Improving the Reliability and Validity of Wizard-of-Oz Methods

Xiangdong Li

A thesis submitted to the University of Huddersfield

in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

School of Computing and Engineering

University of Huddersfield

April 2012

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and he has given The University of Huddersfield the right to the use such Copyright for any administrative, promotional, educational and/or teaching purposes.

II. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

III. The ownership of any parents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such intellectual property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproduction.

Publications

- Andol X. Li and John V H Bonner, Designing smart domestic applications using wizard of oz methodology, In 2nd International Workshop on Human-centric interfaces for Ambient Intelligence, Nottingham, UK, 2011

- Andol X. Li and John V H Bonner, Improving control panel consistency of wizard of oz design and evaluation studies, In 17th International Conference on Automation and Computing (ICAC11), Huddersfield, UK, 2011

- Andol X. Li and John V H Bonner, Enhancing social relationships through human-like intelligences, The 9th International Workshop on Social Intelligence Design, Egham, UK, 2010

- Andol X. Li and John V H Bonner, Improving prototype consistence for wizard-of-oz simulations and evaluations, In the Proceedings of the Computing and Engineering Annual Researchers' Conference, Huddersfield, UK, 2010

- Andol X. Li and John V H Bonner, Designing interfaces to visualise domestic communication patterns, In the Proceedings of the Computing and Engineering Annual Researchers' Conference, Huddersfield, UK, 2010

- Andol X. Li and John V H Bonner, Smart control panel: developing conventional domestic infrastructures into ambient media, In the 2nd International Workshop on Semantic Ambient Media Experience (NAMU series), Salzburg, Austria, 2009

- John V H Bonner, Andol X. Li and Joanne Robinson, Designing and evaluating smart domestic technologies which use infrequent interaction, In the 8th International Conference on Pervasive Computing, Nara, Japan, 2008

Abstract

Wizard-of-Oz (WoZ) is a flexible, efficient and cost-economic method to the design and evaluation of interaction systems, particularly such of natural dialogue and smart systems. However, the literature review in the beginning of this research indicated that researchers struggled to implement WoZ and be able to gain reliable and valid experimental results in terms of system facilitation consistency; and WoZ has been criticised for a lack of systematic assessment of influence variables, especially when it was used to study new emerging information and communication technologies. Hence, this research aimed to understand and improve the reliability and validity of WoZ.

The research consisted of a series of empirical studies to incrementally deepen the understanding of influence variables. The main body of research comprised studies investigating (1) the impact of schema as WoZ study guidelines, (2) the impact of control panel in system facilitation, (3) the variables affecting evaluator's interpretation of schema, control panel and subject activity, and (4) the differences in multiple evaluators' system facilitation.

The results indicated that neither rigorous nor general schemas supported highly reliable system facilitation; rather, schemas should be accordingly proposed on the base of predictable or unpredictable user interactions. Also the results revealed the hidden relationships between control panel and system facilitation through identifying the control panel influence factors such like layouts and functions and their connections with system facilitation. Additionally, despite the difficulty of synchronising evaluators' individual expertise and experiences was admitted, the research findings suggested practical measurements to address the individual differences at acceptable levels through applying additional assistance and constraints to evaluator's system facilitation judgement and execution. And the results also provided secondary understanding towards smart system design for domestic communication and the development of WoZ system.

To my family, for their great supports

Contents

Copyright Statement ..... **i**

Publications ..... **ii**

Abstract ..... **iii**

Chapter 1 Introduction **1**

1.1 Research Motivation 2

1.2 Research Agenda 4

1.2.1 Research Objectives 4

1.2.2 Research Procedure 5

1.2.3 Contribution to HCI Knowledge 5

1.3 Thesis Structure 6

Chapter 2 Literature Review **7**

2.1 Reliability and Validity of HCI Methodologies 7

2.1.1 Scope of Research 7

2.1.2 Reliability and Validity of Design Methodologies 8

2.1.3 Reliability and Validity of Evaluation Methodologies 10

2.1.4 Lessons learnt from the Review of Methodologies 12

2.2 Reliability and Validity Requirements for Technology Studies 14

2.2.1 Requirements for Domestic Communication Technology Studies 14
2.2.2 Requirements for Novel Communication Technology Studies 17
2.2.3 Lessons Learnt from the Review of Interactive Technologies 19
2.3 Wizard-of-Oz, Why and How 22
2.3.1 Why Wizard-of-Oz 22
2.3.2 How WoZ Was Used in Previous Studies 24
2.3.3 Research Questions about the Reliability and Validity of WoZ 27
2.3.4 Lessons Learnt from the Review of WoZ Studies 29
2.4 Summary of Literature Review 31
Chapter 3 Specifications of WoZ Platform for Proposed Studies 34
3.1 Introduction of WoZ Platform 34
3.2 Refinement of WoZ Study Objectives 35
3.3 Specifications of WoZ Platform Construction 38
3.3.1 Requirements for Intelligent System Development 38
3.3.2 Requirements for Conversational System Development 39
3.3.3 A Trial Configuration of WoZ Platform 40
3.4 Specifications of Proposed WoZ Study 42
3.4.1 Conditions of Effective WoZ Study 42
3.4.2 Subjects (Participants) 43
3.4.3 Evaluation Criteria for Evaluator Operation 44
3.4.4 Ethical Issues 45
3.5 Summary 47
Chapter 4 Schema and the Impact in WoZ Study 48

4.1 Introduction of Study Objectives and Scope 48
4.1.1 Study Objectives 49
4.1.2. Study Scope 49
4.2 Study 1 - Method of Proposed Study 51
4.2.1 System Development 51
4.2.2 Study Structure and Evaluation Variables 53
4.2.3 Study Procedure 55
4.3 Study Results 57
4.4 Analysis: Evaluator's Operation and Impact 58
4.5 Analysis: Subject Activity and Impact 63
4.6 Analysis: Control Panel and Impact 66
4.7 Analysis: Study Strategy and Impact 69
4.8 Discussion of Study Findings 71
4.9 Conclusion 76
Chapter 5 Control Panel, Subject Activity Interpretation and the Impact in WoZ Study 77
5.1 Introduction of Study Objectives and Chapter Structure 77
5.1.1 Study Objectives 77
5.1.2 Chapter Structure 78
5.2 Approach for WoZ System Development 79
5.3 Study 2 – Consistency of Control Panel Operation 81
5.3.1 System Development and Design of Proposed Studies 81
5.3.1.1 System Development 82
5.3.1.2 Study Variables and System Components 85

5.3.1.3 Study Procedure 87
5.3.2 Study Results 87
5.3.3 Analysis: Difference in Evaluator's Operation 88
5.3.4 Analysis: Difference in Subjects' Reaction 95
5.3.5 Discussion of Study Findings 96
5.4 Study 3 – Consistency in Subject Activity Interpretation 99
5.4.1 Review of System Development 99
5.4.2 Method of Proposed Studies 100
5.4.2.1 System Development 101
5.4.2.2 Study Structure and Evaluation Variables 101
5.4.2.3 Study Procedure 103
5.4.3 Study Results 104
5.4.4 Analysis: Evaluator's Operation Judgement and Impact 105
5.4.5 Analysis: Subject Activity Interpretation and Impact 111
5.4.6 Discussion of Study Findings 112
5.5 Contribution 114
Chapter 6 Multiple Evaluators and the Impact in WoZ Study 115
6.1 Introduction of Objectives and Chapter Structure 115
6.1.1 Study Objectives 116
6.1.2 Chapter Structure 118
6.2 Study 4 – Identifying Variables in Multiple Evaluators WoZ Studies 119
6.2.1 Method of Proposed Studies 120
6.2.1.1 System Development 120

