnecessary. Omission errors refer to the errors that participants missed steps of tasks that is supposed to be completed. Sequence errors refer to errors in which the order of the steps was wrong. Mistakes refer to the errors that do not fall into any of the above categories. Repeated measures ANOVA is conducted with commission errors per step, omission errors per step, sequence errors per step and mistakes per step as the dependent variable and cognitive complexity, physical complexity, user experience, and machine order as the independent variable, with repeated measures on task order. ANOVA results are presented in Table 4.10. Results show that for the commission types of error, the interaction effect of machine order and cognitive complexity has a significant impact on

the number of commission errors participants made per step ($F(1, 201) = 10.92$, $p = .0011$). Post-hot comparison shows that when the mini-lathe machine was used first, low cognitive complexity tasks exhibited significantly higher commission errors than the rest of combinations. Physical task complexity and cognitive task complexity also has a significant interactive effect on the number of commission errors per step ($F(1, 201) = 4.51$, $p = .0349$). Post-hoc analysis shows that low cognitive and low physical complexity tasks exhibit the highest commission errors.

For omission types of error, task order and machine order have an interaction effect on the number of omission errors made per step ($F(5, 201) = 10.27$, $p < .0001$) and number of sequence errors made per step ($F(5, 201) = 23.03$, $p < .0001$). Task order and cognitive complexity also have an interaction effect on the number of omission errors made per step ($F(5, 201) = 2.58$, $p < .0277$). At last, cognitive complexity has a main effect on the number of mistake made per step in tasks ($F(1, 201) = 5.01$, $p = .0263$). Post-hoc comparison shows that low cognitive complexity tasks exhibited significantly higher mistakes during tasks.

Errors per step is also examined in terms of SRK model. Repeated measures ANOVA is conducted with skill, rule, and knowledge based errors per step as the dependent variables and cognitive complexity, physical complexity, user experience, and machine order as the independent variable, with repeated measures on task order. Results are showed in Table 4.11. Results indicate that the main effect of cognitive task complexity ($F(1, 201) = 8.05$, $p = .0050$) , physical task complexity ($F(1, 201) = 4.23$, $p = .0410$), and user experience ($F(2, 201) = 4.88$, $p = .0085$), has a significant impact on users' skill based errors per step. In addition, the interaction effect of task order and

machine order has significant impact on the users' skill based errors per step ($F(5, 201) = 9.65$, $p < .0001$).

Table 4.11 Repeated measures ANOVA results for the four types of error C/O/S/M.

		Commission		Omission		Sequence		Mistake	
Source	**DF**	**F value**	**p value**	**F value**	**p value**	**F value**	**p value**	**F value**	**p value**
Cog_Complexity	1	10.62	**0.0013**	0.01	0.9203	0.85	0.3568	5.01	**0.0263**
Phy_Complexity	1	5.73	**0.0176**	7.71	**0.006**	0.04	0.8378	0.09	0.7697
Experience	2	2.39	0.0942	6.05	**0.0028**	1.86	0.1582	2.60	0.0768
Machine Order	1	23.41	**<0.0001**	6.44	**0.0119**	13.79	**0.0003**	0.51	0.4772
Task Order	5	1.75	0.1249	0.73	0.6002	1.33	0.2528	1.19	0.3147
Machine Order*Phy_Comp	1	1.25	0.2644	1.58	0.2109	2.86	0.0922	0.33	0.5651
Machine Order*Cog_Comp	1	10.92	**0.0011**	0.16	0.6886	0.23	0.6311	0.01	0.9869
Phy_Comp*Cog_Comp	1	4.51	**0.0349**	0.66	0.4180	2.25	0.1349	1.94	0.1651
Task Order*Machine Order	5	0.74	0.5965	10.27	**<0.0001**	23.03	**<0.0001**	0.40	0.8478
Task Order*Phy_Comp	5	1.56	0.1742	0.35	0.8835	0.97	0.4383	1.14	0.3430
Task Order*Cog_Comp	5	1.27	0.2771	2.58	**0.0277**	0.46	0.8058	0.61	0.6901
Machine Order*Experience	2	2.82	0.0619	0.14	0.8686	0.14	0.8693	0.29	0.7493
Phy_Comp*Experience	2	2.61	0.0763	0.18	0.8315	0.75	0.4733	2.49	0.0852
Cog_Comp*Experience	2	1.53	0.2192	1.74	0.1787	1.85	0.1597	0.47	0.6272
Experience*Task Order	10	0.48	0.903	0.75	0.6777	0.61	0.8048	0.60	0.8125
Error	201								
Total	245								

Post-hoc analysis with Tukey's adjustment indicates that low physical complexity tasks exhibited a significantly higher number of skill based errors per step compared to high physical complexity. Low cognitive complexity tasks also exhibits significantly higher skill based errors per step as compared to high cognitive complexity tasks. High experience participants exhibit significantly lower skill based errors per step as compared to medium and low experience participants.

Results also show that the main effects of machine order ($F(1, 201) = 12.66$, $p = .0005$) and user experience ($F(2, 201) = 4.24$, $p = .0158$) have a significant impact on users' rule based errors per step. No significant interaction effect is identified for users' rule based errors per step.

Post-hoc analysis with Tukey's adjustment indicates that when drill press is used first, the participants exhibit significantly lower rule based error per step compared to when mini-lathe machine is used first. High experience participants exhibit significantly lower rule based errors per step as compared to medium and low experience participants.

For the knowledge based errors per step, results also show that there are main effect of user experience ($F(2, 201) = 4.57$, $p = .0114$), and interaction effect of task order and machine order ($F(5, 201) = 8.07$, $p < .0001$) on knowledge based errors per step. Post-hoc analysis with Tukey's adjustment shows that high experience participants exhibit significantly lower knowledge based errors per step as compared to low experience participants.

Table 4.12 Repeated measures AVOVA results for the S/R/K types of errors

		Skill Based		Rule Based		Knowledge Based	
Source	**DF**	**F-value**	**P-value**	**F-value**	**P-value**	**F-value**	**P-value**
Cog_Complexity	1	8.05	**0.005**	0.18	0.6694	0.01	0.9475
Phy_Complexity	1	4.23	**0.041**	3.50	0.0626	0.01	0.9446
Experience	2	4.88	**0.0085**	4.24	**0.0158**	4.57	**0.0114**
Machine Order	1	7.08	**0.0084**	12.66	**0.0005**	0.87	0.3513
Task Order	5	1.73	0.1285	0.63	0.6780	0.70	0.6237
Machine Order*Phy_Complexity	1	0.96	0.3282	2.09	0.1502	0.05	0.8327
Machine Order*Cog_Complexity	1	0.23	0.6318	0.69	0.4070	2.01	0.1575
Phy_Complexity*Cog_Complexity	1	2.45	0.1192	0.02	0.8841	0.44	0.5062
Task Order*Machine Order	5	9.65	**<0.0001**	0.23	0.9508	8.07	**<0.0001**
Task Order*Phy_Complexity	5	0.35	0.8807	0.84	0.5222	1.67	0.1433
Task Order*Cog_Complexity	5	1.12	0.353	0.77	0.5738	0.35	0.8823
Machine Order*Experience	2	2.75	0.0665	1.54	0.2160	2.82	0.0619
Phy_Complexity*Experience	2	0.24	0.7897	0.74	0.4785	1.14	0.3204
Cog_Complexity*Experience	2	2.07	0.1295	0.06	0.9460	0.24	0.7849
Experience*Task Order	10	0.97	0.4675	1.50	0.1410	0.33	0.9709
Error	201						
Total	245						

Discussion

Reliability of STQ

The confirmative factor analysis shows a relative consistent factor patterns and loadings of STQ question items as compared to the results in Chapter II. This shows that STQ is robust and can measure a consistent construct of transferability when applied in machine devices. The only question that fell under a different factor pattern is Q12: "The second machine presents information that is consistent to the first machine". This question item was designed for use in information technology devices to elicit users' perception on information consistency. When used with a machine device, this question becomes ambiguous as "machine" itself presents very limited "information". A possible revise of the question would be "The second machine is operated in a way that is consistent with the first machine." In this way, the question still asks about the users' consistency perception, but eliminates the ambiguity regarding the platform.

The STQ also has a high overall Cronbach's alpha value, indicating that the internal reliability is evidenced. In addition, both of the sub-factor TE and OE shows a high internal reliability. Therefore, to answer the first research question: *With minor modification, the STQ is considered robust and reliable to be used in a machine device usability evaluation.*

Effect of task complexity

Physical task complexity and cognitive task complexity was found to have no impact on the total system usability score. This result is not expected because many studies found that task complexity has modulating effect on people's perception and performance (Bystrom and Jarvelin, 1995; Todd and Benbasat, 1999). There are two

possible reasons for this result. First possible reason is all of the usability constructs are not significantly impacted by task complexity. This indicated that total usability score is a device level characteristic and does not change based on the level of task complexity. A second possible reason is that the four usability constructs was under different impact of the task complexity. When these four components were linearly combined in to a single usability score, the significant effects of independent variables were counteracted. However, in either reason, hypothesis one was not supported: *Higher task complexity (cognitive complexity and physical complexity) has no significant effect on the overall usability of the system.*

To explore the modulating effect of task complexity, the four sub-factors of usability are further examined. Results show that low cognitive complexity tasks lead to higher transferability. This is also not expected as low cognitive task complexity will easily make users automated in performing the tasks. This will make users vulnerable to any transfer impact (Strawderman and Huang, 2012). The reason for this result is that compared to computer based tasks, machine tasks involve both physical and cognitive tasks which makes the task complex enough to prevent users from entering automated processing. This result does support the cognitive resource theory. During transfer of learning, users need to adopt cognitive resources to observe, comprehend, and react to the changes. When cognitive task complexity is high enough to occupy most of the cognitive resources, the transfer process will have to be sacrificed.

Low cognitive complexity tasks are also found to lead to higher satisfaction. This result is expected. Physical task complexity is not found to have a significant effect on transferability or user satisfaction.

For performance measures, conflicting results are identified. Low cognitive complexity tasks are found to lead to lower performance time per step but higher errors per step. Low physical complexity tasks are found to lead to higher performance time per step but lower errors per step. One possible reason for this is that low cognitive complexity tasks make users do the task without thinking, but leading to a lot of errors. High physical complexity tasks make users do repeated steps which speed up the task and reduce the errors. Therefore, hypothesis two and three are not fully supported: *Lower cognitive task complexity leads to higher transferability of the system, higher satisfaction, faster performance time and higher errors. Physical task complexity has no effect on transferability or satisfaction. Lower physical task complexity leads to longer performance time and lower errors.*

Effect of User Experience

User experience levels have no effect on the overall usability score. This result is not expected. It is expected that higher experienced user group would have a better mental model of the machine device, which will lead to easier use and transfer between the devices. As the total usability construct is composed of transferability, satisfaction, performance time per step and errors per step, it is critical to also investigate the effect of user experience on these usability constructs.

Regarding the subjective component of individual constructs, user experience is found to have no significant effect on satisfaction and transferability. This may happen because both satisfaction and transferability are measured subjectively using questionnaires. High experienced users may have a more complete mental model of the machines, thus be able to identify more usability or transferability issues that

inexperienced users are not aware of. This may have offset the effect of better performance of the high experience users. In addition, different experienced user has different objective and standard when using the machines. Therefore, the subjective ratings are affected by users' experience. This result shows that both high and low experienced uses should be used in a usability test because both user groups are able to identify different usability issues and the results would not bias the evaluation outcome.

Experience levels do not have significant main effect on the performance time per step. However, high experience participants exhibit significantly lower errors per step as compared to medium and low experience participants. This effect holds true for skill, rule and knowledge based errors. This effect is also significant for omission errors. This shows that user experience is effective in reducing users' errors when operating the machines. All of the skill, rule, and knowledge based errors are reduced for the experienced user group. Omission errors are reduced for experienced user group while commission error, sequence errors and mistakes are not affected by experience level. In addition, medium experienced users show no difference with inexperienced users indicating a trend to divide the user group to either experienced users (experience scores higher than 5) and inexperienced users (experience scores below 5).

The hypothesis four is not supported by the results. Hypothesis five is partially supported by the results: *Experienced users and inexperienced users exhibit same overall usability score of the system. Experienced users would exhibit lower errors per step, but the same transferability, satisfaction and performance time per step as compared to inexperienced users.*

No interaction effect is found between user experience and task complexity. However, user experience interacts with machine order to impact user's performance time per step. Inexperienced users exhibit significantly higher completion time per step when machine order is mini-lathe first as compared to when order is drill press first. This effect is not significant for high experience participants. This result supports our hypothesis. Inexperienced users are prone to the machine order effect. When the machine order causes a disturbance, users will likely take more time completing the task. Therefore, for hypothesis six and seven: *There is no interaction effect between user experience and task complexity. Inexperienced users encounter greater impact in terms of performance time per step by machine order. This effect does not exist for experienced users.*

The results of user experience also show that subjective and objective measures may capture different construct of usability. Objective measures capture aspects of usability that can be explained by the effect of user experience. However, subjective measures capture aspects of usability that based on perception, knowledge, preferences, experiences, etc. There is still no conclusion whether subjective or objective measures provide a better representation of true usability construct. We believe that both measures are key to the UPMDS framework to better present the usability construct.

Effect of Machine Order

The effect of machine order received divergent results from subjective measures and objective measures. The transferability and satisfaction results show that when mini-lathe was used first, the transferability from mini-lathe to drill press is higher and the satisfaction score was higher. However, the objective measures show that when mini-

lathe was used first, users' completion time per step and errors per step was significantly higher than drill press was used first. The possible reason for this discrepancy is that the operation of drill press is closer to user's mental model and easier to be accepted. The operation of mini-lathe is relatively more different from users' mentor model. When drill press is used first, the greatest disturbance occurred during the transfer, leading to a subjective dissatisfaction towards the system. When mini-lathe is used first, the greatest disturbance was imposed at the start. This impact inflicted on task one will lead to the compromise of the performance of users throughout the process. However, the transfer to drill press later would cause fewer disturbances and seems easier, which is the reason of higher transferability score of this machine order.

This result again proves the claim that both subjective and objective measures capture different construct of the usability. This is valuable information for the usability designers. For the functionality and user performance design, usability specialist can collected objective data to inform on design. For user satisfaction and perception, usability specialists can collect subjective data to inform the design process. Sometimes a trade-off between adopting subjective or objective measures will have to be decided on which to promote and which to sacrifice.

Conclusion

In this chapter, a machine usability study is designed to examine the effect of task complexity, user experience and machine order on the total usability and its sub-factors. Results indicate that cognitive task complexity lead to divergent effects on usability constructs. Physical task complexity has no effect on subjective usability measures, but lower physical task complexity leads to longer performance time and lower errors from

the users. User experience level has no effect on subjective measures. High experienced users have significantly lower errors made in tasks. The machine order also has divergent results. When the mini-lathe is used first, users have better subjective transferability results but poor performance outcomes as compared to when drill press is used first.

This study also has several limitations. First, users did not achieve automated processing with the study tasks. We found that high task complexity is limiting the usability and transferability. Future studies need to examine the effect of the automated processing state caused by low task complexity and identify the lower limit of task complexity to best promote the use of multiple device system. Second, the result of this study should be able to be generalized to a broader spectrum of usability studies and user groups. Future studies should be applied on a wider range of devices and tested on a more diverse user population.

Reference

Annett, J., & Duncan, K.D. (1967). Task analysis and training design. Occupational Psychology, 41, 211-221

Annett, J., and Stanton, N., eds. (2000). Task analysis. London: Taylor & Francis.

Barnard, P. and May, J. (2000). Towards a theory-based form of cognitive task analysis of broad scope and applicability. In: Schraagen, et al., 147-163.

Bystrom, K. & Jarvelin, K. (1995). Task complexity affects information seeking and use. Information Processing and Management, 31(2), 191—214.

Card, S., Moran, T. and Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum.

Campbell, D. J., 1988, Task complexity: A Review and Analysis, *Academy of Management Review, 13,* 40-52.

Chae, M., & Kim, J. (2004). Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. Behaviour & Information Technology, 23(3), 165-181.

Diaper, D. & Stanton, N. A. (2004). The Handbook of Task Analysis for Human-Computer Interaction. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Frese, M., 1987, A theory of control and complexity: implications for software design and integration of computer systems into the work place, in: M. Frese, E. Ulich, & W. Dzida, (Ed.), Psychological issues of human-computer interaction in the workplace, vol. Elsevier Science, Amsterdam, pp.313-337.

Hackos, J. & Redish, J. (1998). User and task Analysis for Interface Design. Chichester: Wiley.

Hollnagel, E. (2006) Task Analysis: Why, What, and How, in Handbook of Human Factors and Ergonomics, Third Edition (ed G. Salvendy), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0470048204.ch14

Holyoak, K.J., Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory & Cognition. 15(4), 332–340

Kahneman, D. Attention and Effort, Prentice-Hall, Englewood Cliffs, NJ, 1973.

Klemz, B.R. and T.S. Gruca, "Dueling or The Battle Royale? The Impact of Task Complexity on the Evaluation of Entry Threat," Psychology & Marketing, Vol. 20 No. 11:999-1016. 2003.

Meister, D., & Rabideau, G. F., (1965). Human Factors Evaluation in System Development. Wiley, New York.

Rasmussen, J. (1986). Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering. New York: North-Hollandh/Elsevier.