6.2.1.2 Study Structure and Evaluation Variables 122
6.2.1.3 Study Procedure 125
6.2.2 Study Results 126
6.2.3 Analysis: Evaluators' System Operation Differences and Impact 127
6.2.4 Analysis: Evaluators' Personal Differences and Impact 131
6.2.5 Discussion of Study Findings 134
6.3 Study 5 – Improving Variables in Multiple Evaluators WoZ Studies 136
6.3.1 Method of Proposed Studies 136
6.3.1.1 System Development 136
6.3.1.2 Study Structure and Evaluation Variables 137
6.3.1.3 Study Procedure 138
6.3.2 Study Results 138
6.3.3 Analysis: Differences in Schema and Subject Interpretation and Impact 139
6.3.4 Analysis: Differences in Control Panel Operation and Impact 141
6.3.5 Discussion of Study Findings 142
6.4 Study 6 – Controlling Multiple Evaluators' System Facilitation 144
6.4.1 Method of Proposed Studies 145
6.4.1.1 System Development 145
6.4.1.2 Study Structure and Evaluation Variables 145
6.4.1.3 Study Procedure 146
6.4.2 Study Results 146
6.4.3 Analysis: Differences in Evaluators' Judgments and Impact 146
6.4.4 Discussion of Study Findings 147

6.5 Comparative Review of Study 4, 5 and 6 149

6.6 Contribution 151

6.7 Conclusion 152

Chapter 7 Research Findings and Contribution 153

7.1 Introduction 153

7.2 Research Questions and Findings 154

7.2.1 Questions from Literature Review 154

7.3 Revisiting the Research Framework 156

7.4 Contribution to Knowledge 158

7.4.1 Improving the Reliability and Validity of WoZ Studies 158

7.4.2 Improving Practical Design and Evaluation 160

7.5 Critical Review of Thesis 162

7.5.1 Fulfilling Research Objectives 162

7.5.2 Research Weakness and Strength 163

7.5.3 Challenges 163

7.6 Future Work 165

Acknowledgements 167

References 168

Appendix 1.1 181

Appendix 3.1 183

Appendix 4.1 188

Appendix 4.2 189

Appendix 4.3 ..... **193**

Appendix 5.1 ..... **195**

Appendix 5.2 ..... **196**

Appendix 5.3 ..... **197**

Appendix 5.4 ..... **199**

Appendix 5.5 ..... **200**

Appendix 6.1 ..... **202**

Appendix 6.2 ..... **204**

Appendix 6.3 ..... **206**

Appendix 6.4 ..... **208**

List of Figures

Figure 2.1: WoZ study key components 24

Figure 3.1: Virtual control panel of domestic central heating 41

Figure 3.2: Control panel interfaces following the colour cube 41

Figure 4.1: White and black patterns on cube surfaces 51

Figure 4.2: Front-end application interface 52

Figure 4.3: Control panel 52

Figure 4.4: WoZ operation system design 53

Figure 4.5: Study structure and components of operation system 53

Figure 5.1: The cube with numbers on its surfaces 82

Figure 5.2: The projector (left) rendered interfaces on the coffee table (right) 83

Figure 5.3: Compact control panel design 83

Figure 5.4: Second control panel design 84

Figure 5.5: The dialogue (left) and multimedia control (right) control panel design with preset messages ... 85

Figure 5.6: Study variables 86

Figure 5.7: Control panels for the system applications 100

Figure 5.8: Study variables 102

Figure 6.1: The control panel for the new function 121

Figure 6.2: Displaying pictures on the coffee table 121

Figure 6.3: Control panel design 122

Figure 6.4: Study variables 123

Figure 6.5: Control panel layouts. Left – new evaluator's, right – experienced evaluator's 128

List of Tables

Table 4.1: Summary comparison of observations in four aspects 61

Table 5.1: Summary of study 2's key components 81

Table 5.2: Summary of study 3's key components 101

Table 6.1: Summary of study 4's key components 120

Table 6.2: Summary of study 5's key components 136

Table 6.3: Summary of study 6's key components 144

Chapter 1

Introduction

Wizard-of-Oz (WoZ) is an integrated design-evaluation method which is flexible, efficient and cost-economic for HCI study. It provides enough flexibility to the design and evaluation of interactive systems at a variety of fidelity levels; it also supports iterative studies through reconfiguring evaluator's management and control of system components; and it saves repetitive development work through human evaluator's simulation. For this reason, WoZ was popularly used in natural language dialogue system and intelligent interface studies, and was useful to draw insights into future interactions beyond current technology levels. Like other evaluator-participatory design and evaluation methods in this field, however, the reliability and validity of WoZ study were potentially variable due to the inconsistency of evaluator's system management and control. Evaluator's participation provided good flexibility of system design and evaluation, but it also introduced risks of inconsistent system simulation (or system facilitation) in the study.

Previous studies have reported preliminary measurements to address these risks, such as additional evaluator training and careful control panel design (Xu et al. 2009). These studies noted that WoZ was such a cognitive-heavy method that researchers struggled to maintain consistent system facilitation. There was a widely spread concern that the evaluator's facilitation, which was affected by study variables, largely determined the reliability and validity of WoZ. Furthermore current trends, such like human-like conversational interaction (Edlund et al. 2008), multimodal interfaces (O'Halloran et al. 2010) and intelligent systems (Hawes et al. 2009), are exaggerating these risks.

Therefore improving the reliability and validity of WoZ becomes a compelling challenge. Given WoZ's potentials of probing novel interactive technologies and future domestic communication, researchers will gain substantial benefits from the understanding and improvement of reliability and validity. Fraser and Gilbert

(1991) had reviewed a small number of potential influence variables in WoZ-based speech simulation systems, such as task variables and subject variables. Although new design and evaluation studies are being continuously proposed, few have systematically assessed the variables and their impact (Dow et al. 2005). Hence, in this research, much effort was aimed at this direction, and the high level aim of this research was to contribute to HCI knowledge regarding the reliability and validity of WoZ method. In particular the research aimed to address the inconsistency of evaluator's system facilitation in WoZ studies, in order to support the design and evaluation of new emerging information and communication technologies within domestic settings.

1.1 Research Motivation

The communication between humans and computers has been shown to be significantly enhanced when smart devices such as natural-language dialogue systems and other intelligent interfaces are used (Aarts 2004). Domestic digital communications are particularly undertaking profound changes such as domestic telecommunication (Anderson et al. 1999). Approaches to address such changes, however, had insufficient considerations on the unique qualities of the domestic communication (Dahlback et al. 1993). Insightful and empirical studies are therefore required for the understanding of this; and one way of achieving this is to use WoZ, an integrated design-evaluation method that has substantial advantages over other methods in terms of flexibility, efficiency and cost.

Given evaluator's system facilitation WoZ is capable to build up rapid prototypes beyond current technologies. For example, WoZ was used to mimic a listening typewriter when the state-of-the-art speech recognition technology was still immature (Gould et al. 1983). In pioneering interaction system studies WoZ is one of important methods to gain insightful understanding of future systems such like spoken dialogue tutorial system (Forbes-Riley and Litman 2011) and intelligent home appliance control system (Hsu et al. 2010).

WoZ also carries risks with its advantages due to the nature of evaluator's system facilitation, which is occasionally improvisational and inconsistent. Although the system facilitation in previous studies produced

convincing systems, the challenges of this research are twofold. In terms of technology faithfulness, the evaluator's operation is required to be rigid to have computer-like performances. In terms of facilitation consistency, the evaluator is required to generate predictable responses to user interactions. Several general factors were described as influence variables, including scope, task, subject, communication channel and the most important 'wizard' (evaluator) variables (Fraser and Gilbert 1991); and some specific variables were also noted, such as evaluator training qualities and subject's awareness of evaluator (Salber and Coutaz 1993). Given the dynamics in system facilitation, a handful studies cannot constitute a body of systematic understanding to improve the reliability and validity of WoZ; and a good number of empirical studies are needed to assess and address relevant influence variables systematically.

1.2 Research Agenda

Previous studies had clear emphasises on cutting edge areas such as ambient intelligence, but these had less focuses on the design and evaluation methodologies. After surveying this field, this research developed a particular interest in the improvement of the reliability and validity of WoZ due to WoZ's potentials to probing future domestic communication.

1.2.1 Research Objectives

To understand and improve the reliability and validity of WoZ a series smart devices were developed in the laboratory in order to support empirical studies. These devices were inspired by surveys of domestic communication routines, and were continuously improved throughout the research with increasing understanding of system facilitation variables.

The primary aim of the smart device development was to provide interaction channels for domestic communication. These applications were designed with human-like intelligence, and integrated with advanced natural-language dialogue interfaces. Meanwhile some novel interactive technologies such as mixed realities were also integrated.

The overall objectives of this research were provided:

- *To gain the understanding of the schema and its impact in WoZ study.* In order to lay a foundation for WoZ studies, the research firstly planned to investigate the schema variable and its impact on the reliability and validity of WoZ; and a secondary aim of this was to reveal guidelines for schema design.

- *To examine the variables relevant with key system components, including user activity interpretation and control panel use*. This aimed to examine the variables that were relevant with system components, and to explore how these variables generated influence and what measurements should be used.

- *To investigate the variables related to multiple evaluators and thus devising methodological recommendations for WoZ study configuration.* The final objective aimed to provide methodological understanding to support future WoZ study with multiple evaluators.

1.2.2 Research Procedure

This research was initially based on the understanding of literature review, where there were a set of variables identified as major research questions. With increasing understanding of the reliability and validity of WoZ, the research kept refining its objectives, system development and study methods. At first, the research was proposed on the fundamental questions from the literature review; it investigated the schema as the fundamental variable. After gaining practical understanding of schema throughout evaluation, the research stepped further to investigate other variables that were relevant with system components such like control panel. And lastly, the study led to the evaluation of variables that concerned multiple evaluators and their individual influences.

1.2.3 Contribution to HCI Knowledge

The research made contributions to HCI knowledge in two main perspectives. Firstly, the contribution included the understanding that neither rigorous nor general schemas supported highly reliable system facilitation; rather, schemas should be accordingly proposed on the base of predictable or unpredictable user interactions. Also the contribution included new understanding of hidden relationships between control panel and system facilitation through identifying the control panel influence factors such like layouts and functions and their connection with system facilitation. Additionally, despite the difficulty of synchronising evaluators' expertise and experiences was admitted in studies, the research contribution comprised suggestions of practical measurements to address the individual differences at acceptable levels through applying additional assistance and constraints to evaluator's system facilitation judgement and execution.

Secondly, the research contributed new understanding of how to build a reliable WoZ system, thus to support high-level intelligent interfaces such as natural-language dialogue system design. Such contribution

included, for example, the understanding of the connections between evaluator, subjects, data collection and analysis methods. In addition to that, contributions were also made in terms of future intelligent system design for domestic communication study. For example, the system required much clearer indication of system capability and progress than traditional computer applications.

1.3 Thesis Structure

The thesis is structured as follows. Firstly, **Chapter 2** presents a literature review of the state-of-the-art of current design and evaluation methodologies, and reviews the reliability and validity of WoZ in previous studies. Following that, **Chapter 3** provides more specific understanding of implications from literature review, and draws explicit objectives based on that. Some technical difficulties, preconditions and ethical issues are described in this chapter. **Chapter 4** describes details of a group of WoZ studies that focused on schema design variables and impact. The study in **Chapter 5** was based on the understanding schema design, and it presented another two groups of WoZ studies. The focus of the first group was on control panel design, and the focus of the other group study was on evaluator's interpretations and prediction of subject activities. **Chapter 6** continues to investigate WoZ variables related to multiple evaluators. Three incremental studies are presented in this chapter. And finally **Chapter 7** presents a discussion of all previous findings, thus identifying the guidance for improving the reliability and validity of WoZ. In this chapter, the overall contribution and critical review of the research are described, along with the future work at last.

Chapter 2

Literature Review

2.1 Reliability and Validity of HCI Methodologies

This chapter presents a review of the backgrounds and relevant research on Wizard-of-Oz (WoZ). Firstly, **Section 2.1** provides an overview of Human-Computer-Interaction (HCI), the research field to which the thesis makes a contribution. A number of HCI design and evaluation methodologies are reviewed in this Section. Secondly, **Section 2.2** describes a series of technical requirements for domestic communication study. In this Section a combined selection of traditional and novel interactive technologies are reviewed. Thirdly **Section 2.3** describes why and how WoZ is suitable for study of future domestic communication. In addition, previous WoZ studies are reviewed to understand the reliability and validity of WoZ. And finally this chapter summarises the understanding in **Section 2.4** in which several research questions are described.

2.1.1 Scope of Research

This research fell in the field of HCI which was concerned with a series of computing system-based design and evaluation activities for all fashions of technologies and related systems and products (Winograd 1997). These activities were claimed relevant with '*the disciplinary concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them*' (Hewett et al. 1996). Two main types of outputs made by these activities were aimed to make contribution to the body of HCI knowledge. One output takes accounts of empirical results such as interactive system design and technologies, and the other output focuses on methodological contribution to guide the design and evaluation of interaction tools and methods.

HCI research is concerned with empirical activities, for example designing and evaluating computer interfaces. The conceptualisation of task-artefact cycle highlighted the significance of practical development in iterative progresses (Carroll et al. 1991). However, that traditional design and evaluation methodologies are short of supporting novel interactive technology and system development study. In this case, more endeavours are required to update methodological understanding. Studies have been carried out to understand both weaknesses and strengths of HCI design and evaluation methodologies. For example, previous studies have identified weaknesses such as being time-consuming, difficult to generalise findings, having unknown biases and low representative data collection (Kjeldskov and Graham 2003). These weaknesses were grouped in several categories, such as reliability, validity, effectiveness and usability. In particular, reliability and validity were the most highlighted concerns amongst these methodologies due to its significant impact on study results.

2.1.2 Reliability and Validity of Design Methodologies

A typical selection of design and evaluation methods was reviewed due to their popularity of use. The intention was to reveal how the reliability and validity of current design methodologies were and how they were addressed. The review divided design methods in three categories according to the taxonomy of Wania et al. (2006), as this taxonomy highlighted the focus of these methods and participatory and evolutionary procedures of method.

Product-Oriented Design Methodologies

These methodologies focused on system theory and software engineering (Helander et al. 1997), and included sketching and paper-prototyping for software interface design (Bailey et al. 2008; Snyder 2003). Sketching was useful in design-oriented disciplines such as visual communication. Advantages of such methodology included creativity and fast problem-solving (Craft and Cairns 2009). As a form of presenting interface ideas, sketching visually rendered alternative ideas of communication system interface design. In this regard sketching supported rapid prototype development and achieved specific goals via cross-discipline thinking (Schutze et al. 2003). Methodologies like this were helpful in identifying preliminary communication

questions, and assessed conceptual prototypes in the early stage of domestic communication system design (Sengers et al. 2005).