Schraagen, J., Chipman, S., and Shalin, V. (2000). Cognitive task analysis. Mahwah, NJ: Lawrence Erlbaum.

Shanteau, J. (1992). Competence in Experts: The Role of Task Characteristics. Organizational Behavior and Human Decision Processes, 53, 252-266.

Speier, C.(2003). The Influence of Query Interface Design on Decision- Making Performance. MIS Quarterly (27), 3, 397-423.

Shiffrin RM, & Schneider W. (1977). Controlled and automatic information processing: II. Perceptual learning, automatic attending, and a general theory. Psychol. Rev. 84, 127-190.

Strawderman, L. & Huang, Y. (2012). Designing Product Feature Upgrades; the Role of User Processing and Task Change. International Journal of Industrial Ergonomics, 42, 435-442.

Todd, P., and Benbasat, I. (1999). Evaluation the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection. Information Systems Research (10), 4, 356-375.

Wood, R. E. (1986). Task complexity: Definition of the construct. Organizational Behavior and Human Decision Processes, 37, 60–82.

Ye, N., & Salvendy, G. (1994). Quantitative and qualitative differences between experts and novices in chunking computer software knowledge. International Journal of Human-Computer Interaction, 6(1), 105-118.

CHAPTER V

CONCLUSION

Summary of Research

Technological advances create new challenges for the interactions between individuals and devices. The context of use is becoming more multi-media. Users' interaction with technology is no longer limited to a single device. Thus, it is important to ensure an easy and usable interface for the users. This study introduces a usability framework (UPMDS) to characterize the usability in multiple device system. This study constructed a system transferability questionnaire (STQ) to supplement the framework. A systematic scoring approach is also introduced to complete the framework. This framework is applied in a machine usability scenario to answer specific research questions.

System Transferability Questionnaire

Questionnaires have been identified to be effective in capturing users' subjective perceptions regarding the device they use. Thus, this study follows a systematic approach to develop the STQ tailored specifically to measure the transferability between multiple devices. When applied to a software usability study, this questionnaire demonstrates high reliability and validity. The STQ is effective in measuring four sub-factor groups of the

transferability: transfer experience (TE), overall experience (OE), consistency perception (CP), and functionality perception (FP).

The STQ fills the usability research gap that no effective method is available to measure transferability. The STQ also fills the gap of UPMDS framework by providing a valid and reliable measure for the subjective usability component: the transferability. STQ is designed to be applicable in any usability scenario that involves multiple devices. The questionnaire items can be modified to suit different devices and context of use. If usability specialists want to understand the details of transferability sub-factors, the score of four sub-factors will provide valuable information regarding transfer experience, overall experience, consistency, and functionality.

Scoring System for UPMDS

One of the objectives of UPMDS framework is to conceptualize the usability constructs in the context of multiple device usage. But more importantly, this framework should be able to guide the theoretical approach to derive a quantitative tool to measure total usability under the framework. This study provides a quantitative measurement tool that fulfills the objective of the UPMDS framework. With this measurement tool, the UPMDS framework is also complete.

For future application, this scoring tool will be a quick and easy measurement for identifying total usability score. Usability specialist will be able to know the usability status of a system by adopting this scoring tool and inputting various subjective and objective measures. Usability specialist could also adjust the weight and variables according to the specific usability context of interest.

Effects of Task Complexity, User Experience, and Machine Order

The UPMDS framework and scoring system is applied in a machine usability study. It is found that cognitive task complexity has no effect on total usability score but lead to divergent effects on usability constructs. Physical task complexity had no effect on subjective usability measures, but leads to longer performance time and lower errors from the users. User experience level has no effect on subjective measures. High experienced users have significantly lower errors made in tasks. The machine order also has divergent results. When the mini-lathe is used first, users have better subjective transferability results but poor performance outcomes as compared to when drill press is used first.

Future Work

First, more studies are necessary to test the robustness of the STQ when applied in other usability scenarios. Although the questionnaire has been tested in software usability and machine usability, and results seem to be consistent and replicable, the reliability of STQ in other scenarios is still unknown.

Second, the UPMDS framework does not incorporate usability aspects such as aesthetics, affection, or task completion. This was because these aspects have limitation in application. For example, aesthetics are helpful in explaining usability for consumer products. But when evaluating usability of machines or medical devices, aesthetics may not be indicative of true usability. Task completion time is less informative when tasks are too complex or too simple. Future study could examine these usability constructs in specific using context and scenarios.

Third, subjective and objective measures were found to be measuring different usability construct. When they were lineally combined, the effects of many variables were masked. Future study should focus on the sub-factor scores when using the UPMDS to answer usability research questions.

APPENDIX A

ONLINE DEMOGRAPHIC SURVEY FOR STUDY I AND II

Software Usability Analysis Exit this survey

1. Software Usability Study Recruitment Survey

Thank you for taking the time to participate in this survey. The purpose of this survey is to select research participants. Research participants will perform software tasks on a desktop computer. Results from the project will be used to improve the design of software usability.

If you agree to participate in this study, we are asking that you take about 3 minutes and complete a brief online screening survey.

Your participation is completely voluntary and you can end the survey at any time by closing the browser. Your responses will remain confidential. All identifying information will be removed from your responses when your survey is submitted.

If you have any questions about this survey or the research study, please contact the researcher listed below. For additional information regarding your rights as a research subject, please feel free to contact the MSU Regulatory Compliance Office at (662)325-3294.

Yunchen Huang
Ph.D. Candidate
Industrial & Systems Engineering
Mississippi State University
yh95@msstate.edu

By entering the survey area, you indicate that you are at least eighteen years old and are giving your informed consent to be a participant in this study. If you would like a print copy of this document, please use the "print" function on your Internet browser.

Next

Software Usability Analysis Exit this survey

2. Basic Information

1. In which year were you born?

Year

2. What's your gender?

- Male
- Female

3. What's your ethnicity?

- Hispanic or Latino
- American Indian or Alaska Native
- Asian
- Black or African American
- White
- Native Hawaiian or Other Pacific Islander

4. What's your educational level?

- 8th grade or less
- Some high school
- High school grad or GED
- Some college or 2-year degree
- 4-year college degree
- More than 4-year degree

Prev Next

Software Usability Analysis Exit this survey

3. Experiences

1. Have you ever worked in the following industry?

- [ ] Construction
- [ ] Manufacturing
- [ ] Auto repair
- [ ] Office/Secretary
- [ ] Art Design/Image processing
- [ ] Other
- [ ] None of these

2. If yes to the above question, for how long?

Construction	
Manufacturing	
Auto repair	
Office/Secretary	
Art design/Image processing	
Other	

Prev Next

1. Do you have the following hobbies?

- [ ] Woodwork
- [ ] Auto repair
- [ ] Image processing
- [ ] Web design
- [ ] None of these

Prev Next

1. Have you used Adobe Acrobat before?

- ◯ Yes
- ◯ No

Prev Next

1. How often do you use Adobe Acrobat? (Openning pdf files using Adobe Reader does not count)

- ◯ Frequently (almost every day)
- ◯ Moderately (around once per week)
- ◯ Periodically (around every other month)
- ◯ Infrequently (less than once per year)

Prev Next

1. Have you used Adobe Photoshop before?

- ◯ Yes
- ◯ No

Prev Next

1. How often do you use Adobe Photoshop?

- ◯ Frequently (almost every day)
- ◯ Moderately (around once per week)
- ◯ Periodically (around every other month)
- ◯ Infrequently (less than once per year)

Prev Next

1. Are you requesting extra credit? (For students from IE 4613 only)

- ○ Yes.
- ○ No.

2. Please indicate your typical available time during a week.

	Day	Time
First choice		
Second choice		
Third choice		

3. Please enter your name, email address, and telephone number so that we can contact your regarding the experiment schedule.

Prev Next

APPENDIX B

ORIGINAL SYSTEM TRANSFERABILITY QUESTIONNAIRE (STQ)

Participant ID #_____

The System Transferability Questionnaire (STQ)

Instructions:

This questionnaire, which starts on the following page, gives you an opportunity to tell us your reactions to the software packages you used. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with **both software** while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, please write "N/A" in comments.

Please provide additional comments to elaborate on your answers.

After you have completed this questionnaire, I'll go over your answers with you to make sure I understand all of your responses.

Thank you!

1. Overall, I am satisfied with how easy it is to use the second software package after using the first software package.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

2. It is simple to use the second software package after using the first software package.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

3. I can quickly complete the task when using the second software package after using the first software package.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

4. I can correctly complete all tasks when using the second software package after using the first software package.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

5. I felt comfortable using both software packages and transferring between them.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

6. I felt frustrated using the second software package after using the first software package.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

7. I can quickly learn how to use the second software package after I changed from using the first software package to the second software package.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

8. Using the first software package helped me learn to use the second software package faster.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

9. The visual display and layout are generally consistent between the two software.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

10. I felt more efficient using second software package after using the first software package.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

11. The process of transferring to use the second software package after using the first software package is frustrating and makes me lost.

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

12. The second software package presents information that is consistent to the first software package

STRONGLY DISAGREE **1** **2** **3** **4** **5** **6** **7** **STRONGLY AGREE**

COMMENTS:

13. I will easily confuse some functions in the second software package with the functions in the first software package.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

14. Overall, I enjoy the experience of using both software packages

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

15. Overall, I am satisfied with using both software packages

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

16. Overall, I'm frustrated and confused with using both software packages.

STRONGLY DISAGREE 1 2 3 4 5 6 7 STRONGLY AGREE

COMMENTS:

APPENDIX C

POST-STUDY SYSTEM USABILITY QUESTIONNAIRE (PSSUQ)

1. Overall, I am satisfied with how easy it is to use this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

2. It was simple to use this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

3. I could effectively complete the tasks and scenarios using this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

4. I was able to complete the tasks and scenarios quickly using this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

5. I was able to efficiently complete the tasks and scenarios using this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

6. I felt comfortable using this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

7. It was easy to learn to use this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

8. I believe I could become productive quickly using this system.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

9. The system gave error messages that clearly told me how to fix problems.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

10. Whenever I made a mistake using the system, I could recover easily and quickly.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

I

11. The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

12. It was easy to find the information I needed.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

13. The information provided for the system was easy to understand.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

14. The information was effective in helping me complete the tasks and scenarios.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

15. The organization of information on the system screens was clear.

STRONGLY DISAGREE	1	2	3	4	5	6	7	STRONGLY AGREE

COMMENTS:

Note: The interface includes those items that you use to interact with the system. For example, some components of the interface are the button, the lever, the outlook (including their use of graphics, signs and language).

16. The interface of this system was pleasant.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

17. I liked using the interface of this system.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

18. This system has all the functions and capabilities I expect it to have.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

19. Overall, I am satisfied with this system.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

APPENDIX D

SINGLE ITEM QUESTIONNAIRE

Participant ID #_____

One-Question Survey

To as great a degree as possible, think about the **two software system you just used**. How satisfied are you with the system?

1. Please rate your overall satisfaction level with the two software system:

NOT AT ALL SATISFIED	**1**	**2**	**3**	**4**	**5**	**6**	**7** **VERY SATISFIED**

COMMENTS:

APPENDIX E

TRAINING TASKS FOR STUDY I AND II

Training Tasks For Adobe Acrobat

1. Open a PDF file named “file1.pdf” from the desktop.
2. Using the “Tool” tab on the upper right corner, rotate the current document 180 degrees.
3. Using the “Tool” tab on the upper right corner, insert all pages from “file2.pdf” to the current document, placing the inserted page at the start of the document.
4. Using the “Edit” menu, find the word “Rationale” in the current document.
5. Using the “Comment” tab on the upper right corner, highlight the word “Rationale”.
6. Using the “Edit” menu, take a snap shot of the Figure 1 and paste it into a Word document.
7. Using the “Comment” tab on the upper right corner, draw a rectangle on the bottom of the current document.
8. Using the “View” menu, turn on the grid option.
9. Using the “Comment” tab, cross out the text in the introduction section.
10. Turn off the grid and save the document as “training task.pdf” on desktop.

Training Tasks For Adobe Photoshop

1. Open image file “Training Image.psd”using Adobe Photoshop.
2. Using the “image” menu, change the image width to 1400 pixels width. The rest of the parameters will be automatically adjusted.
3. Using the “Layer” menu, add one extra layer named “layer1” to the image, keeping the rest of the parameters as default. (from lower right corner you can see the layer added)
4. Using the “Image” menu, rotate the image 180 degrees.
5. Using the left column tool bar, select a rectangle area of the image.
6. Choose a color you like using the upper right color selection area.
7. Select the brush tool from the left column tool bar, and change the brush size to 40 pixels. Brush the rectangular area into the color you like.
8. Using the text tool from the left column tool bar, type the text “Adobe Photoshop” in the text box, adjust the font to 48 pt Times New Roman.
9. Using the blur tool from the left column tool bar, click the background layer and blur the dog’s face in the image.
10. Save the image into a PNG format file named “training task.png”.

APPENDIX F

EXPERIMENT TASKS FOR STUDY I AND II

Experiment Tasks for Adobe Acrobat

1. Open file "Rotate.pdf". Turn the grid on. Rotate the first page of the file 90 degrees clockwise. Resize the file to 150% of original size. Save the file as its original name. Turn the grid off and close the file.
2. Open the file "Page.pdf". Delete the first two pages and add file "Last page" to the current document. Place the "Last page" at the end of file "Page.pdf". Add a header and place it in the center to the current document with today's date. Save the file as its original name and close the file.
3. Open file "Find.pdf". Find the word "error" in the document. Replace the word "error" with "mistake". Find the sentence "Training delivery method was found to significantly impact the number of correct responses for scenario questions" and highlight the sentence. Strike though the conclusion section. Save the file as its original name and close the file.
4. Open file "Copy.pdf". Copy Figure 1 and paste it to the word document. Copy the "Exploratory Result" section and paste it to the same word document (Don't worry about formatting). Underline the Reference Section in "Copy.pdf". Save the file as its original name and close the file.
5. Open file "Draw.pdf". Draw a rectangle and a circle separately. Insert a text box and type "this is a text box" into the text box. Make the text ***Bold Italic***. Save the file as its original name and close the file.
6. Covert the word document "Word.doc" into a PDF file "Word.pdf".

Experiment Tasks for Adobe Photoshop

1. Open the file "Fish.psd". Rotate the image 90 degrees counterclockwise. Adjust the image width to 600 pixels. Adjust the canvas size to 10 inches width and 6 inches height. Save the image as its original name and close the image.
2. Open the file "Layer.psd". Add a new layer named "edit layer" with red color, dissolve mode, and 80% opacity. Save the image as its original name and close the image.
3. Open the file "Koala.psd". Select an elliptical area. Choose any blue color and paint the elliptical area in blue. Save the image as its original name and close the image.
4. Open the file "Horse.psd". Use the magnetic lasso tool to cut the horse out. Delete the horse and use "content aware" option with normal mode and 100% opacity. Save the image as its original name and close the image.
5. Open the file "Duck.psd". Insert a horizontal text box and type "photoshop" into the text box. Make the text ***Bold Italic***. Insert a rectangle filled with the color of your choice. Save the image as its original name and close the image.
6. Open the file "Penguin.psd". Blur left penguin and sharpen the right penguin. Save the image into a BMP format "Penguin.bmp" and close the image.

APPENDIX G

VARIMAX ROTATED FACTOR PATTERN FOR 3, 5, AND 6 FACTORS

Rotated Factor Pattern for 3-factors Structure

	Factor1	Factor2	Factor3
s3	**0.90896**	-0.03322	0.17594
s2	**0.89024**	0.10376	0.24085
s1	**0.88543**	0.21168	0.02167
s4	**0.81348**	0.13690	0.06433
s7	**0.80793**	0.33726	0.15886
s6	**0.75292**	0.48027	-0.17843
s10	**0.75233**	0.11400	0.30532
s11	**0.72760**	0.44363	0.01099
s13	0.27707	-0.10890	0.17720
s15	0.20717	**0.84281**	0.22328
s14	0.25900	**0.82611**	0.12841
s5	0.18568	**0.80223**	0.19461
s16	-0.02972	**0.74228**	0.13833
s8	0.08247	0.27896	**0.77475**
s9	0.02756	0.20748	**0.70718**
s12	0.38751	0.09800	**0.70687**

Rotated Factor Pattern for 5-factors Structure

	Factor1	Factor2	Factor3	Factor4	Factor5
s3	**0.90303**	-0.05169	0.17119	0.08805	0.11025
s1	**0.89897**	0.18538	0.05774	-0.02920	0.01783
s2	**0.89338**	0.08418	0.20504	0.04852	0.18055
s4	**0.82843**	0.11888	0.00966	0.06427	0.14276
s7	**0.79821**	0.32359	0.19470	0.11345	0.02049
s6	**0.75167**	0.46189	-0.04019	0.06254	-0.23575
s10	**0.71877**	0.09046	0.50929	-0.08248	-0.13774
s11	**0.69799**	0.44028	0.11216	0.26248	-0.16585
s15	0.22281	**0.83853**	0.19560	-0.02945	0.10894
s14	0.30253	**0.81589**	0.02205	-0.08744	0.21345
s5	0.22938	**0.78561**	0.14389	-0.23062	0.17869
s16	-0.07415	**0.76779**	0.19001	0.35502	-0.13624
s12	0.33936	0.09205	**0.79640**	-0.02763	0.12163
s8	0.02882	0.29114	**0.78377**	0.10850	0.20532
s13	0.19055	-0.05707	0.03115	**0.92347**	0.09141
s9	0.06549	0.21974	0.25797	0.08967	**0.86952**

Rotated Factor Pattern for 6-factors Structure

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6
s3	**0.91026**	-0.07059	0.17743	-0.02737	0.13192	0.06938
s1	**0.90656**	0.15803	0.07317	0.04563	0.02822	-0.04849
s2	**0.89693**	0.06885	0.21129	0.00504	0.19790	0.03511
s4	**0.81202**	0.18750	0.04581	-0.09103	0.10923	0.10260
s7	**0.79326**	0.30091	0.21340	0.12512	0.02571	0.10935
s6	**0.76786**	0.36819	-0.02711	0.26020	-0.21746	0.01676
s11	**0.72429**	0.28995	0.09946	0.36216	-0.11552	0.18639
s10	**0.68091**	0.15424	0.56327	-0.05113	-0.14902	-0.02881
s14	0.26759	**0.88803**	0.06105	0.14204	0.15011	-0.02320
s15	0.18623	**0.88018**	0.23096	0.22281	0.06334	0.02536
s5	0.27291	**0.58363**	0.08882	0.46201	0.25318	-0.34209
s12	0.28642	0.17939	**0.83344**	-0.05253	0.11456	0.04720
s8	0.06035	0.07261	**0.70665**	0.42235	0.33169	-0.00426
s16	0.01278	0.35809	0.07633	**0.85339**	0.02222	0.13300
s9	0.07430	0.20147	0.20002	0.04910	**0.89456**	0.07095
s13	0.17666	-0.04210	0.03119	0.09344	0.07437	**0.94639**

APPENDIX H

REORDERED SYSTEM TRANSFERAILITY QUESTIONNAIRE

Participant ID #_____

The System Transferability Questionnaire (STQ)

Instructions:

This questionnaire, which starts on the following page, gives you an opportunity to tell us your reactions to the software packages you used. Your responses will help us understand what aspects of the system you are particularly concerned about and the aspects that satisfy you.