Process-Oriented Design Methodologies

The philosophy behind these methodologies was to involve the users of future system. Such methodologies, to name a few, included card sorting, culture probes, and technology probes. Studies adopted card sorting to generate information about the association of specific data items (Chaparro et al. 2008; Hudson 2007). Users were employed to participate in the organisation of unsorted items into groups. As an early-stage design method, it provided insights of interactive system architecture that was used by a practitioner. However, from domestic communication study's point view, its advantages were twofold. Firstly, the understanding was implicit in terms of card sorting's validity as a means of system architecture generator. It might be used without in-field user interactions, and thus leading to an architecture that was not useable for real tasks. Secondly, the reliability of card sorting was variable due to its results were influenced by evaluator's operations.

Culture probes used a set of tools, such as pads, cameras and postcards, to enable users to record experiences in real environments (Gaver et al. 1999). As a result, understanding in respect to culture and personal activities were captured, and sensitive routines for domestic communication design could also be uncovered. For this reason, culture probes can be used to investigate strong-tie relationships in non-intrusive way (Vetere et al. 2005), and to probe longitudinal influence of domestic communication technology emergence (Arnold 2004). However, careful experiment control was not well supported in culture probe study, neither was data collection under evaluator's control. The whole study was dependent on user's self-awareness. In this regard, technology probes was also claimed to have similar risks in lowering of reliability and validity (Hutchinson et al. 2003).

Use-Oriented Design Methodologies

These methodologies focused on actual use situation and aimed to assess the quality of design system, such as ethnography (Hughes et al. 1997). Ethnography gained some disputed prominence as an important

methodology of '*requirement elicitation*' (Hughes et al. 1994). It adopted highly developed prototypes for in-field study, and collected study results that reflected real system use. For example it captured practical use of interactive television in the home environment (Bernhaupt et al. 2008). The data of practical use supported comprehensive understanding of users perceptions and interactions with systems (Hughes et al. 2000).

However, ethnography methodology was incompatible with domestic communication study in terms of flexibility and data reliability. Ethnography was not a flexible method due to the fact that it was proposed with high fidelity prototypes for required for real-use. In addition, it was considered an intrusive method that disturbed domestic communication routines. Data capture devices needed to be installed in the home, and the data collection lasted for a long period. In this case the validity of data collection might be affected. For example, installing and calibrating a new system in the home for communication observations might affect the ancient ecology of domestic communication (Venkatesh 2001), as the invasion into private space was more sensitive than public places (Bernhaupt et al. 2008). Other risks of variable reliability and validity were that the data interpretation was done by researchers who were not in presence of study site, thus generating misunderstanding of study results and unaware biases (Blythe and Monk 2002).

2.1.3 Reliability and Validity of Evaluation Methodologies

This review of evaluation methodologies was carried out within the context of domestic communication. Two taxonomies were combined in the review, in order to group these methodologies into four concise categories, including predictive, experimental, empirical and informal evaluation methodologies (Balbo et al. 1993; Nielsen 1994). These categorises highlighted the way of evaluator's participation in evaluation study.

Predictive Evaluation Methodologies

Predictive evaluation methodologies were influenced by practitioner's expertise for system prediction and data interpretation. GOMS (goals, operators, methods and section rules) was such a method that characterised tasks and relevant knowledge of performances by using a whole family of models (Bonnie 1995; Card et al. 1980). Each model referred to different items as defined in (Card et al. 1980)'s study. For example, the 'goals' predicted what tasks users intended to achieve. GOMS provided both qualitative and

quantitative data relevant with parameters such as performance time, error frequency, learning time and working memory errors. However, validity and accuracy of the prediction had strong dependencies with the adding-up of model predictions. These predictions were affected by evaluator's individual experiences, although it was economic, flexible and efficient in system evaluating.

Experimental Evaluation Methodologies

Cognitive walkthrough (CW), as suggested in (Bonnie 2004), supplemented the weakness of reliability and validity of GOMS. CW was a usability evaluation method which inspected levels of ease of system use (Lewis and Wharton 1997). In terms of system learnability CW hardly needed to train experts to assess the system (Wharton et al. 1994). The limitation of CW for domestic communication study was that it required fully-developed communication prototypes. Consequently CW might not be very suitable for intelligent and multimedia communication study in the home (Huart et al. 2004).

Empirical Evaluation Methodologies

Heuristic evaluation (HE) was a non-specialist-required method (Nielsen and Molich 1990), which involved several reviewers, preferably experts though, to criticise prototypes. Since no formal training was needed, the range of reviewer is broad such as students and programmers. In addition it fitted well with various stages of system, which introduced fast turnaround time. On the other hand, (Nielsen 1994) noted the huge differences between reviewers with as low as 9 percentage of comments overlapped. To address the diversity of reviewers a large number of reviewers were required to guarantee the accuracy and reliability. In (Slavkovic and Cross 1999) the study mentioned that HE was not scalable for sophisticated system interfaces such as natural language-based (NL) dialogues. And in (Kantner and Rosenbaum 1997) the study added that HE also had risks of reviewers emulating users for actual user feedbacks which were not properly addressed.

Informal Evaluation Methodologies

Think aloud (TA) was an informal evaluation method which was widely used in usability evaluations. It asked users to work on specific tasks and verbalise task performances and thoughts (Ramey et al. 2006). However,

when TA was used in domestic communication study, users had concerns about the reactivity – the verbalising action affected users' task performances. In this regard a low level reliability occurred due to the dynamic nature of user's task performance.

2.1.4 Lessons learnt from the Review of Methodologies

The review confirmed observations that there was a strong need to update methodological understanding for studies of novel interactive technologies and systems. Especially the review highlighted how the reliability and validity varied from evaluator's point of view.

Firstly, the review indicated that design and evaluation methodologies incurred variable reliability and validity when evaluator-participatory operations were concerned. Variables that posed influence included evaluator's experiences and expertise, number of evaluators, and evaluator's manipulation of methods. Most of these variables were relevant with the evaluator who managed and controlled the study. The difference in evaluators' expertise was a widely spread consensus that required extra measurements to cope with. For example, employing a large number of evaluators could noticeably improve reliability and accuracy of heuristic evaluation. This might be also applicable to other methods such as GOMS. Nevertheless there remained another difference in evaluator's overall operations throughout the study. There was no guarantee that each evaluator could generate consistent operations throughout design and evaluation progresses. In this regard this type of difference was actually more critical than the differences that caused by other variables, since this appeared negligible to researchers when the evaluator was a specialist.

Secondly, the review showed that the way of how evaluator participated in design and evaluation study had different influence on reliability and validity. Evaluator had two main roles of participation. One was in real-time system operation such as interview. This type of participation required rapid response to system interactions, and evaluator had no chance to alter the response after the operation. The other type was prediction and interpretation of study results. For example, in GOMS evaluator generated outputs via model predictions, and in ethnography study recorded study footage was interpreted by evaluators. Prediction and

interpretation were both concerned with evaluator's different experiences and expertise as described previously (see **Section 2.1.3**), and thus generate variable reliability and validity.

And finally, the review noted that the extent to evaluator's manipulation of study also caused variable reliability and validity. In this research 'evaluator's manipulation' was interpreted as the progress of system interaction control and data collection. Methodologies with assistant tools, such as system models in GOMS, were shown to have better reliability and validity results. For example, cognitive walkthrough generated more reliable evaluation results when the study had a system with fully developed functions.

In summary design and evaluation methodologies were reviewed for the understanding of variable reliability and validity. The understanding in itself did not make solid contribution to HCI knowledge, but it pictured current state of reliability and validity in design and evaluation methodologies. In the following sections, these benchmark considerations were applied more specifically to reflect how reliability and validity was coped with in WoZ studies, and based on which to shape future research directions.