To as great a degree as possible, think about all the tasks that you have done with **both software** while you answer these questions.

Please read each statement and indicate how strongly you agree or disagree with the statement by circling a number on the scale. If a statement does not apply to you, please write "N/A" in comments.

Please provide additional comments to elaborate on your answers.

After you have completed this questionnaire, I'll go over your answers with you to make sure I understand all of your responses.

Thank you!

1. Overall, I am satisfied with how easy it is to use the second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY AGREE**

COMMENTS:

2. It is simple to use the second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY AGREE**

COMMENTS:

3. I can quickly complete the task when using the second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY AGREE**

COMMENTS:

4. I can correctly complete all tasks when using the second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**STRONGLY AGREE**

COMMENTS:

5. I felt frustrated using the second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

6. I can quickly learn how to use the second software package after I changed from using the first software package to the second software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

7. I felt more efficient using second software package after using the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

8. The process of transferring to use the second software package after using the first software package is frustrating and makes me lost.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

13. The visual display and layout are generally consistent between the two software.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

14. The second software package presents information that is consistent to the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

15. I will easily confuse some functions in the second software package with the functions in the first software package.

STRONGLY DISAGREE	**1**	**2**	**3**	**4**	**5**	**6**	**7** **STRONGLY AGREE**

COMMENTS:

APPENDIX I

SYSTEM USABILITY SCALE (SUS)

Participant ID #_____

Survey ID #_____

Software ID #_____

System Usability Scale (SUS)

Instructions:
Please think about **the software you just used,** read each statement below and indicate how strongly you agree or disagree with the statement by circle a number on the scale. If a statement does not apply to you, please circle "N/A".

	Strongly Disagree				Strongly Agree	N/A
1. I think I would like to use this software frequently.	1	2	3	4	5	N/A
2. I found the software unnecessarily complex.	1	2	3	4	5	N/A
3. I thought the software was easy to use.	1	2	3	4	5	N/A
4. I think I would need Tech Support to be able to use this software.	1	2	3	4	5	N/A
5. I found the various functions in this software were well integrated.	1	2	3	4	5	N/A
6. I thought there was too much inconsistency in this software.	1	2	3	4	5	N/A
7. I would imagine that most people would learn to use this software very quickly.	1	2	3	4	5	N/A
8. I found the software very cumbersome to use.	1	2	3	4	5	N/A
9. I felt very confident using the software.	1	2	3	4	5	N/A
10. I need to learn a lot about this software before I could effectively use it.	1	2	3	4	5	N/A

APPENDIX J

TOTAL USABILITY SCORE CALCULATION SHEET

Table J.1 TUS Score Calculation

	Raw Score				Standardized Score				
Subject ID	Avg SUS Score	Avg STQ Score	Avg CTPS	Average Errors	Avg SUS Score	Avg STQ Score	Avg CTPS	Average Errors	Total Usability Score
1	56.25	4.25	8.22	16.83	0.25	0.53	0.67	0.59	0.49
2	90.00	5.02	9.51	7.58	0.86	0.70	0.55	0.86	0.76
3	67.50	4.69	9.72	17.42	0.45	0.63	0.54	0.57	0.54
4	66.25	4.88	9.57	6.33	0.43	0.67	0.55	0.90	0.63
5	83.75	5.25	7.74	6.75	0.75	0.76	0.71	0.88	0.78
6	60.00	3.00	7.98	10.75	0.32	0.24	0.69	0.77	0.49
7	57.50	3.75	12.74	37.08	0.27	0.41	0.27	0.00	0.24
8	60.00	5.00	13.71	17.33	0.32	0.70	0.19	0.58	0.45
9	58.75	3.44	6.52	10.58	0.30	0.34	0.81	0.77	0.53
10	82.50	5.88	14.13	10.42	0.73	0.90	0.15	0.78	0.67
11	76.25	4.75	9.95	13.17	0.61	0.64	0.52	0.70	0.62
12	88.75	5.19	9.15	14.67	0.84	0.74	0.59	0.65	0.72
13	47.50	3.69	13.53	30.75	0.09	0.40	0.21	0.18	0.21
14	55.00	3.19	8.73	12.00	0.23	0.29	0.62	0.73	0.45
15	58.75	2.44	8.02	14.00	0.30	0.11	0.68	0.67	0.42
16	73.75	4.19	10.94	8.42	0.57	0.51	0.43	0.83	0.59
17	68.75	4.06	13.09	9.33	0.48	0.49	0.25	0.81	0.52

Table L.1 (Continued)

18	60.00	3.31	9.91	14.50	0.32	0.31	0.52	0.66	0.44
19	67.50	5.06	9.79	15.33	0.45	0.71	0.53	0.63	0.58
20	68.75	5.63	10.71	17.17	0.48	0.84	0.45	0.58	0.59
21	42.50	2.44	11.91	15.33	0.00	0.11	0.35	0.63	0.26
22	62.50	3.75	7.77	6.75	0.36	0.41	0.70	0.88	0.57
23	97.50	5.94	7.18	3.00	1.00	0.91	0.76	0.99	0.93
24	51.25	5.00	7.45	6.42	0.16	0.70	0.73	0.89	0.59
25	66.25	3.44	12.29	17.42	0.43	0.34	0.31	0.57	0.42
26	56.25	4.81	14.04	18.33	0.25	0.66	0.16	0.55	0.41
27	62.50	3.50	10.71	16.25	0.36	0.36	0.45	0.61	0.44
28	66.25	4.19	10.74	14.33	0.43	0.51	0.45	0.66	0.51
29	65.00	4.56	14.56	17.08	0.41	0.60	0.12	0.58	0.44
30	51.25	4.69	15.92	5.67	0.16	0.63	0.00	0.92	0.43
31	62.50	2.88	9.20	13.92	0.36	0.21	0.58	0.67	0.45
32	68.75	3.94	9.91	21.33	0.48	0.46	0.52	0.46	0.48
33	53.75	1.94	10.24	18.92	0.20	0.00	0.49	0.53	0.29
34	63.75	4.81	6.97	11.83	0.39	0.66	0.77	0.74	0.62
35	61.25	4.00	7.99	15.92	0.34	0.47	0.69	0.62	0.51
36	53.75	3.25	6.81	10.33	0.20	0.30	0.79	0.78	0.49
37	72.50	5.19	9.27	14.17	0.55	0.74	0.57	0.67	0.63
38	65.00	5.00	6.15	7.58	0.41	0.70	0.85	0.86	0.68
39	72.50	5.56	7.31	9.33	0.55	0.83	0.74	0.81	0.72
40	71.25	3.13	5.90	9.75	0.52	0.27	0.87	0.80	0.60

Table L.1 (Continued)

41	62.50	3.44	11.86	22.75	0.36	0.34	0.35	0.42	0.37
42	92.50	6.00	4.36	2.75	0.91	0.93	1.00	1.00	0.95
43	47.50	3.63	8.93	18.00	0.09	0.39	0.60	0.56	0.38
44	70.00	4.50	8.76	10.92	0.50	0.59	0.62	0.76	0.61
45	58.75	3.63	7.23	5.75	0.30	0.39	0.75	0.91	0.56
46	50.00	4.38	5.38	7.67	0.14	0.56	0.91	0.86	0.58
47	75.00	6.25	8.02	8.92	0.59	0.99	0.68	0.82	0.77
48	60.00	2.69	5.04	8.08	0.32	0.17	0.94	0.84	0.54
49	57.50	2.61	12.09	15.92	0.27	0.15	0.33	0.62	0.34
50	48.75	3.25	11.51	12.58	0.11	0.30	0.38	0.71	0.36
51	48.75	2.69	6.52	2.83	0.11	0.17	0.81	1.00	0.49
52	75.00	6.31	11.49	23.08	0.59	1.00	0.38	0.41	0.61
53	56.25	2.00	12.53	19.33	0.25	0.01	0.29	0.52	0.27
54	70.00	4.44	10.14	23.00	0.50	0.57	0.50	0.41	0.50

APPENDIX K

ONLINE DEMOGRAPHIC SURVEY FOR STUDY III

Machine Usability Study Exit this survey

1. Machine Usability Study Recruitment Survey

Thank you for taking the time to participate in this survey. The purpose of this survey is to select research participants. Research participants will perform some tasks using some simple machines. Results from the project will be used to improve the design of machine usability.

If you agree to participate in this study, we are asking that you take about 5 minutes and complete a brief online screening survey.

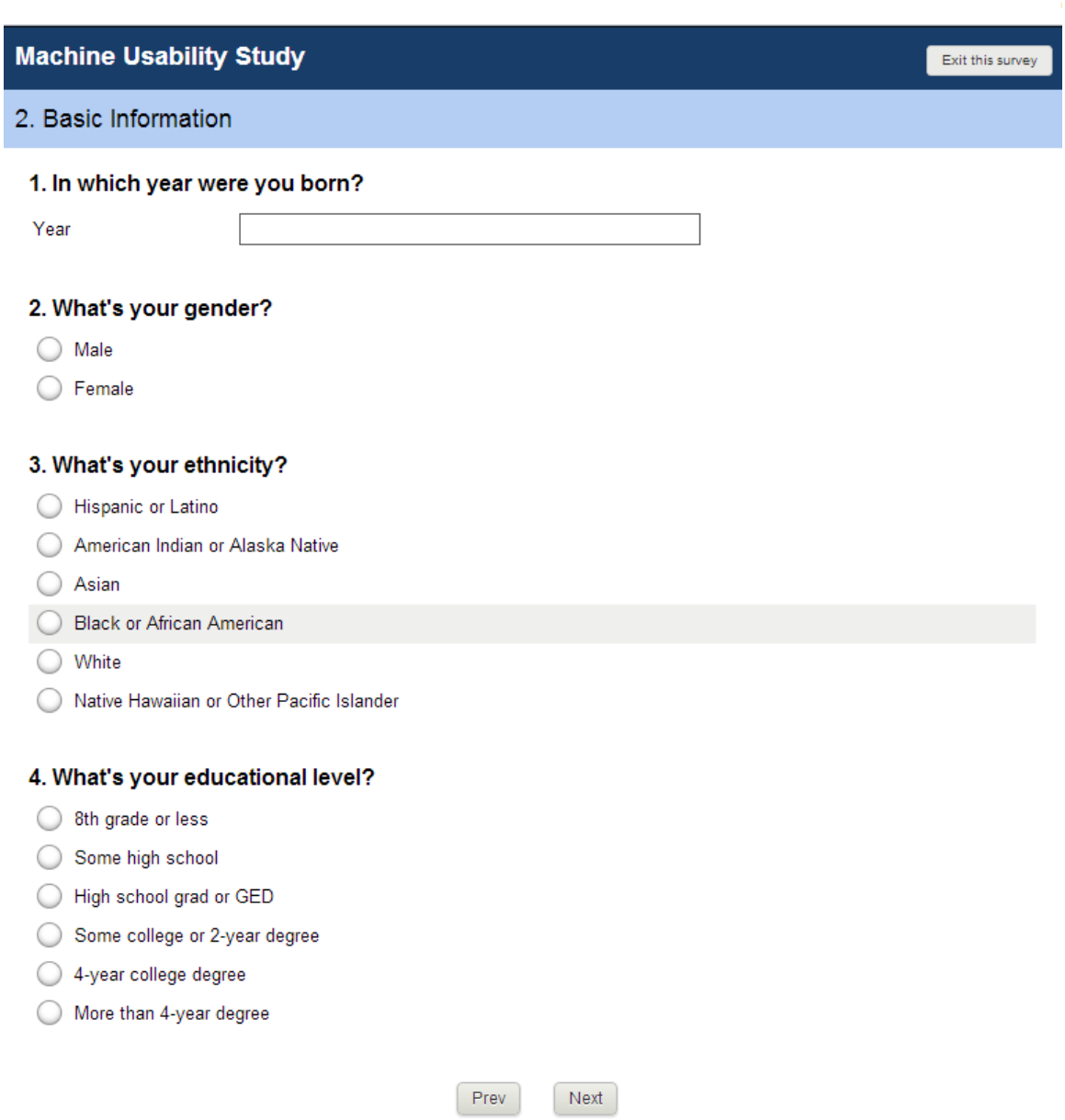
Your participation is completely voluntary and you can end the survey at any time by closing the browser. Your responses will remain confidential. All identifying information will be removed from your responses when your survey is submitted.

If you have any questions about this survey or the research study, please contact the researcher listed below. For additional information regarding your rights as a research subject, please feel free to contact the MSU Regulatory Compliance Office at (662)325-3294.

Yunchen Huang
Ph.D. Candidate
Industrial & Systems Engineering
Mississippi State University
yh95@msstate.edu

By entering the survey area, you indicate that you are at least eighteen years old and are giving your informed consent to be a participant in this study. If you would like a print copy of this document, please use the "print" function on your Internet browser.

Next



1. Have you ever worked in the following industry?

- [ ] Construction
- [ ] Manufacturing
- [ ] Auto repair
- [ ] Office/Secretary
- [ ] Art Design/Image processing
- [ ] Other
- [ ] None of these

2. If yes to the above question, for how long?

Construction	
Manufacturing	
Auto repair	
Office/Secretary	
Art design/Image processing	
Other	

Prev Next

1. Do you have any of the following hobbies?

- [ ] Woodwork
- [ ] Auto repair
- [ ] Welding
- [ ] None of these

Prev Next

1. Have you used a drill press before?

- ( ) Yes
- ( ) No

Prev Next

1. You mentioned you've used a drill press before. Have you used a drill press within the last 3 years?

- ( ) Yes
- ( ) No

Prev Next

1. On average, how often have you used a drill press within the last three years?

- ( ) Frequently (almost every day)
- ( ) Moderately (around once per week)
- ( ) Periodically (around every other month)
- ( ) Infrequently (less than once per year)

Prev Next

1. Have you used a lathe machine before?

- ○ Yes
- ○ No

Prev Next

1. You mentioned you've used a lathe machine before. Have you used a lathe machine within the last 3 years?

- ○ Yes
- ○ No

Prev Next

1. On average, how often have you used a lathe machine within the last three years?

- ○ Frequently (almost every day)
- ○ Moderately (around once per week)
- ○ Periodically (around every other month)
- ○ Infrequently (less than once per year)

Prev Next

1. Please indicate your typical available time during a week.

	Day	Time
First choice		
Second choice		
Third choice		

2. Please enter your name, email address, and telephone number so that we can contact your regarding the experiment schedule.

Prev Next

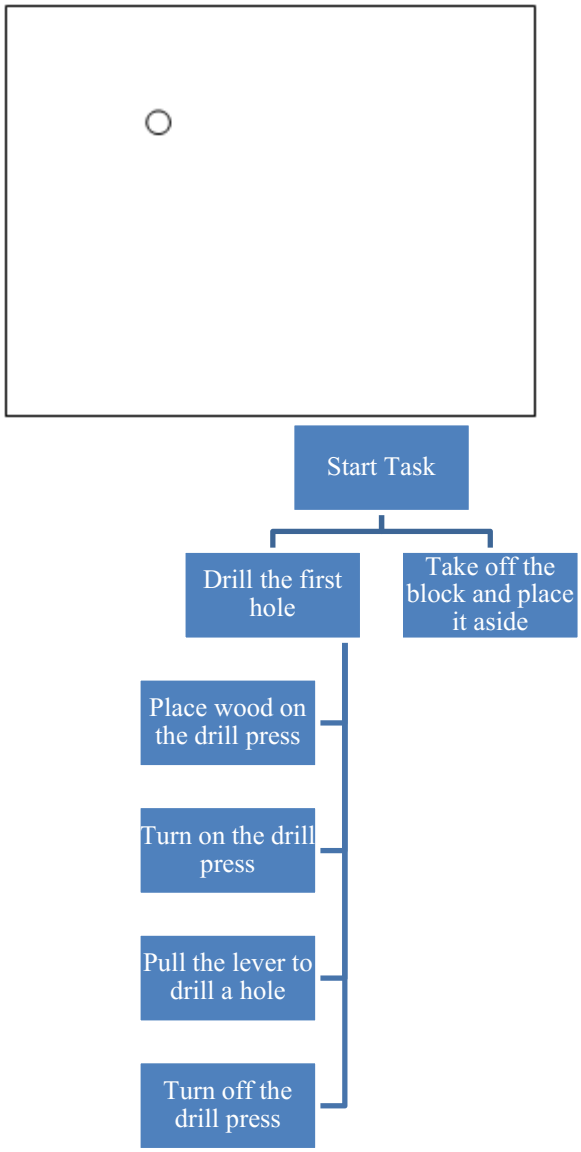
APPENDIX L

EXPERIMENT TASKS FOR STUDY III

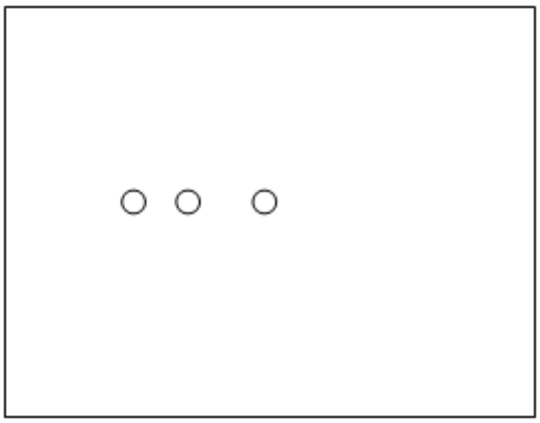
Low cognitive and low physical complexity

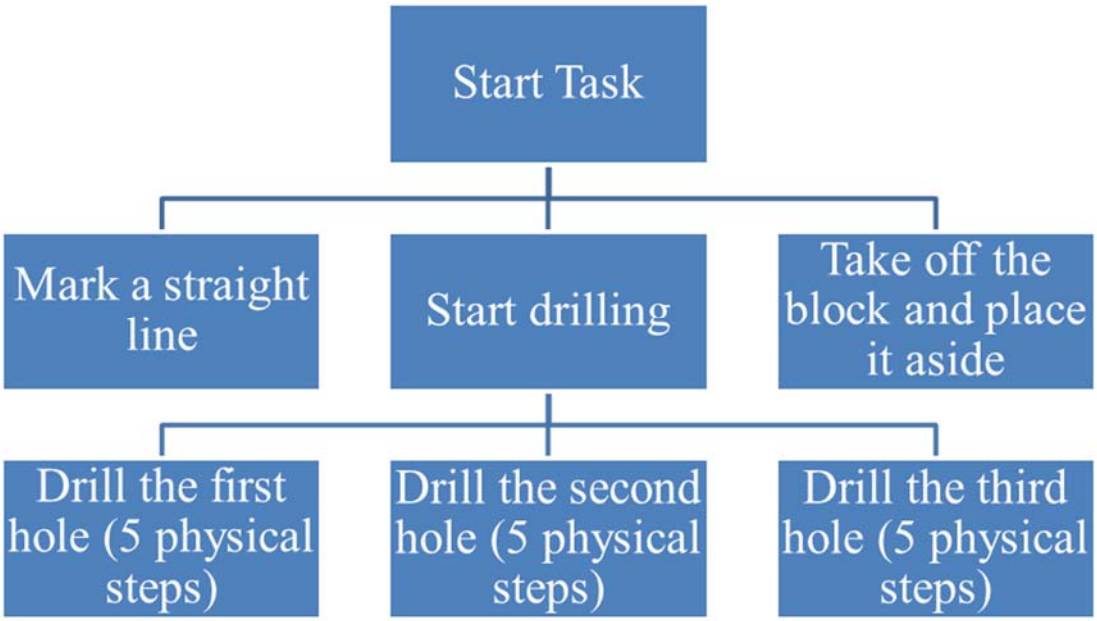
Drill Press Tasks and Hierarchy

1. Use the drill press to drill a hole anywhere in the work piece. The depth of the hole does not matter. (A: 5 physical steps 0 cognitive steps)

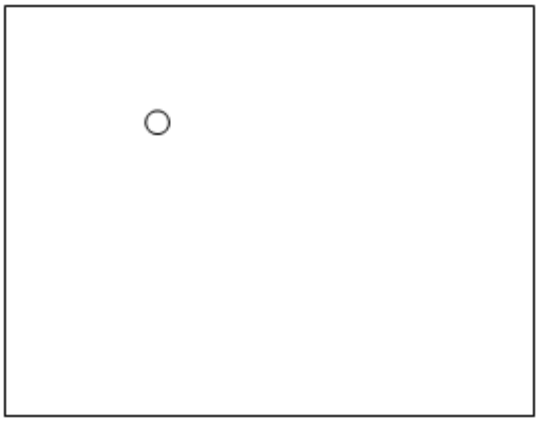


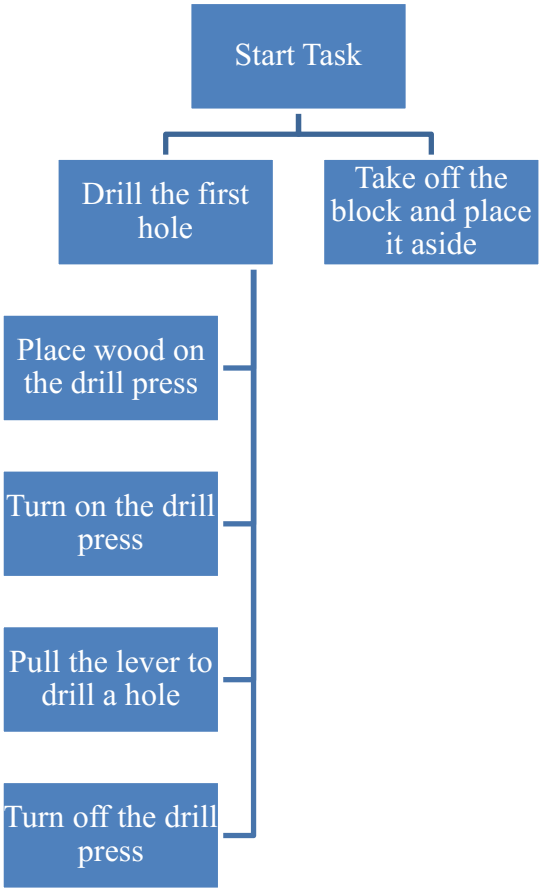
2. Use the drill press to drill three holes in a straight line in the work piece. The depth and spacing of the holes do not matter. (B: 17 physical steps 0 cognitive steps)





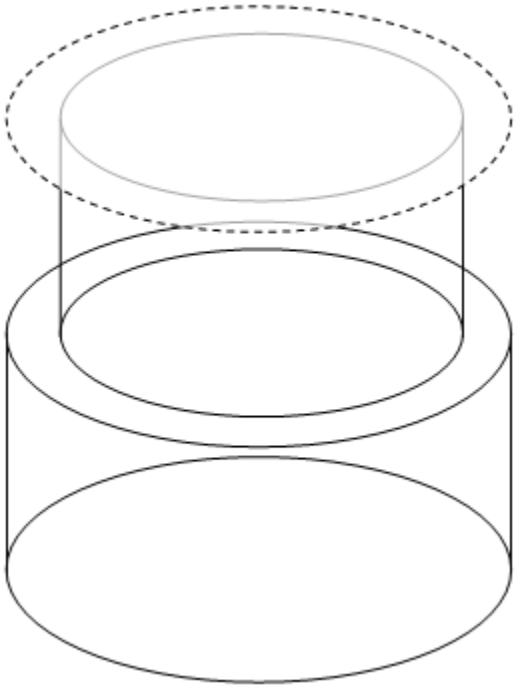
3. Use the drill press to drill one hole in one of the corner of the wood block, as shown in the figure below. The depth and exact location of the hole does not matter. (C: 5 physical steps 0 cognitive steps)

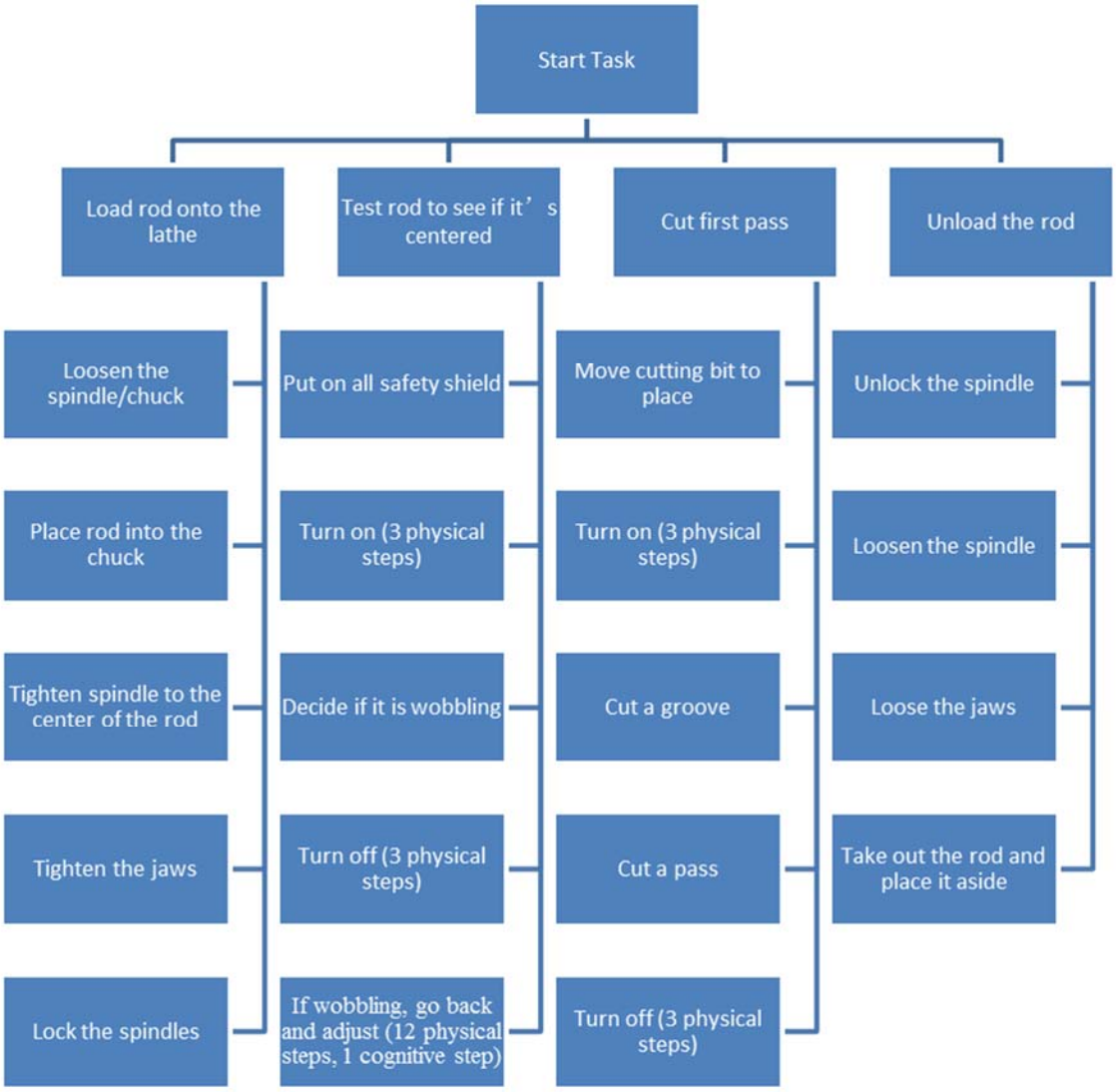




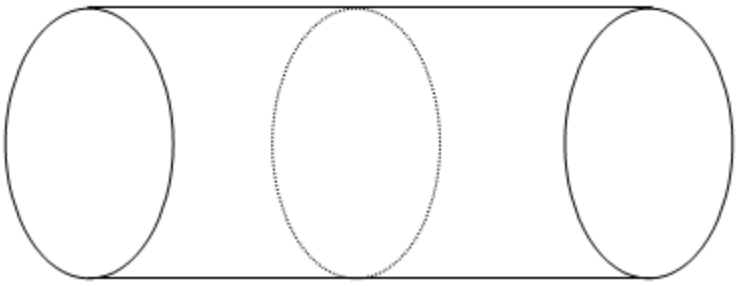
Mini-lathe Machine Tasks and Hierarchy

1. Use the lathe machine to carve the wood rod into the shape shown in the figure below. Dimension does not matter (But please note that lathe can only carve 2mm depth at a time). (D: 37 physical steps 2 cognitive steps)

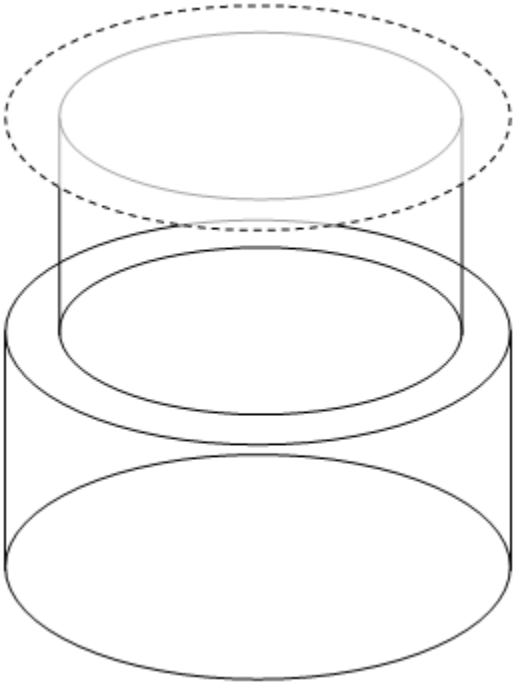




2. Use the lathe machine to carve one groove in the wood rod. The depth and width of the groove do not matter. (E: 36 physical steps 2 cognitive steps)



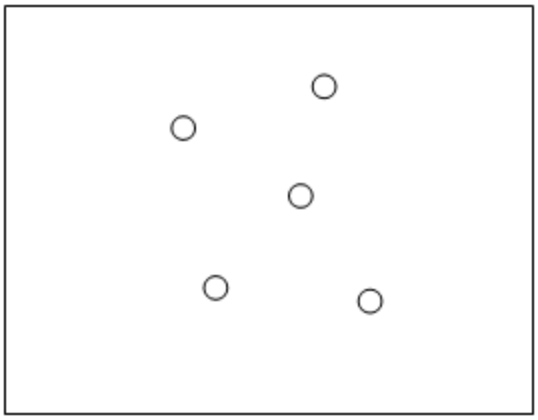
3. Use lathe machine to carve the wood rod into the shape shown in the figure below. Dimension does not matter. (But please note that lathe can only carve 2mm depth at a time). (F: 37 physical steps 2 cognitive steps)