2.2 Reliability and Validity Requirements for Technology Studies

This section reviews the first body of research related to the adoption of domestic communication technologies and their requirements of reliability and validity of methodologies in study. The primary aim of this work was to understand what requirements needed to meet to address variable reliability and validity of methodologies with fast technology emergence, and how that evolved from previous studies.

2.2.1 Requirements for Domestic Communication Technology Studies

Over last decades a number of communication technologies have been introduced into the home, such as portable computers and wireless networks. To understand how design and evaluation methodologies were applied to address the technology emergence, a typical range of interactive technologies and relevant studies were highlighted as follows.

Initial Domestic Communication Technologies

Blythe and Monk (2002) reported some ethnographic findings related to early domestic technologies (see definition in **Appendix 1.1**) such as privatisation of personal space. According to the taxonomy of Vall (1988), these technologies had limited communication capacity, and were designed to reduce mundane chores in the home. The focus of these technologies was on the respect to functionality, and their '*complexity and compatibility*' were well understood (Rogers 2003). Although these 'labour-saving' technologies saved much housework (Blythe and Monk 2002), an unexpected consequence was that householders spent more time on housework such as laundry and vacuuming, regardless of that high effectiveness for housework was provided by these technologies. In this regard, householders' behaviours of domestic communication were not very much affected by these technologies which, however, provided possibility for householders to spend more time on communication.

Television and Domestic Communication

The rise of communication needs in the home accelerated the diffusion of 'time-consuming' technologies (Bowden and Offer 1994). Television was probably an ultimate time-consuming technology that dominated most free time in the home. In the appearance of colour television in 1970s, it attracted householders coming home earlier due to the colour television reduced the attraction of '*doing nothing particularly*' in workplace (BBC Statistics, 1978). The increasing ubiquity of television continuously changed householders' conventional magazine and book readership (Vitalari et al. 1985), on the base of which, it caused a noticeable 'time-shifting' (Irani et al. 2010) in the home. For example, to date the average time for family meal has lost much to television viewing (Bowden and Offer 1994), and the conventional sports and cinema time is also compressed. Both time and site of conventional domestic communication were affected due to television's temporary 'broadcast structure' (Irani et al. 2010). Following technologies such as VCR and DVR generated additional impact on domestic communication behaviours. For example, householders no longer needed to catch up timetables for specific TV programme. Some side effects are also received in terms of householders' communication behaviours such as TV addiction (Kubey and Csikszentmihalyi 1990) and social isolation (Bickham 2006).

Telephone and Domestic Communication

Another technology that posed important impact on domestic communication was the telephone, a basic tool to provided accessibility and assistance for social life organisation in decades ago (Noble 1987). By connecting two people over distance, telephone was a '*private and independent* (Lacohée and Anderson 2001)' communication tool to continue housewives' ancient gossip habits (Kline 2003). Telephone helped mediate strong-tie relationships such as families and friends geographically dispersed (Vetere et al. 2005). Meanwhile, Lacohée and Anderson (2001) noticed that new communication patterns were irritated by telephone, such as duty calls and '*pseudo maintenance calls*'. Furthermore, communication with external bodies such as bank and hospital was also affected, especially in terms of communication medium and effectiveness. For example, it gave efficient access to family doctors for medical consultations (Bunn et al. 2009). Rather than seeing the telephone as an efficient communication tool, Hjorthol and Gripsrud (2009) regarded it as an extension of social communication hub. Furthermore, Lo and Lie (2008) added that,

householders had different levels of trust in telephone communication, between long-distance and short-distance communication situations.

Broadband and Domestic Communication

Papacharissi and Zaks (2006) seemed broadband in the home as the next step of internet diffusion, although some barriers such as third party infrastructures were still concerned (Edwards and Grinter 2001). Indeed, Horrigan (2009)'s national survey indicated the increasing importance of broadband use in the home as a new communication technology. '*All flavours of high-speed digital voice, data and video services*' (Kirstein et al. 2001) to date have been provided through broadband. That caused another shift of communication site in the home – from televisions to internet-based devices. Horrigan (2009)'s report of home broadband stated some qualitative changes of everyday life patterns relevant with broadband technology progress. It described householders' growing degree of engagement in internet services such as e-shopping and online file sharing. It also described projections of real life's communication experiences on virtual community, such as Second Life. The appearance of new communication patterns were reported such as larger social networks (Humberto T. Marques et al. 2004).

A number of studies focused on broadband's social impact on domestic communication behaviour change. Kraut and Kiesler (2003)'s longitudinal study claimed that the use of broadband influenced its impact on communication behaviours over time. For example, using the search engine to find information had a different impact than using it for talking with families. Firth and Mellor (2005)'s analysis in broadband's benefits and problems indicated that, in terms of social relations, highly socially-isolated individuals might still be excluded from the communication online. A similar conclusion made by Kraut and Kiesler (2003) stated that greater use of internet hardly led to more positive outcomes related to social engagement. Whereas they also added, because the broadband was a social technology it would bring similar effects as conventional forms of communication such as less loneliness.

Mobile Phone and Domestic Communication

Also, studies mentioned the impact of mobile phone, as its use in the home gained special places. It was noticed that this technology changed perception of communication, for example parents perceived their phones as a means of staying connected with children (Palen and Hughes 2007). As well, how teenagers related to their parents using mobile phone was also affected (Ling and Yttri 2002). Studies noted that teenagers had few simultaneous multiple conversations via mobile texts, and they communicated with a few friends via mobile phones (Grinter and Eldridge 2003). Mobile phones not only loosened constraints of communication time and sites, but also helped to augment the coordination of domestic routines (Ling 2004). In this regard, mobile phone refined the way of relationship maintenance and development between families (Aoki and Downes 2003).

2.2.2 Requirements for Novel Communication Technology Studies

Ubiquitous Computing

Studies envisioned pioneering concepts of future domestic communication technologies. Weiser (1991)'s ubiquitous computing was one of these. Similar concepts such as ambient intelligence, pervasive computing and the internet of things were often interchangeably used, referring to the ultimate human-computer interactions that were invisible but everywhere. Ubiquitous computing suggested numerous small, wirelessly interconnected sensors, which were embedded invisibly in everyday objects. In terms of technical basis of ubiquitous computing, the communication technology was particularly emphasised as the key technology (Friedewald and Raabe 2011). With which, the system could sense the environment where devices were embedded with data exchange with other terminal devices and between devices and users (Lagasse and Moermann 2005).

On the one hand, ideas such as '*intimate assistant*' and '*human assistant*' were around in the vision of future domestic communication (ACM 1993). This included a revolutionary change of communication style from 'physical actions – reactions' to 'signal and signs' (Bruns 2006). In the new style, the signs regarded as real world perception and changes had more connections with real world's objectives than cyber objects. In order

to interact with invisible embedded systems, novel interfaces were developed to enable '*natural*' communication such as speech and gesture recognition. Systems were envisioned not only to capture environmental parameters but also to sense users' intended emotions and actions. In addition, systems enabled the communication between users and physical objects via object identifications and localisations within the context, due to the Radio Frequency Identification (RFID) was an important basis for applications of ubiquitous computing (Friedewald and Raabe 2011).