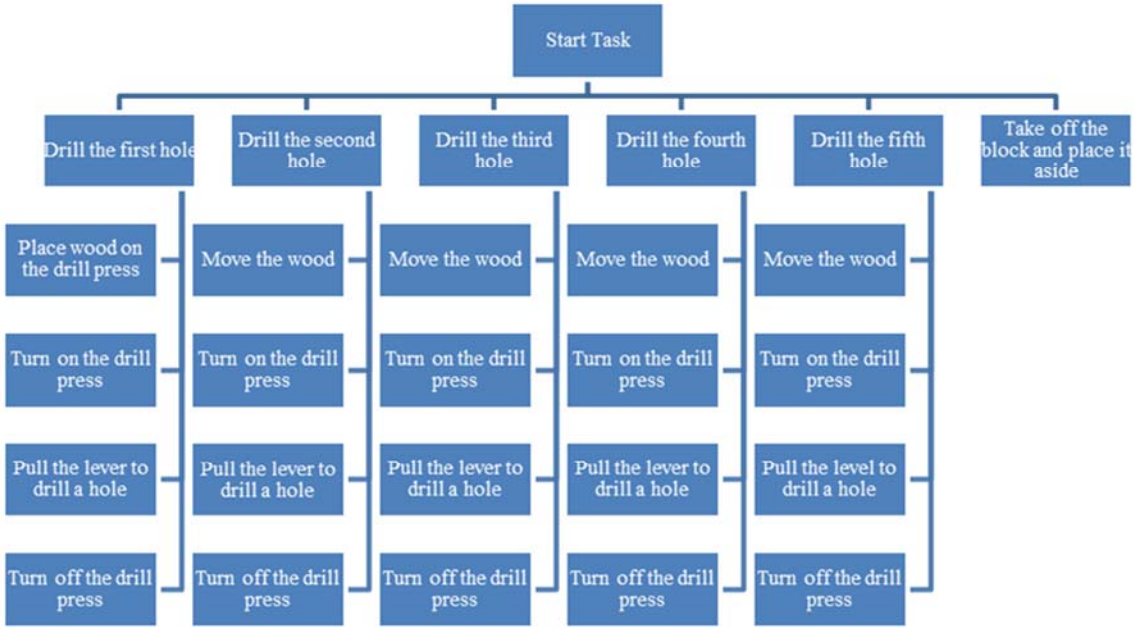


Low cognitive and high physical complexity

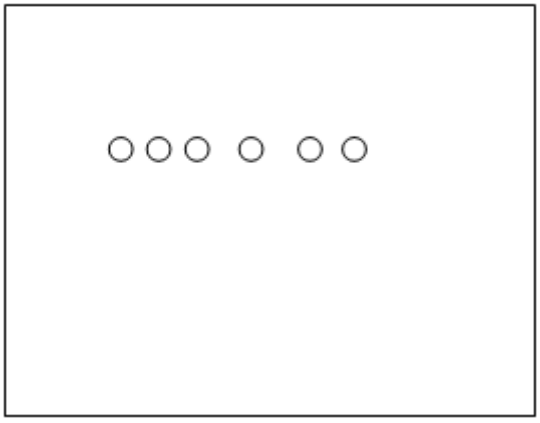
Drill Press Tasks and Hierarchy

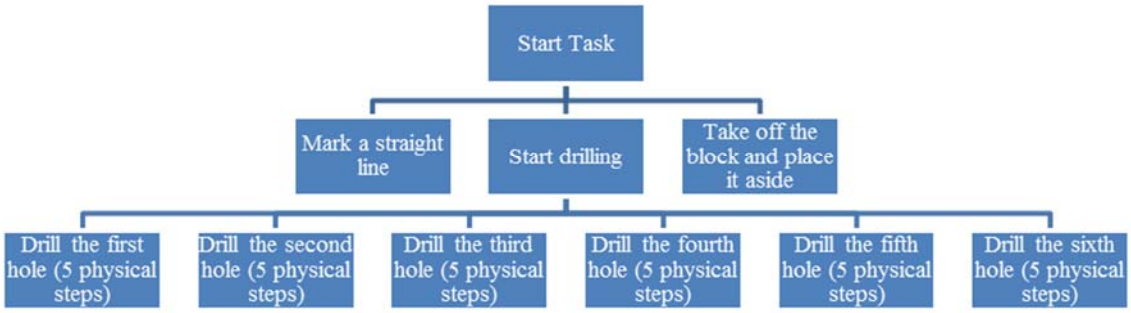
1. Use the drill press to drill five holes in the work piece. The positions and depths of the holes do not matter. See figure below. (C: 21 physical steps 0 cognitive steps)



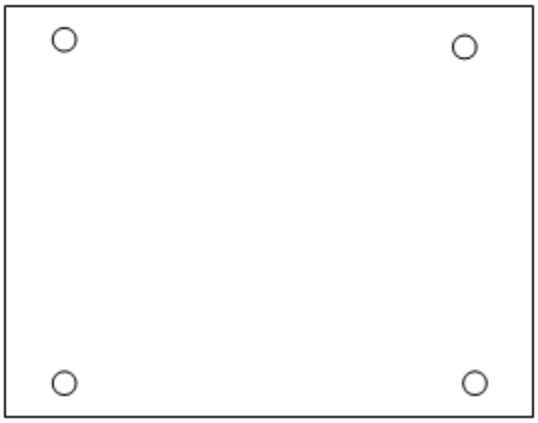


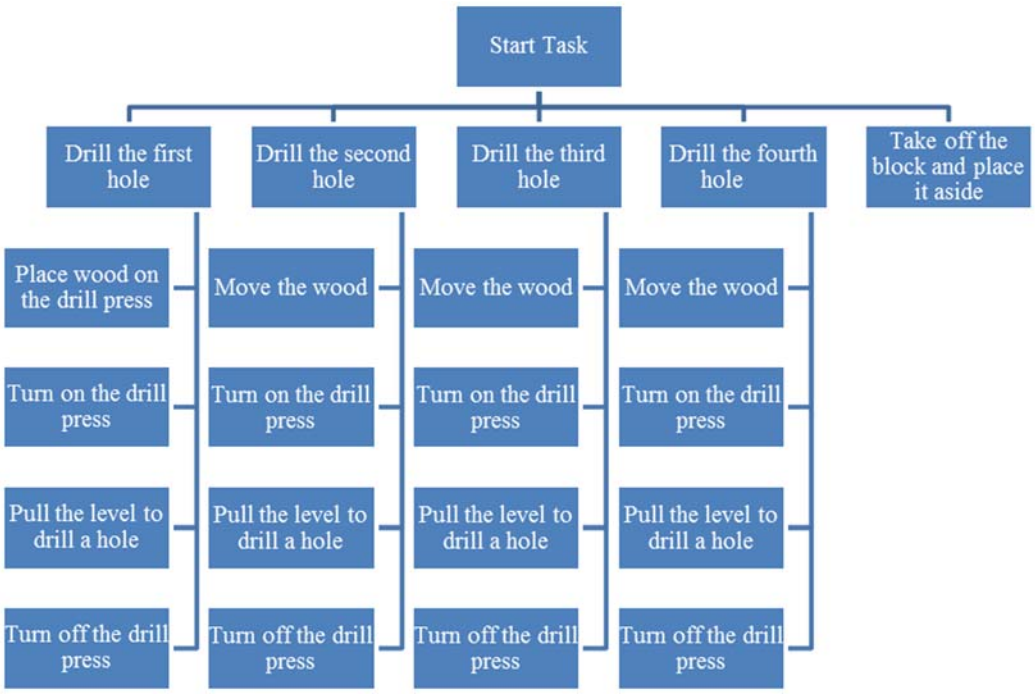
2. Use the drill press to drill six holes in a straight line in the work piece. The depth and spacing of the hole do not matter. See figure below. (B: 32 physical steps 0 cognitive step)





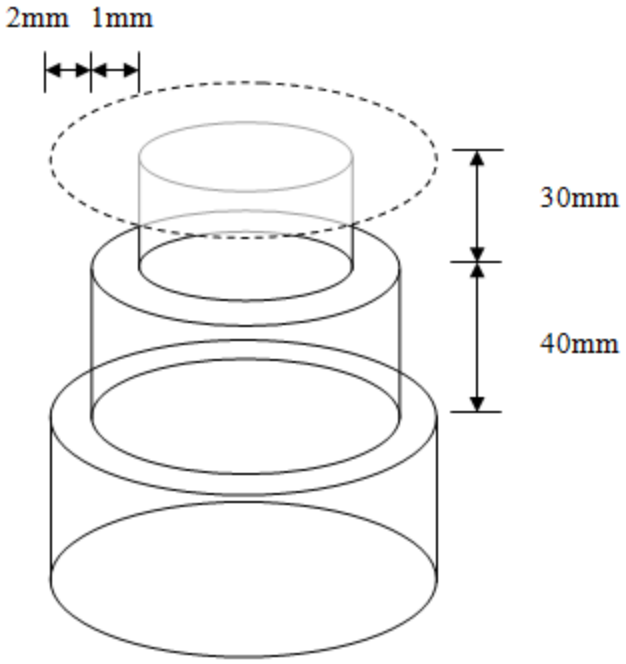
3. Use the drill press to drill one hole in each of the four corners in the work piece in the places noted in the figure below. Dimension does not matter. (A: 17 physical steps 0 cognitive steps)

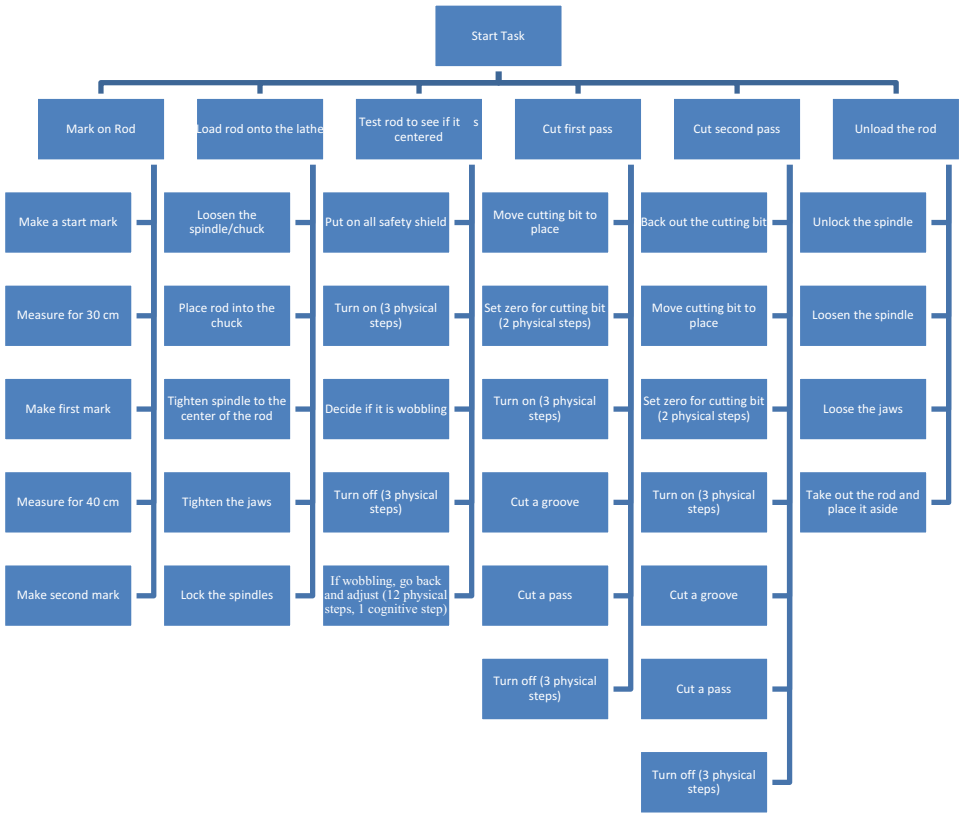




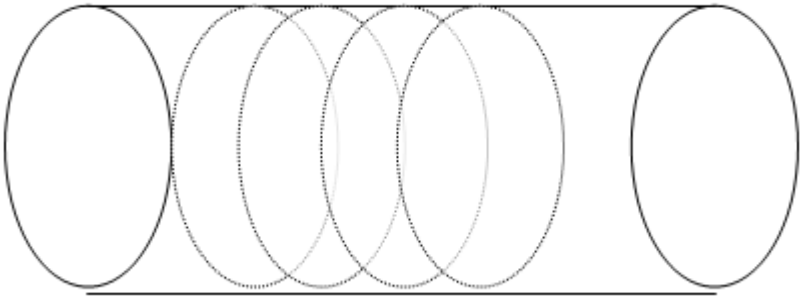
Mini-lathe Tasks and Hierarchy

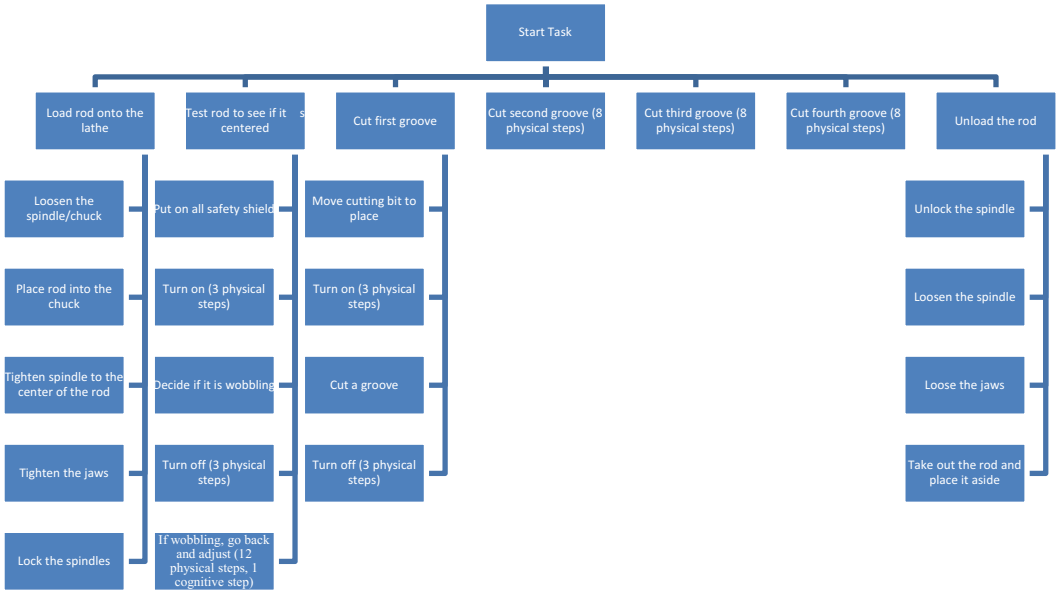
1. Use the lathe machine in manual feeding mode to **manually** carve the wood rod into the shape shown in the figure below (Please note that lathe machine can only carve a depth of 2mm at a time). (D: 52 physical steps 6 cognitive steps)



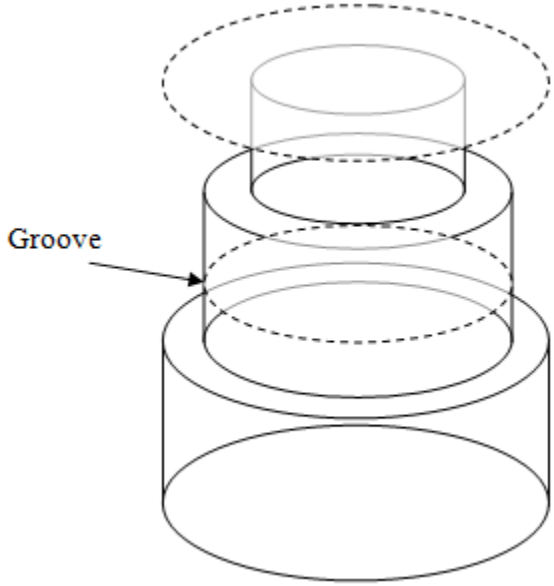


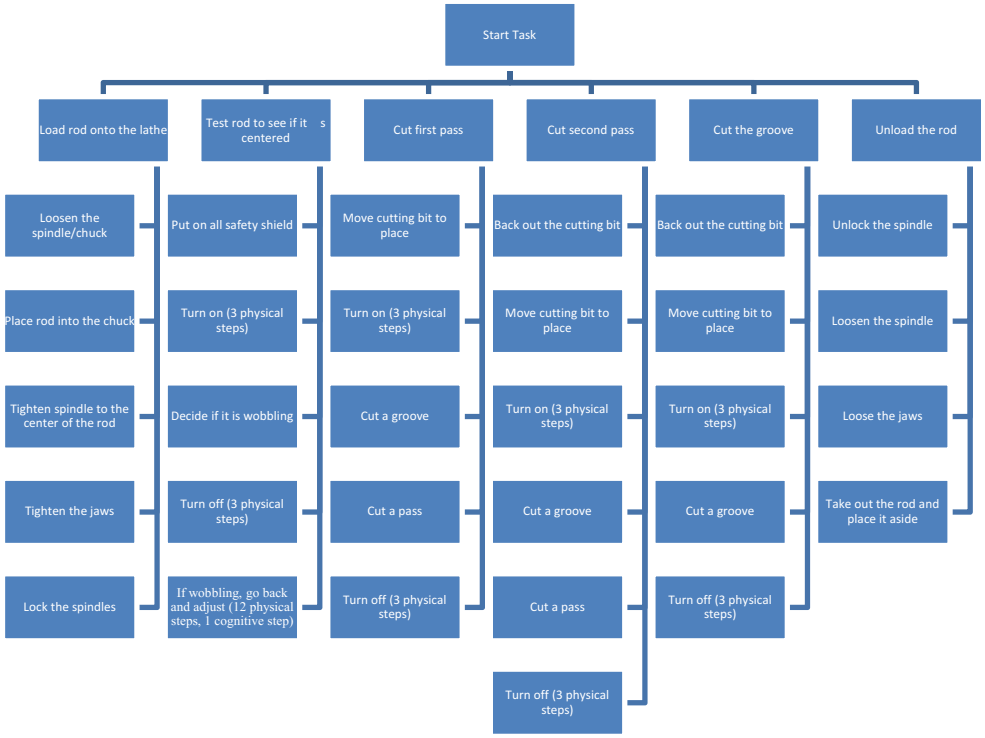
2. Use lathe machine to carve four grooves in the wood rod. The depth, width, and spacing of the grooves do not matter. Please see the figure below. (E: 60 physical steps 2 cognitive steps)





3. Use lathe machine in manual feeding mode to **manually** carve the wood rod into the shape shown in the figure below. Then carve one grove in the position noted in the figure. The dimension does not matter. (Please note that lathe machine can only carve a depth of 2mm at a time) (56 physical steps 2 cognitive steps)

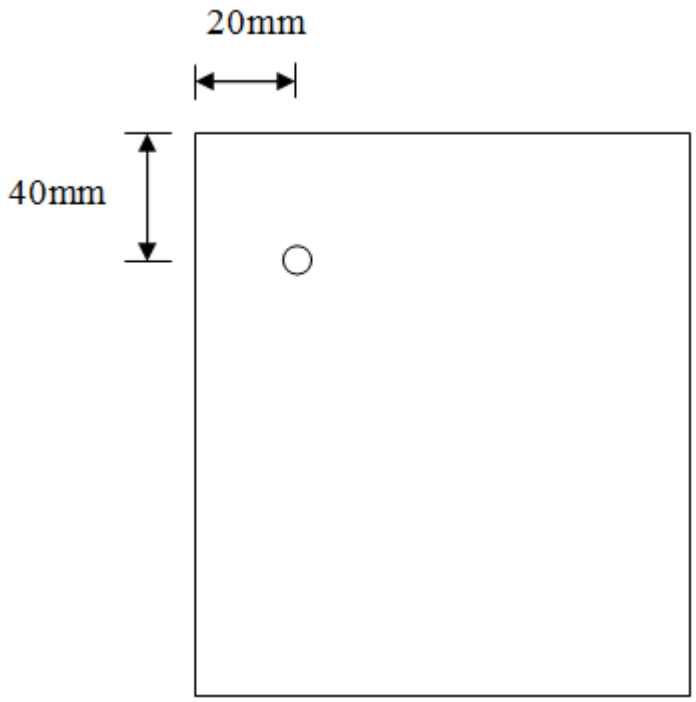


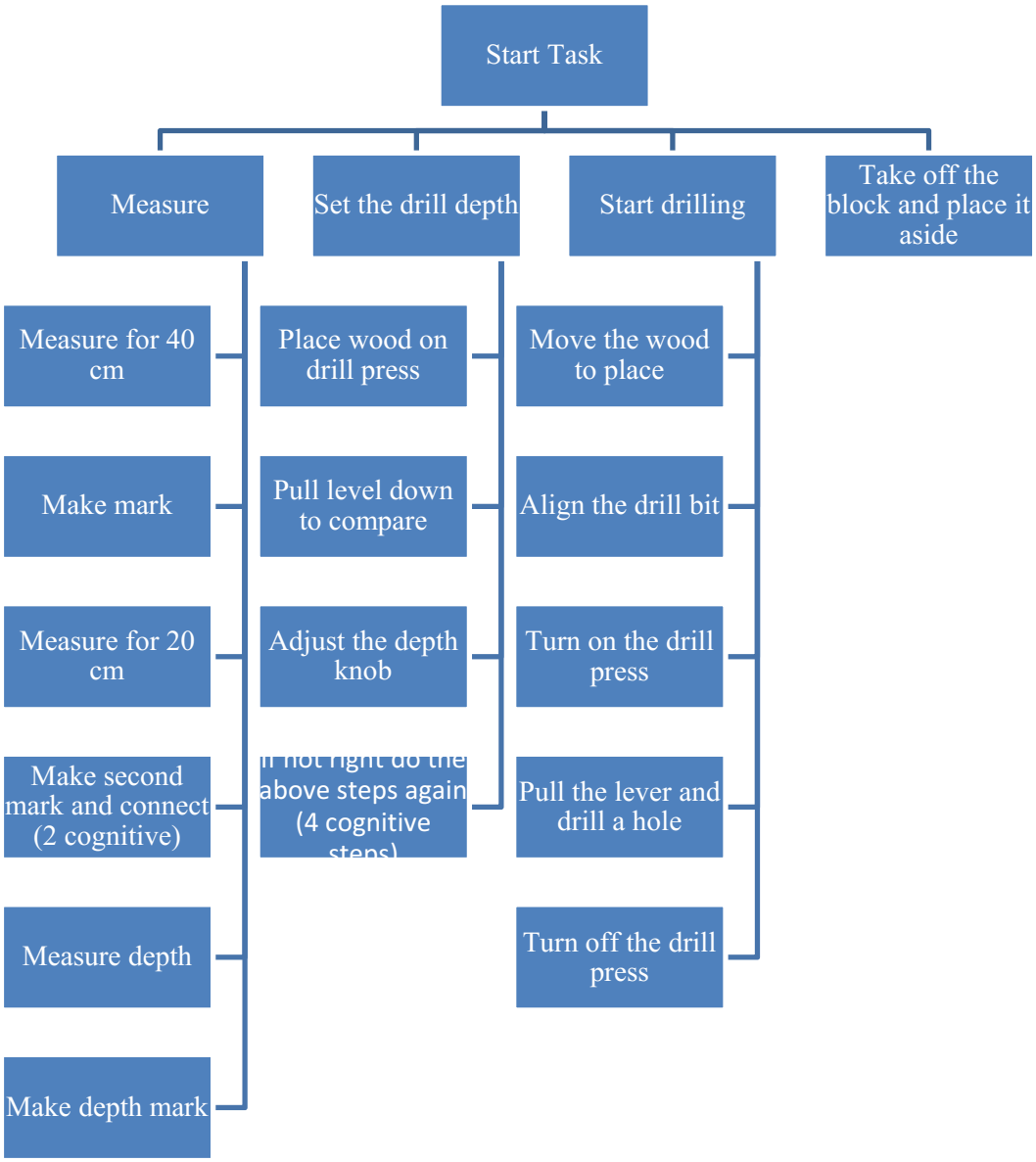


High cognitive and low physical complexity

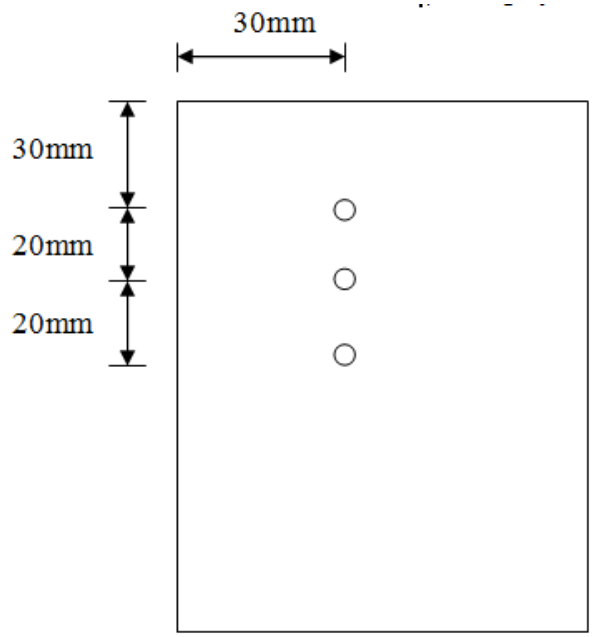
Drill Press Tasks and Hierarchy

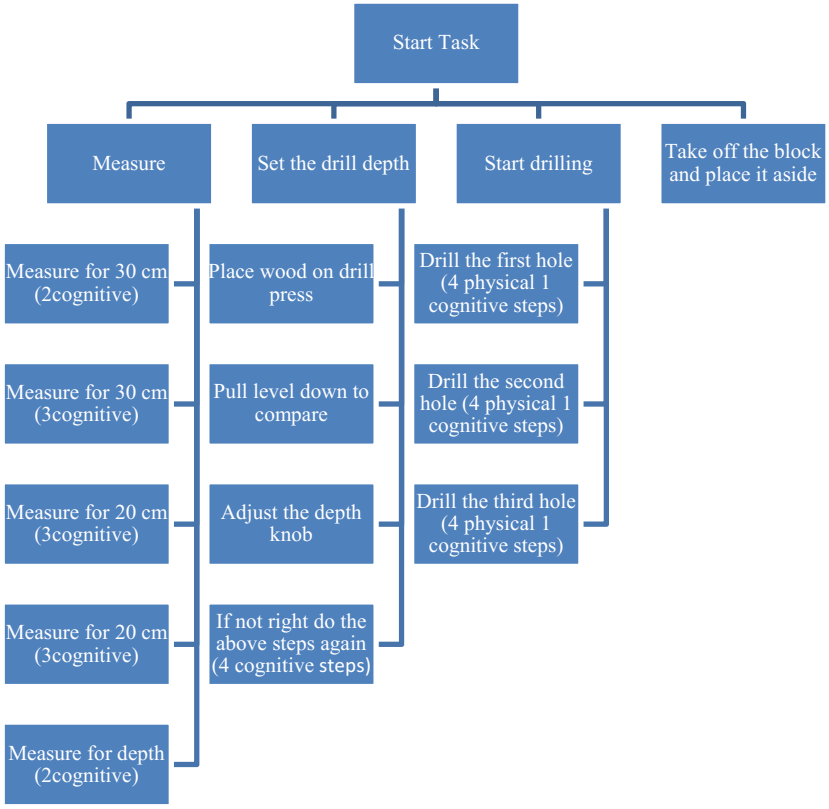
1. Use the drill press to drill a 5mm depth hole in the work piece at the place noted in the figure below. (A: 6 physical 14 cognitive)



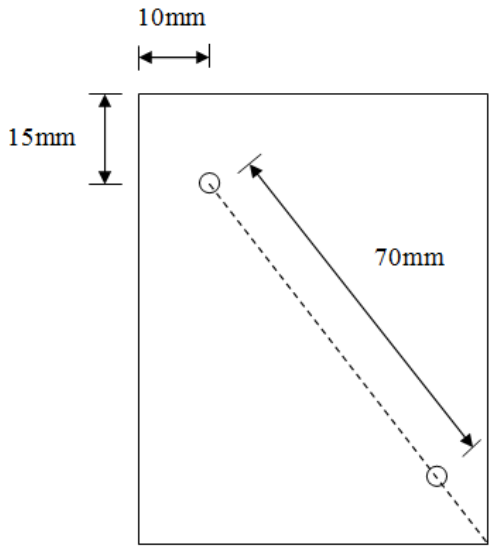


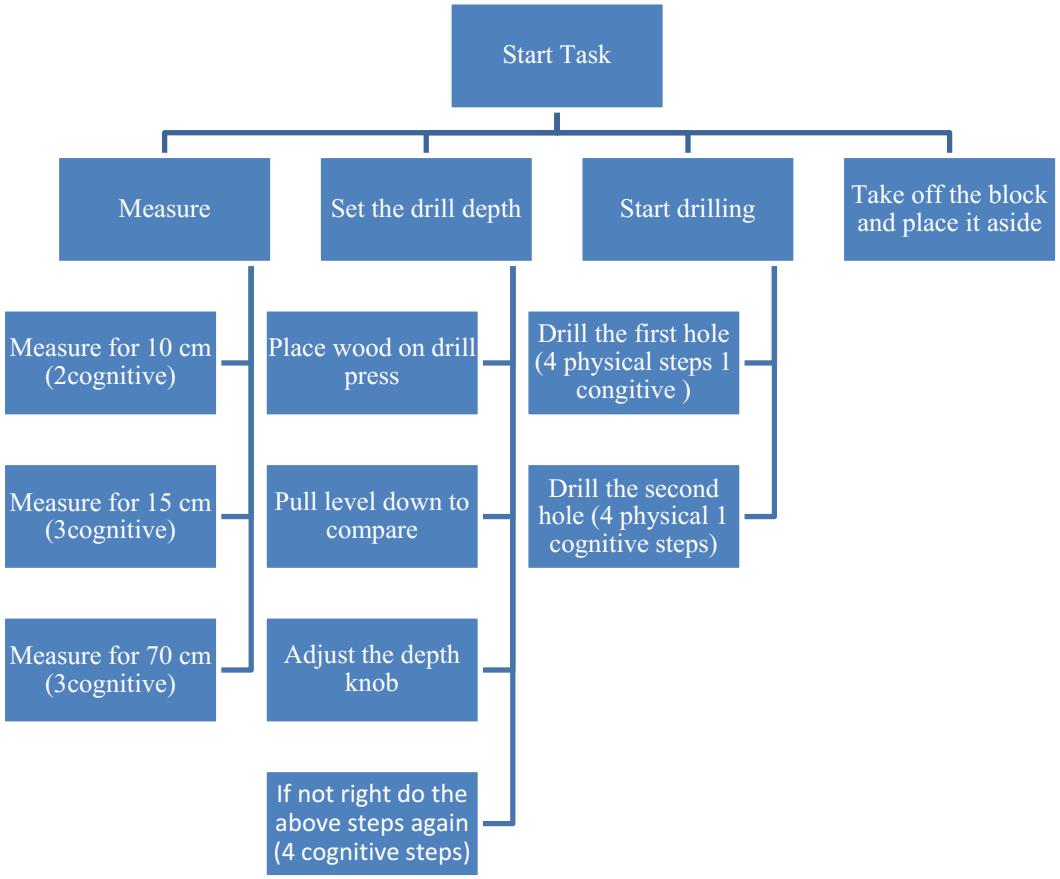
2. Use the drill press to drill three holes in a straight line in the work piece at the place noted in the figure below. The depth of the hole is 8mm. (B: 14 physical 22 cognitive)





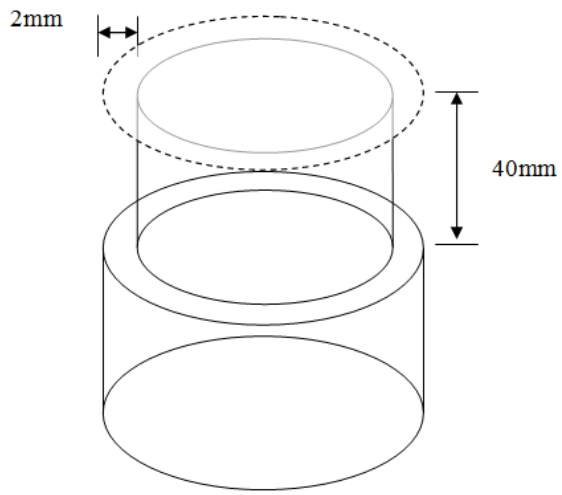
3.Use the drill press to drill two 6mm depth hole in the work piece at the place noted in the figure below. (C: 10 physical 16 cognitive)