On the other hand, even if some technical development was relatively simple, such as the RFID, a number of open questions related to privacy, security and trust still hindered massive diffusion of ubiquitous computing for domestic communication. Bardram (2005)'s study highlighted the incompatible issue of security login in ubiquitous computing environments, as users were constantly access a wide variety of devices. Schmandt and Ackerman (2004) argued that the security and privacy issue were deeply intertwined in ubiquitous computing due to the concerns such as system control and information sharing in communication. Shortly after the rise of ubiquitous computing, in Weiser (1993)'s study, privacy, identify, security and trust were summarised as key issues in ubiquitous computing vision, and attracted a number of studies with different emphasises on domestic communication, such as data collections, searchable large-scale database construction, data mining (Bohn et al. 2004) and private information management (Schmandt and Ackerman 2004).

Smart Homes

Studies in the field of domestic communication have been, recently, heavily concerned with the concept of smart home (see definition in **Appendix 1.1**). In terms of domestic communication, the smart home is a promising way to building and improving access to devices and families. Although Harper (2003) claimed that smart home had not been a hit due to the reasons such as the old house stocking, the tendency of little networked connectivity, and little attention to users, pioneering studies have envisioned its impact on domestic communication, to name a few, such as Georgia Tech's Aware Home, MIT's Place Lab, Microsoft' MS Home and Samsung's Smart Home Project. Taylor et al. (2007) investigated surfaces in the home, and argued that these should be treated not only as places of digital capabilities but also a part of the smart home ecology. Ramos et al. (2008)'s study envisioned the '*context awareness of users*' in communication

within the house. For example, pertinent social and emotional factors would be considered to match users' mood in a conversation.

Mixed Reality (MR)

Milgram and Kishino (1994) coined mixed reality in their study in interaction issues, referring to a mixture of virtual reality (VR) and augmented reality (AR) (Tamura et al. 2001). Due to its combined virtual-real interface, the technology enhanced user perceptions with real object experiences.

MR integrated with communication technologies and used to build enhanced communication experiences. Henrysson et al (2005), developed a MR system based on mobile phones, and found that the system brought better face-to-face communication experiences than normal mobile phone experiences. To help users distinguish virtual and real communication, Billinghurst and Kato (1999) described a shared-space system overlaying virtual tools on physical objects. More commonly, MR was used to visualising information, such as 2D/3D conference environments (Regenbrecht and Wagner 2002), and '*Home Window*' for domestic energy consumption visualisation (Lapides et al. 2009).

2.2.3 Lessons Learnt from the Review of Interactive Technologies

The last section reviewed domestic communication technologies and extracted technological requirements for design and evaluation methodologies. The selection of technologies was not exhaustive, but was typical enough to represent challenges of rapid technology emergence. The understanding of these requirements were summarised as following:

Firstly, technology emergence in the home was rapid, and this required high flexibility of design and evaluation methodologies to build rapid prototypes at various fidelity levels, while also maintaining reliability and validity during the progress. For example, the telephone took nearly half century to become pervasive while to date social networks take much shorter time. Improvement of reliability and validity of these methodologies was aimed to accomplish with the acceleration of technology emergence. However criticism was directed at much of the work that reliability and validity of these methodologies received insufficient

support (Cairns and Cox 2008). Reliability and validity were crucial to technology design and evaluation, since these reflected whether a methodology was '*better or just different*' (Newman and Lamming 1995).

Technologies reviewed above were innovative, and added new ways of interaction to domestic communication routines. Innovative methodologies for HCI research was indeed rare, an alternative approach was incremental improvement (Carroll 2000). By applying limited and controlled changes to a methodology, this research could build both empirical experiences and novel techniques in methodological use (Newman and Lamming 1995).

Secondly, it was essential for design and evaluation methodologies to steadily control design and evaluation procedures, particularly the evaluator's participation in study. There had several challenges to meet this requirement. One challenge was to place evaluator's participation within the context of domestic communication. Plomp and Tealdi (2004) claimed that '*a right application of technology may have revolutionising effects on the way we spend our days, the efficiency we can deal with routine jobs*'. Consequently, methodologies needed to understand domestic environments in order to inform the development of technologies. This required integration of domestic communication technologies and domestic communication routines, such as telephone and gossip habits, paper mails and emails (Elliot et al. 2005; Whittaker and Sidner 1996).

Another challenge was to address the increasing degree of information richness. The richness in this research refers to '*the amount of information that can be conveyed through a communication medium*' (Lo and Lie 2008; Purdy and Nye 2000). It was considered an important variable due to domestic communication study observed that householders had clear preferences to interactive technologies when doing different tasks (Lo and Lie 2008). Thus, design and evaluation methodology needed to support selection of interactive technologies for domestic communication. In this regard, switching between interactive technologies for domestic communication raised high requirements for evaluator's manipulation of design and evaluation procedures.

Other challenges also included addressing increasing connectivity and sociability. The review noted that television had single-way communication, telephone had dual-way communication, and broadband had

limitless users simultaneously such as video conferences (Lagasse and Moermann 2005). This trend required design and evaluation methodologies being able to deal with complicated relationships and private activities (Silverstone et al. 1992). In this regard, evaluator's participation needed to join current domestic communication relationships while not influencing them, which was particularly difficult(Stewart 2003). To overcome this difficulty, design and evaluation methodologies needed to '*open the black box*' of domestic communication and intensive connections in a manner of non-intrusive participation (Morley and Silverstone 1990). Thus, the aim of this challenge was to strike a balance between evaluator's participation and interference.

2.3 Wizard-of-Oz, Why and How

Given the lessons learnt from the review of methodology and technological requirements, this section discusses why WoZ methodology was selected over other design and evaluation methodologies, and why was more suitable for domestic communication study. Furthermore, this section reviews studies that adopted WoZ as primary design and evaluation methodology, and identifies reliability and validity problem of WoZ in practical use.

2.3.1 Why Wizard-of-Oz

The magic wizard (evaluator) in WoZ '*is a small shy man but who operates a large artificial image of* himself' (Sharp et al. 2006). By taking its name from a story in which a little girl was swept away by storm and had an adventure in the land of Oz (Baud and Denslow 1900), WoZ was considered a light-weight methodology that employed a human(s) to facilitate a pseudo system and thus making users think they were interacting with an genuine intelligent system (Dahlback et al. 1993).

By intercepting the communication between system and users, WoZ produces human-like intelligence (Edlund et al. 2008), and supports a wide range of interaction systems and scenarios without having to incur heavy development costs. WoZ was commonly used to investigate natural interfaces (NI) such as speech (Forbes-Riley and Litman 2009), gesture (Hoysniemi et al. 2004) and emotion interfaces (Barliner et al. 2003). In this regard, by mimicking diverse natural interaction (see definition in **Appendix 1.1**) systems at different intelligence levels, WoZ was useful to support future domestic communication study that involved smart devices.

WoZ provides rich flexibility to support smart device development and evaluation. In terms of design it mimicked functions that appeared as genuine smart devices. System components were flexible to switch between various fidelity levels by reconfiguring evaluator's operation. The intelligence level provided by the evaluation's operation covered a wide range from machine-like system to human-like system. Dahlback et al

(1993) considered WoZ a significant solution to empirical studies of natural language interactions, since WoZ provided diverse levels of speech recognition capabilities.

WoZ supported iterative studies by combining design-evaluation and thus shortening the period of design-evaluation-design cycle. Its flexibility of operation enabled rapid system design in cost-effective way. In terms of evaluation, WoZ was an efficient method. Experiment settings can be configured rapidly by changing operation parameters, such as the schema, operation tool, and pre-experiment training. Dow et al. (Dow et al. 2005) claimed that WoZ '*helped designers avoid getting locked into a particular design or working under an incorrect set of assumptions*' due to it enabled fast explorations before considerable investment of design work.

WoZ generated real field-based experiment data, and provided a non-intrusive way of system evaluation. It supported in-field observations on evaluator(s) and subjects, and was compatible with other data collection methods such as video recording and self-reporting (Goldman et al. 2003). WoZ was compatible with wide experimental scenarios, such as the living room and the kitchen, and thus fitted with both laboratory and field studies. WoZ was not only an evaluator-participatory but also a user-participatory methodology. Although light ethical concerns were raised (see **Section 3.4.4**), it provided a natural way to probe users' communication behaviours.

In summary, the reason WoZ methodology was highlighted for future domestic communication study is twofold.

- WoZ is flexible enough to offer various fidelity levels of prototypes with novel domestic communication technologies, thus to set up evaluation experiments in a wide range of study sites, and to control the cost of system design and evaluation.

- WoZ is efficient enough to support iterative studies in non-intrusive ways, to collect field-based data from evaluator- and user-participatory studies, and then to present quick and cheap analysis and discussion mechanisms.

2.3.2 How WoZ Was Used in Previous Studies

WoZ studies focused on three main research areas, including natural interaction (Tscheligi et al. 2009) such as spoken dialogues (Edlund et al. 2008) and gesture manipulations (Paiva et al. 2003), intelligent systems such as conversational robots (Ruyter et al. 2005), and multimodal systems such as wearable multimedia clothing (Mann 1996). These studies were concerned with cutting edge technologies that were still dwelling in laboratories, such as semantic speech understanding (Lapides et al. 2009) and mind reading (Realo et al. 2003). These studies employed one or more human evaluators to facilitate system components, and thus supported investigations in future interactive technologies and their impact.

In order to demonstrate the key system components and flows, **Figure 2.1** illustrates a typical WoZ study.

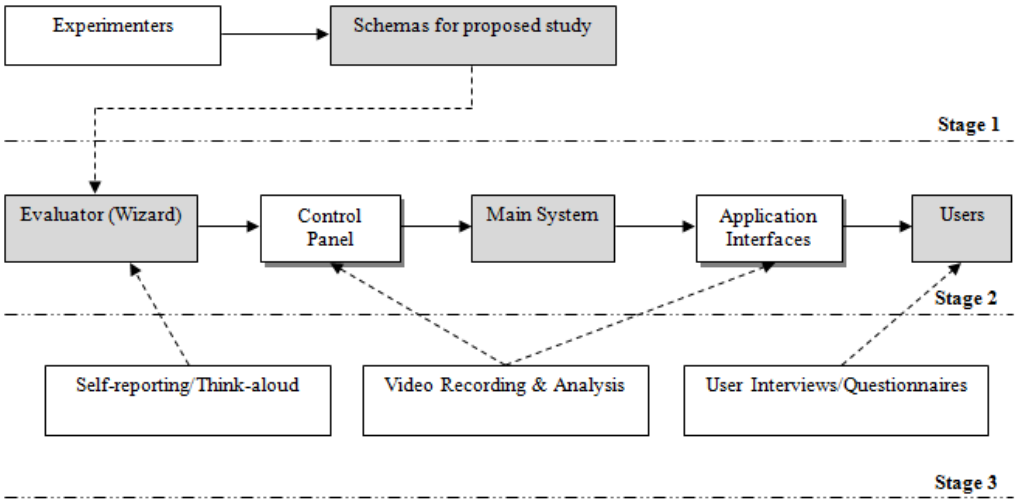

Figure 2.1: WoZ study key components

WoZ was used in various ways across studies. **Firstly**, WoZ was used to mimic simple sensors that detected advanced events such as anger. Hudson et al. (2003)'s study simulated a range of possible sensors based on human coding, and instructed these 'human sensors' to respond to user activities. Given these tasks were specific and simple in the study, the impact of human simulation was not concerned in depth. The human evaluator was seen as a system component, but additional measurements were applied to stabilise evaluator's outputs, such as a vocabulary dictionary. Hoysniemi et al. (2004) carried out a study in a similar

way, they made a human evaluator exclusively responsive to children's actions for game control. Studies using WoZ in the same way also included Voida et al. (2005)'s 2D objects manipulations and Bretan et al. (1995)'s simulation-based dialogue design study. All of these studies gave simple tasks to the evaluator, and were not concerned with autonomous operations. Although evaluator was very capable of dealing with these tasks, a few of these studies had noticed the necessity to constraining evaluator's operation.

Secondly, WoZ was used as an efficient tool for rapid system design. In Clarizio et al. (2006)'s study an evaluator mimicked an embodied conversational agent with a young woman's appearance. A dialogue plan was applied to assure the evaluator followed a consistent logic. Gamm et al. (1997)'s study designed a control panel for the development of a command-based speech interface for telephone-answering. And by using the control panel evaluator was able to play a whole system. Zhang et al. (2006) described similar use of WoZ simulation as an integrated intelligent system. The study provided multiple communication channels for evaluator to use, however how the evaluator facilitated these channels was not described in depth. In Forbes-Riley and Litman (2011)'s study WoZ was used to take the place of multiple system components, the combination of which presented a full intelligent tutor system. These studies employed WoZ to play whole systems, however, details of evaluator's operations were rarely provided. These normally developed a control panel for system operation, and used this to constrain evaluator's operations by providing specific functionalities in control panels.

Thirdly, WoZ was a reliable tool for in-field experiment data collection. Liu et al. (2009) presented an open-source data-collection tool based on WoZ to study information presentation strategies. Paiva et al. (2003) and Andersson et al. (2002) adopted WoZ as a means of data collection to inform further system development. In Paiva et al. (2003)'s study evaluator recognised users' toy movements and returned respective emotions, thus collecting users' preliminary preferences to emotional toy use. In Novielli et al. (2010)'s study, evaluator was used in the same way to collect users' attitudes towards an embodied conversational agent. Similarly, Yildirim et al. (2011) and Rosis et al. (2006) used WoZ to capture rough data of emotional states for further conversational interface design. Studies like the ones mentioned above were not planned with a strict operation routine, as various user actions were expected to be collected. However, light-weight constraints to evaluator's operation were applied in some of these studies. For example, the

smart conversational toy must be possible and convincing (Wooffitt et al. 1997), to make users produce real data.

And **finally**, WoZ was a flexible evaluation tool. Whittaker et al. (2002) used WoZ to work on evaluation of spoken dialogue in the application domain of a restaurant. In the study a complex control panel was developed for user interactions, and the evaluator was required to follow rigorous operation procedures. In Salber and Coutaz (1993)'s evaluation study the system operation was divided into three parts, and each part was assigned a respective evaluator to maintain high quality operations. In addition, pre-study training was provided to evaluators to reduce improvisation. Simpler evaluations were carried out by Balbo et al. (1993), in which a WoZ platform was used to capture and mix digitised and analogical data automatically. Fraser and Gilbert (1991) considered more details related to speech-based WoZ evaluations, such as task variables and subject variables. The study particularly highlighted the dynamic nature of evaluator-related variables, such as recognition and production variables. WoZ was also commonly used to evaluate robotic intelligent systems, such as Ruyter et al. (2005)'s assessment on the effects of social intelligent robot interface. These studies presented limited flexibility in terms of evaluator's operation. More strict strategies of WoZ use were observed across studies, such as additional evaluator training and rigorous operation instructions.

In summary, questions are summarised from the review of WoZ use as follows.

- How should the evaluator training be assessed prior to study, and what requirements should the evaluator meet before conducting a study? Little evidence was provided in previous studies that described details of pre-study training and evaluation.