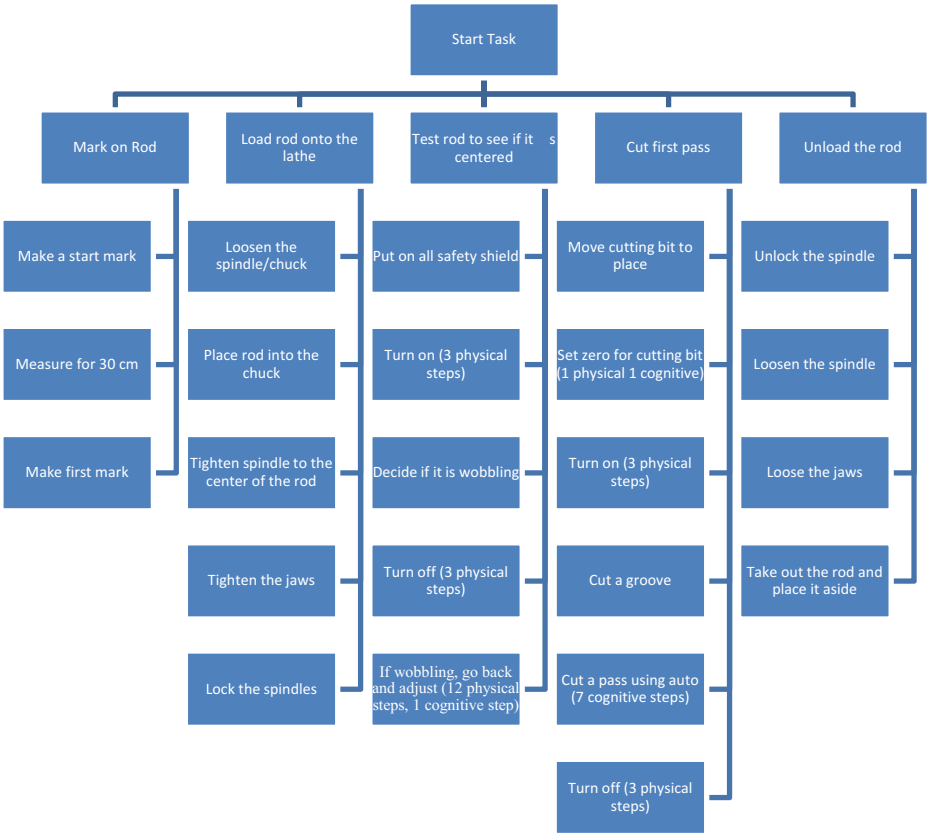




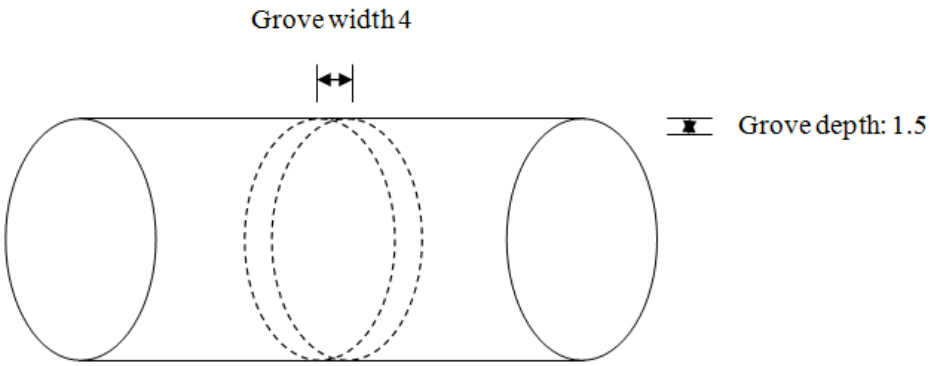
Mini-lathe Tasks and Hierarchy

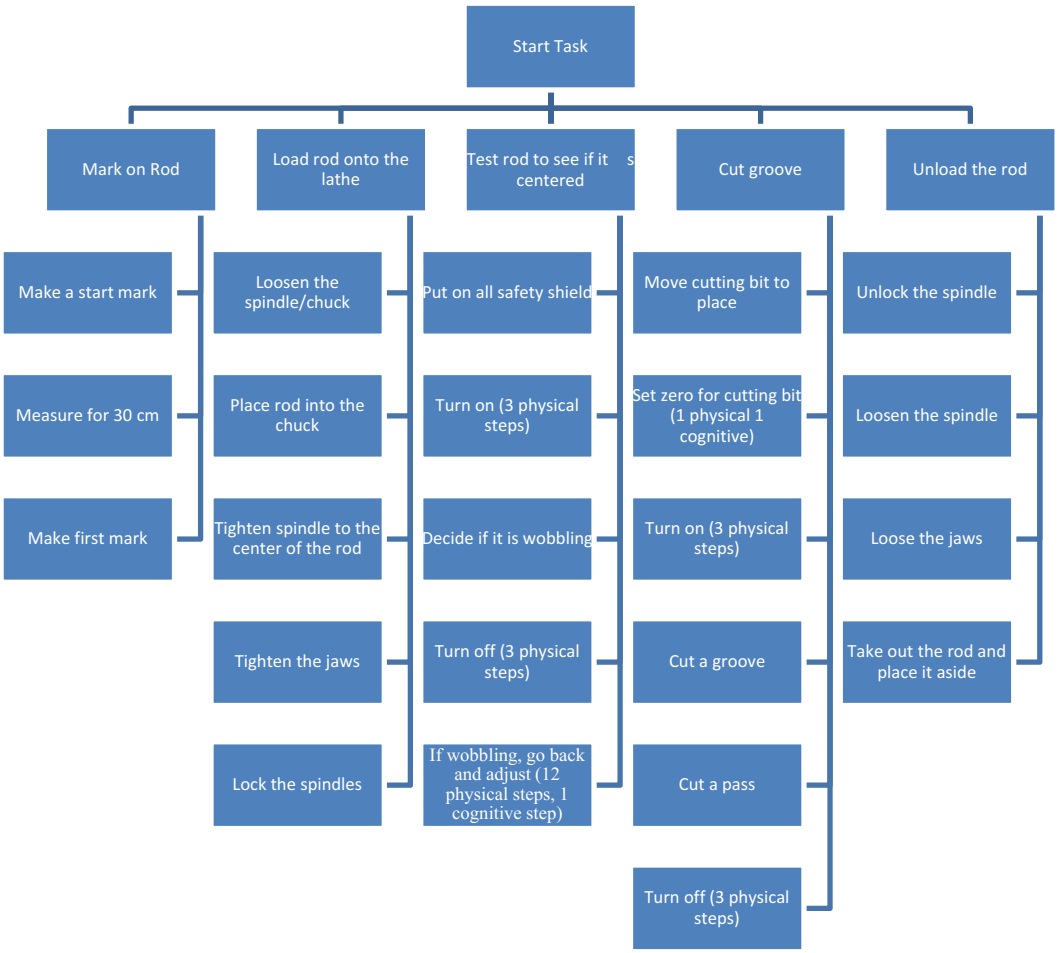
1. Use lathe machine in **automatic feeding mode** to carve the wood rod into the shape in the figure below. (D: 33 physical 13 cognitive)



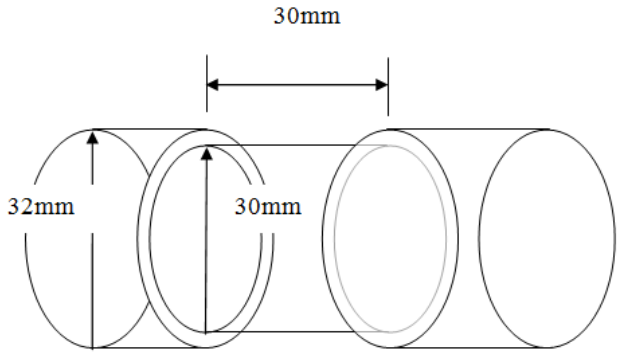


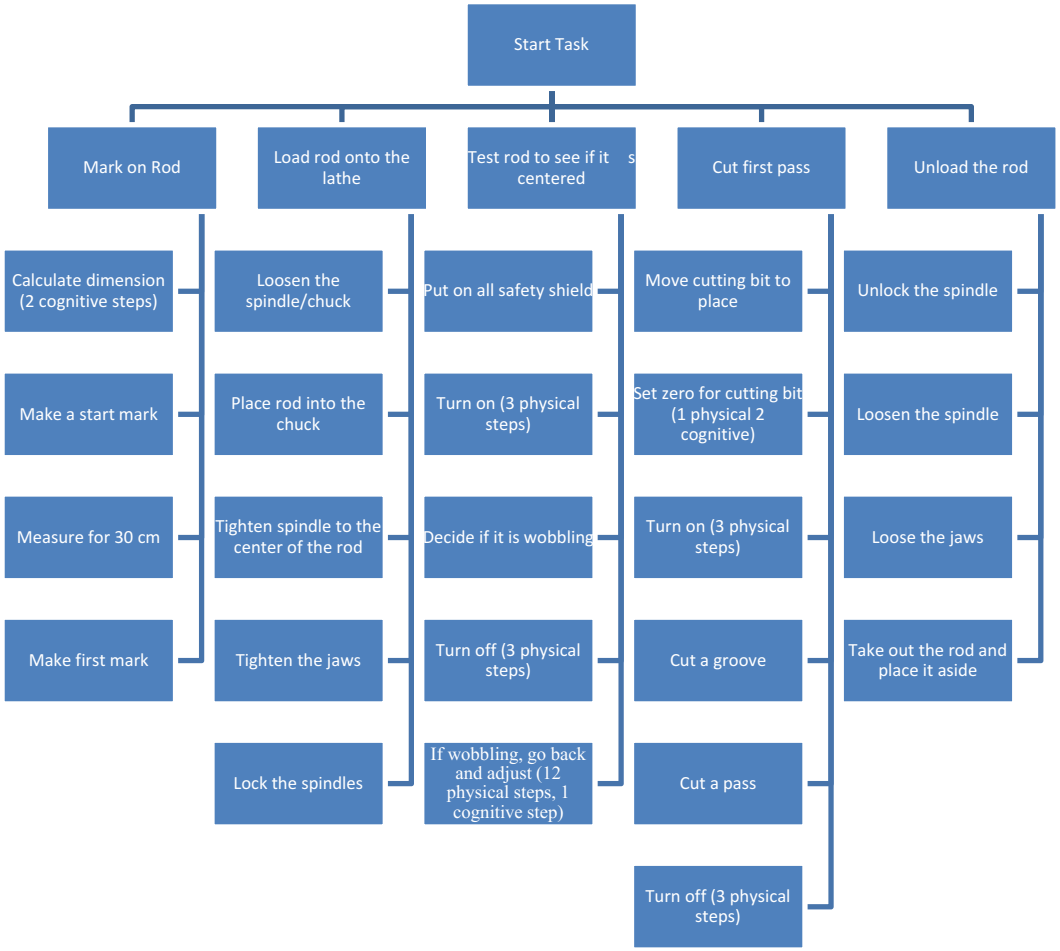
2. Use lathe machine to carve one grove in the wood rod. Please follow the dimension in the figure below. (E: 37 physical 7 cognitive)





3. Use the lathe machine in **automatic feeding mode** to carve out a rod section that has a length of 30 mm and a diameter of 30mm, as shown in the figure below (assume current diameter of the raw wood material is 32mm). (F: 37 physical 10 cognitive)

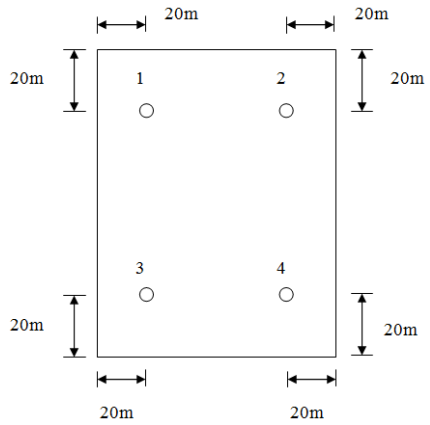


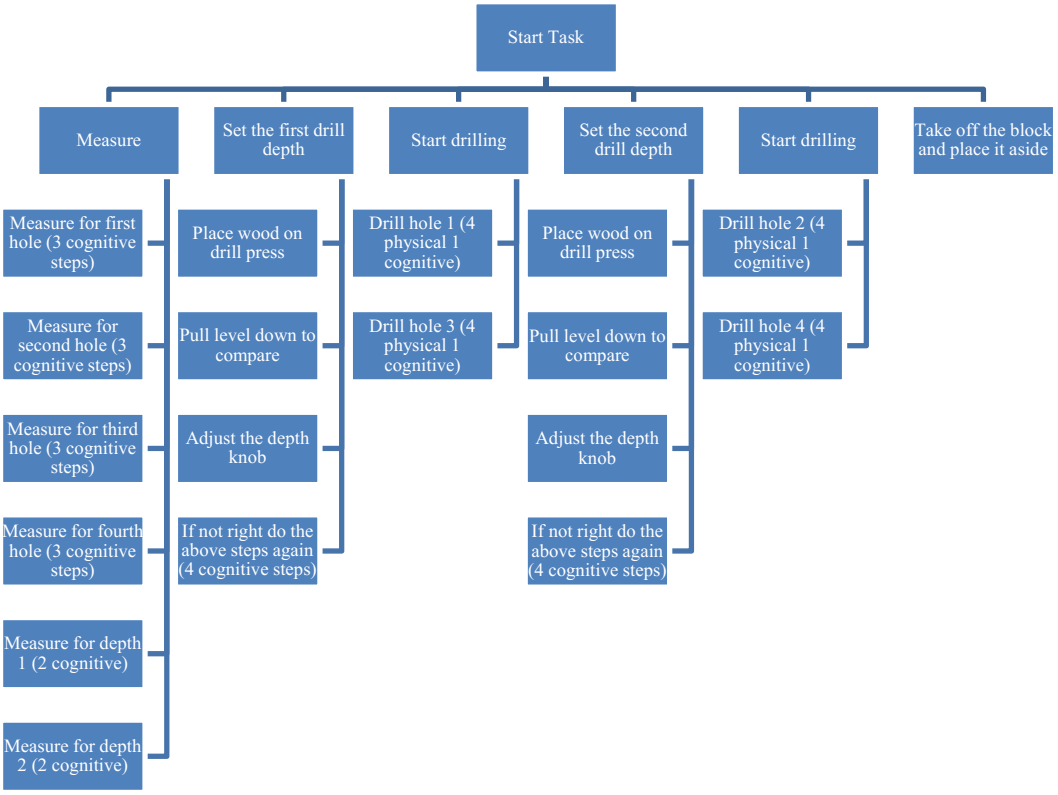


High cognitive and high physical complexity)

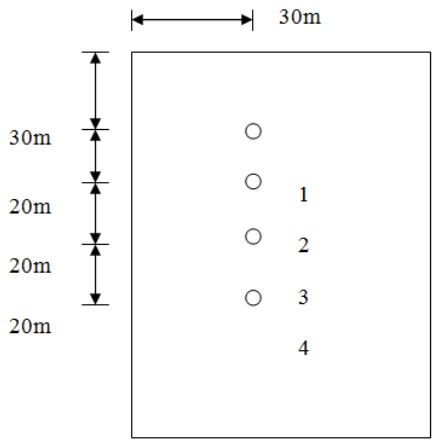
Drill Press Tasks and Hierarchy

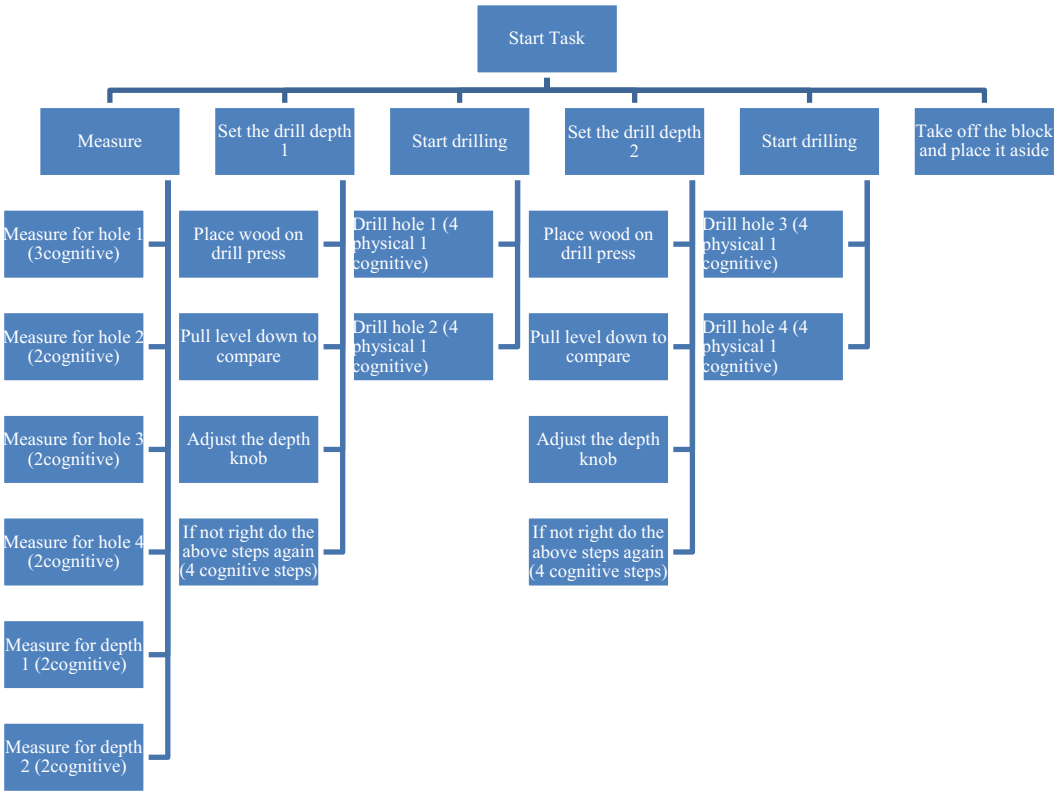
1. Use the drill press to drill four holes in the work piece in the places noted in the figure below. Hole 1 and hole 3 have a depth of 5 mm. Hole 2 and hole 4 have a depth of 8 mm (A: 19 physical 32 cognitive)



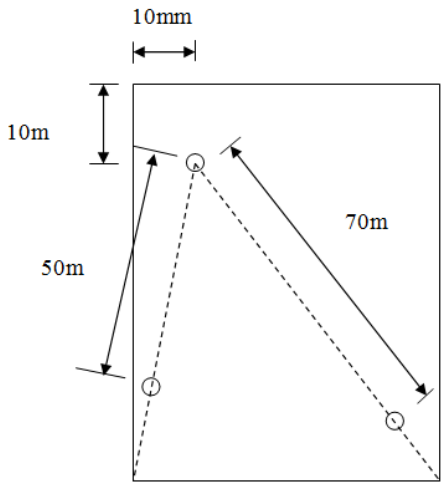


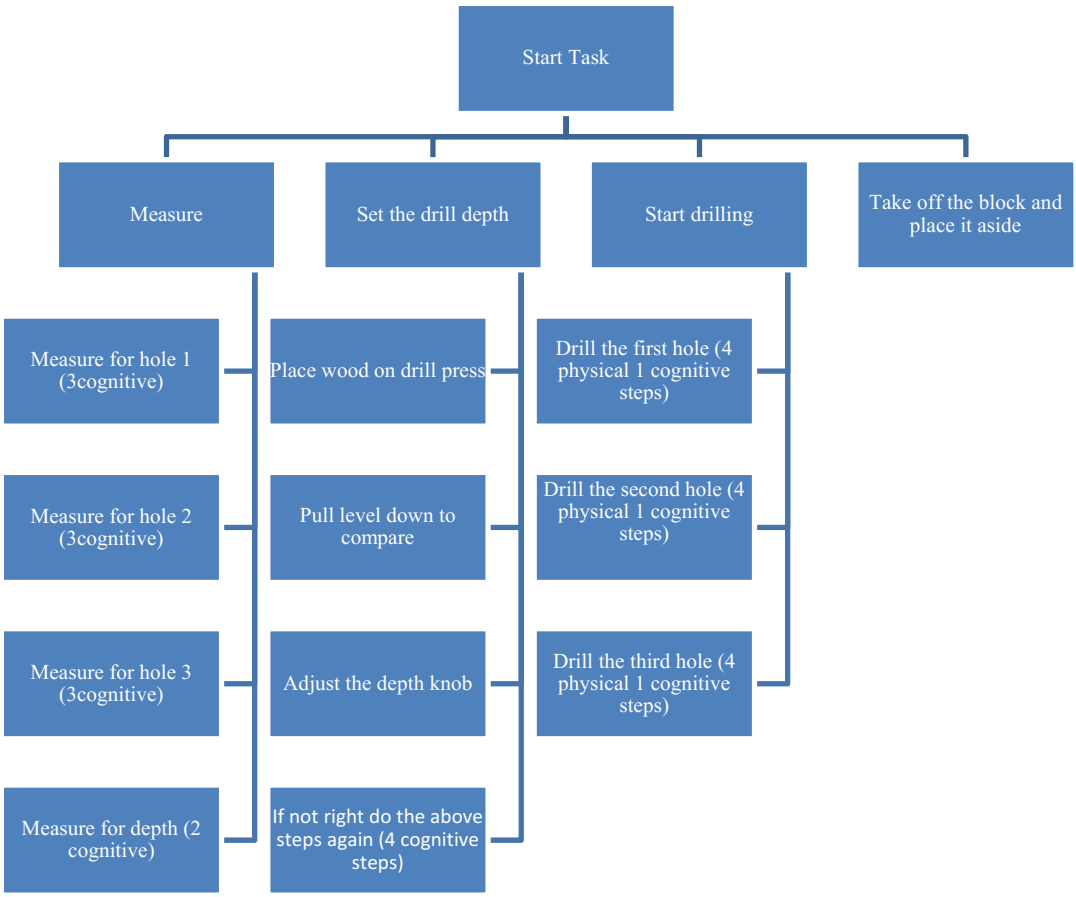
2. Use the drill press to drill four holes in a straight line in the work piece at the places noted in the figure below. Hole 1 and hole 2 have a depth of 9 mm. Hole 3 and hole 4 has a depth of 4 mm. (B: 19 physical 29 cognitive)





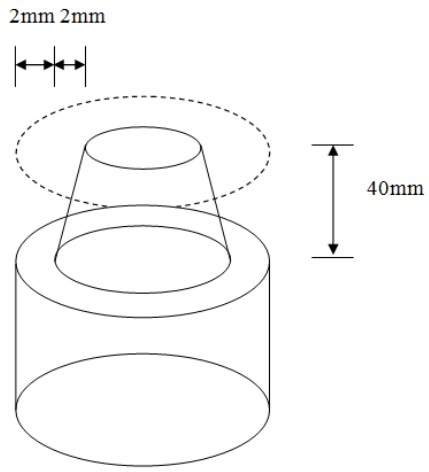
3. Use the drill press to drill three 6mm depth hole in the work piece at the place noted in the figure 3. (C: 14 physical 20 cognitive)

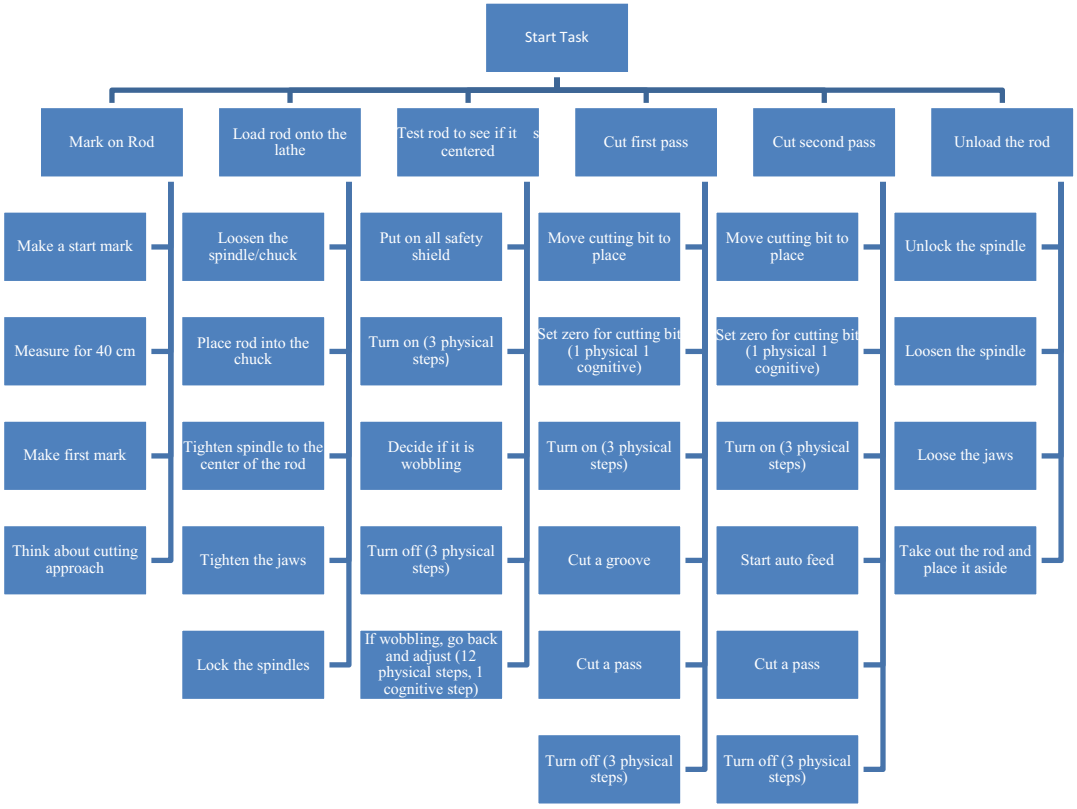




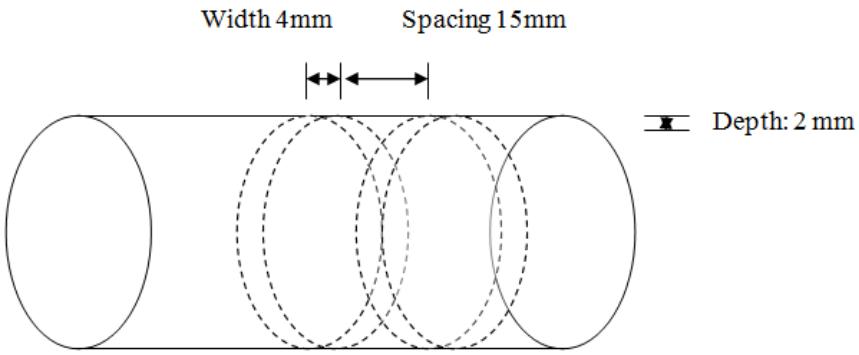
Mini-lathe Tasks and Hierarchy

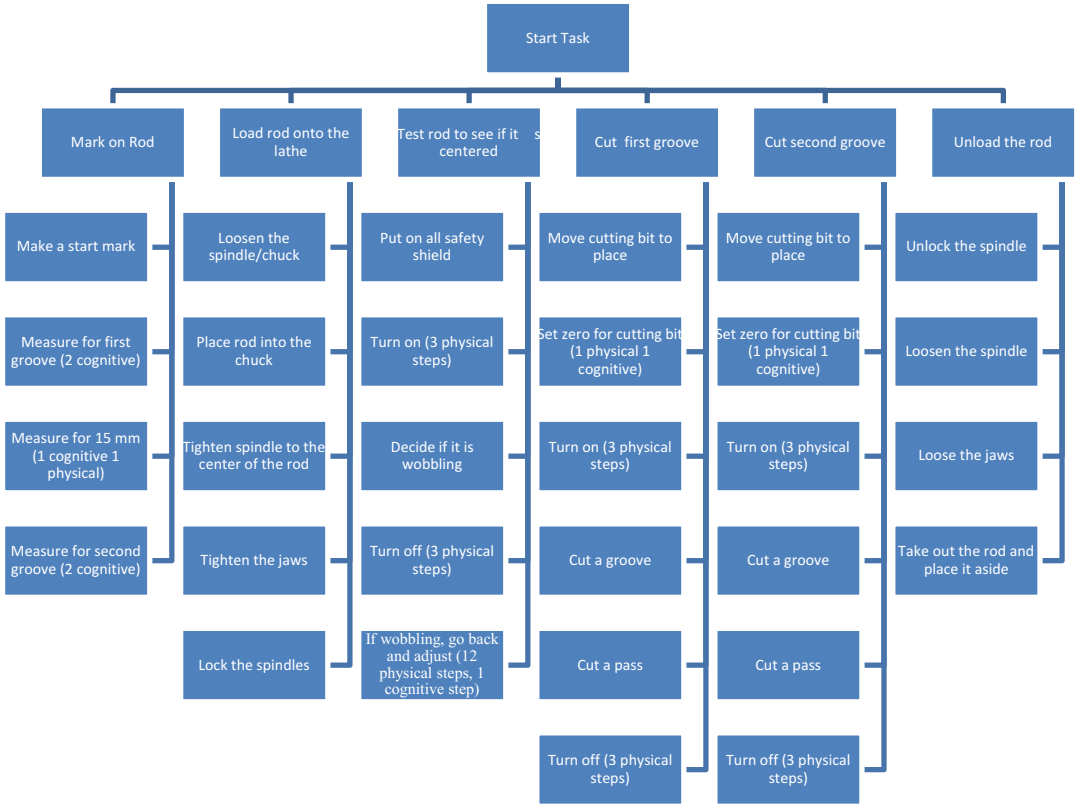
1. Use the lathe machine to carve the wood rod into the shape shown in the figure below. (F: 45 physical 11 cognitive)



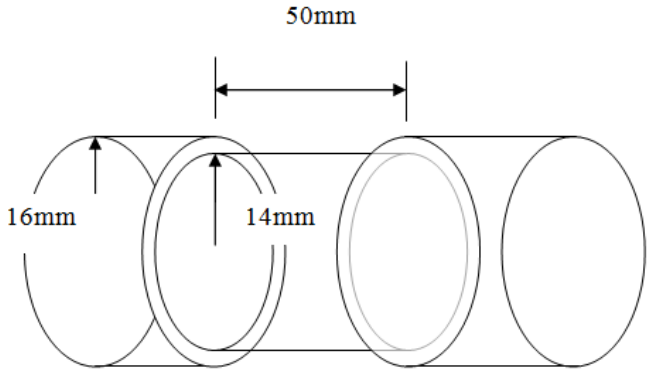


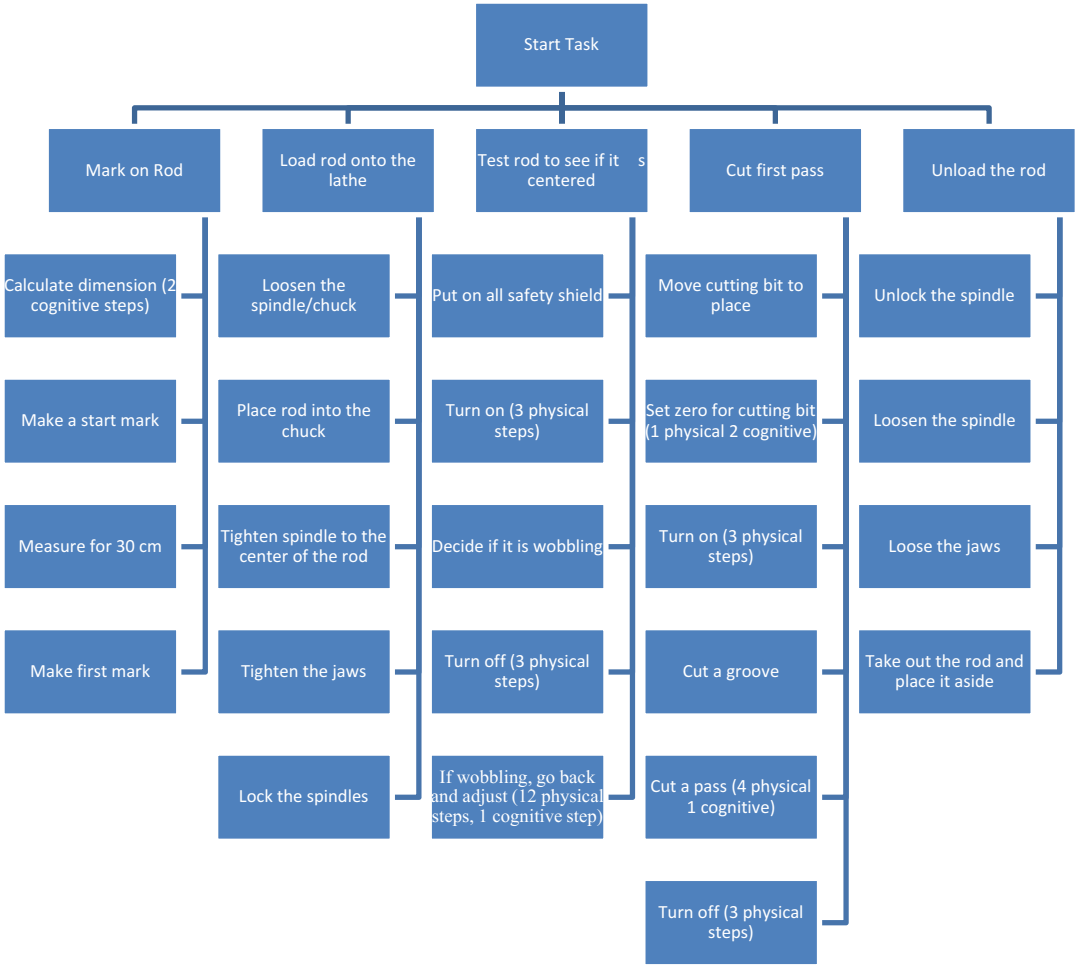
2. Use the lathe machine to carve two grooves in the wood rod. Please follow the dimensions shown in the figure below. (D: 47 physical 12 cognitive)





3. Use the lathe machine to carve out a rod section that has a length of 50 mm and a radius of 14mm, as shown in the figure below (assume current radius of the raw wood material is 16mm). (E: 41 physical 11 cognitive)





APPENDIX M

EXPERIMENT PROTOCOL FOR STUDY III

Safety reminder in scheduling email:

Dress code for experiment: to ensure your safety during the experiment, please avoid wearing loose clothing when coming to the experiment. If you have long hair, you will need to have it pulled back during the experiment. Also, you will not be allowed to wear any watches or jewelry during the experiment. If you wear them to the study, we will provide you a place to store them while you participate.

Before participants come

- Determine which task combination the participant is on.
- Prepare related document (2 SUS, 1 STQ, 1 think aloud note sheet, 2 consent forms, task lists and figures)
- Prepare raw material (3 wood blocks and 3 wood rods). Have the finished product sample ready.
- Reset machines (vacuum previous used machine, clean the desk, reset machine settings to default setting)
- Check camera (both cameras and computer, make sure the schedule on computer control software is correct and on)

When participants come

- Take him/her to the table to measure the required table height (elbow height should be between 2-8 inches above table height)
- Talk about the study, consent form. Give participant some time to read and sign the form. Then change the table height if necessary while participants do the consent form.

- After consent, reiterate the safety precautions. Check with participants, put watches and jewelry into a box aside.

Start Experiment

- Training on drill press/lathe
- Experiment on drill press/lathe
- 5 minutes to fill out SUS
- Flip-flop training
- Flip experiment
- 10 minutes to fill out SUS and STQ

End Experiment

- Debrief and payment/receipt
- Check data file (all questionnaires, note sheet, save and backup video file)
- Number the finished material and store them appropriately
- Clean up workstation and reset all machines

Training Script for Drill Press

Here is the drill press produced by Shopmate. It is used to drill holes in wood work pieces. It is composed of the press lever, the drill bit, and some control keys. The control keys include a knob to change the drilling speed and an on/off switch (in this experiment, you don't need to change the drilling speed, you can keep the same drilling speed when doing the tasks).

To do a drilling task, you can put a work piece on the platform, start the drill by turning on the switch, and press down the drill by pressing on the control lever. The drill bit will start to drill a hole into the work piece. When you feel you've drilled a certain

depth of hole, you can release the lever and finish drilling. Turn the machine to off, and brush the dust off of the work piece. A hole has been drilled.

There is one feature on this drill press that can control the depth of a hole you are drilling. This knob on the side is the control mechanism. It limits how deep you are going to drill. To set a certain drill depth, adjust this knob to the required position and stabilize the knob. Then when you drill, the press lever will stop at the depth that you set (show participants how it works)

Safety issues. Please never touch the drill bit with your hand. When doing a drilling task, try to hold the work piece firmly. When holding the work piece, do not leave your hand too close to the drill bit. Make sure you ware safety glasses when doing tasks.

Let participants test drill several holes. Pass them if they can do that comfortably. Otherwise correct them on their mistakes until they are comfortably doing that.

Training Script for Lathe Machine

Here is the 7" x 10" Precision Mini-lathe machine. It is used to change the diameter of a wood rod. We can also carve different shapes using the lathe machine. The lathe machine is composed of the spindle (holds the wood rod and makes it spin), controlling panel (start/stop, spinning speed, rotating direction), carving bit, and other controlling mechanisms (auto-feeding on/off, auto-feeding direction, carving depth control, horizontal distance control). (Show the participant each part)

We can perform several operations using the lathe machine, including carving a groove in the wood rod and carving the wood rod into different shapes (do some simple task and show them to the participants.)

You can also use the auto-feeding feature to complete the tasks (show participant how to do that)

Safety issues. Please make sure the rotation speed is at “0” when you turn on the machine. Always turn off the machine before adjusting the settings of the machine or loading/unloading work pieces. If you want to mark something on the work piece, do it before you load it on to the machine. Never touch the carving bit with your hand. Remember to put on safety glasses when doing the tasks. Keep safety shield in correct places (after loading the work piece and before start the machine). Never put your hand into the working area when the machine is on.

Let participants test run the lathe machine. Pass them if they can do that comfortably. Otherwise correct them on their mistakes until they are comfortably doing that.