- How does the evaluator follow the planned schemas, and what measurements were adopted to reflect the extent of schema following? The level of understanding and execution of schemas on the part of the evaluator was implicit in previous studies, since most of these studies employed experienced researchers to operate the system. It was assumed that experienced researchers could follow schemas strictly.

- How should control panel component be designed, and how would this affect the evaluator's operation? Throughout the review of WoZ use, few studies noted the impact of control panels on evaluator's operation. However, across different control panels, different evaluation operation effectiveness, reliability and validity were observed.

- How does the evaluator's personal behaviour affect system operation? Evaluator was considered a variable that might affect system operation in (Fraser and Gilbert 1991)'s speech simulation study. However, understanding of what were these specific variables and how system operation was affected was still primitive.

2.3.3 Research Questions about the Reliability and Validity of WoZ

The review of WoZ study confirmed that the reliability and validity of WoZ were influenced by variables in the study. To further understand the influence variables, a set of questions were described as follows.

Firstly, how these variables affected evaluator's operation was not sufficiently considered in previous studies, although advantages such as flexibility and effectiveness were described? Hudson et al. (2003)'s human-coding sensors might be affected by user interactions, and thus had risks of inconsistent operation. Gould et al. (1983)'s speech recogniser had few considerations on the impact of subject's interactions due to the fact that they were located in a same room. In this case subjects were aware of evaluator, and system interactions collected by WoZ might be inaccurate. In addition, previous studies rarely mentioned the impact of different operation devices, such as header mounted displays and monitors. Therefore, further investigation in influence variables was compelling. These variables included evaluator's exposure, location and operation devices.

Secondly, how the supports to multimodal system operation were proposed, although the adaptive nature of WoZ was demonstrated in intelligent system development? As descriptions of technical requirements, reliability and validity might be more variable in multimodal system operation. Salber and Coutaz (1993) employed multiple evaluators to address this concern, however this introduced higher system maintenance costs and more operation variables. Balbo et al. (1993) attempted to set up automatic mechanisms for

multimodal system operation and evaluation, but the results showed that similar reliability and validity problems existed, especially in terms of control panel operation. In summary, it was claimed that evaluator's operation for intelligent systems still faced weak multimodal operation supports.

Thirdly, what were the risks of using different evaluators for system operation which were not highlighted in previous studies? The study overview showed that most evaluators used to be played by a member of research groups. Advantages of this were that evaluators were familiar with schemas, operation system and study purposes. On the other hand the risk was that the selected evaluator might be too familiar with system operations and neglect crucial operations by instinct. In addition, the impact of switching different evaluators – even if they were well trained – on WoZ reliability and validity was still unclear.

Fourthly, what were other experimental variables affected reliability and validity of WoZ methodology? Some of these variables were described in **Section 2.3.2**. The first was the lack of assessment of evaluator's pre-study training. Salber and Coutaz (1993)'s study split system operation tasks into three independent parts with different evaluators. Although the aim of the split was to reduce operation burden, it actually increased load of extra evaluator training and had the training without assessment. In this case the uncertainty of evaluator training might have risks of improvisational operations. The second was evaluator's schema complying. Schemas were popularly used in WoZ studies to guide evaluator's operations. However, little evidence was found about how evaluator interpreted and executed these schemas. As reported in Fraser and Gilbert (1991)'s study, some occasions such as unrecognised speech could cause improvisational operations which were not planned in schemas. The third was that operation facilities such as control panel had influence on evaluator's operation. Since control panel was the direct tool used by evaluator, system operation was vulnerable to be affected by control panel use. And finally several personal variables of evaluator were considered relevant with the reliability and validity of WoZ. Evaluator might discriminate system operations due to individual preferences.

2.3.4 Lessons Learnt from the Review of WoZ Studies

As claimed previously (see **Section 2.3.2**) WoZ was a suitable method for domestic communication study. Due to evaluator's operation WoZ had high flexibility and effectiveness in intelligent system design and evaluation. However, as other evaluator-participatory design and evaluation methodologies, WoZ also had similar issues of variable reliability and validity. A wide range of variables were identified across previous WoZ studies, as well as their influence on the reliability and validity of WoZ. These variables were distributed around all stages of study. Therefore a summary was provided to categorise these variables according to relevance with system components, and based on which to present understanding the reliability and validity of WoZ.

Firstly, variables described above were relevant with the schema, including schema design, schema interpretation and schema execution. In terms of schema design, the anticipation of system interactions might affect the extent to which evaluator needed to make improvisational operations. In terms of schema interpretation, it was a great matter of how the evaluator understood these schemas. Previous studies showed evident examples of evaluator's speech misunderstanding, and similar occasions also existed in schema interpretation. And in terms of schema execution, evaluator's operation plan was also affected according to schema's requirement of output operation.

Secondly, variables were identified to be relevant with control panel. Studies described control panel as a tool to facilitate system operations. However control panel had more influence on system operation. One was that control panel might be useful to apply constraints of evaluator's improvisational operations. The 'Y/N' control panel design provided predictable and consistent operations, although these operations were simple. However, similar examples indicated that control panel design – especially function design – was able to affect the reliability and validity of WoZ. Another influence of control panel was the operation effectiveness. As the evaluator in WoZ study was expected to present rigid operations as computers did, variable response speed caused inconsistent operations, and lowered reliability and validity of study results.

And finally, the evaluator's personality and preferences were considered crucial matters for the reliability and validity of WoZ. These variables were adherent with individual evaluators, as previous studies admitted

the difference between evaluators and system operations. However, consensus was not reached on how much individual evaluator differences remained after training and how that difference could affect the reliability and validity of WoZ.

In summary, all variables described above generated inconsistent operations in WoZ study, which led directly to variable reliability and validity. Improvement of one category of variables may not address the problem, and systematic studies are required to carefully control and investigate the impact of these variables.

2.4 Summary of Literature Review

The review presented a series of technical and methodological requirements for variable reliability and validity study. The review especially examined the state of variable reliability and validity in WoZ studies. Taking these together, WoZ and its use in empirical studies were surveyed, and this raised several questions.

The understanding of technological and methodological requirements for reliability and validity study was described in respective sections. In this section a summary was provided to discuss the key research gaps and questions for the proposed study. The subsequence of this discussion pointed to the consistency of evaluator operation at three levels.

Validity of WoZ, defined as the degree to which the researchers could gain accurate and consistent results of any design and evaluation made through WoZ operation, were discussed as follows.

- The unspecified quality of evaluator's pre-study training generated unaware improvisations, thus leading to inconsistent operations. These inconsistent operations are most relevant to schemas, since the trainings used to be based on experimental scenarios. However, no severe impact was reported across current WoZ studies. This was justified for two reasons. First, the poorly trained part was heavily used in operation, and second, the inconsistency receives little attention due to its unawareness.

- The unknown impact of schema design on evaluator's operation may be underestimated by current studies. A wide variety of schemas have been observed across WoZ studies. Even if the target operations were the same between two studies, schemas were constituted with different rules. This was rarely estimated in current studies. The reason may be that most WoZ studies employed research team members and thus covered this issue due to their familiarity of schemas and systems. However, if, to make WoZ a non-specialist required methodology, this validity issue may need further improvement.

- The control panels may generate considerable impact on system operation. Control panel is commonly used but less investigated in current studies. One exception was (Fraser and Gilbert)'s study, which

indicated the potential impact of control panel on speech system operation. However, the impact on other systems such as affective agents (Paiva et al. 2003) requires more attention.

Reliability, defined as the levels of consistency in yielding the same results of system operation styles and content across different WoZ operation trials, was discussed as follows.

- Multimodal systems may generate noticeable inconsistent operations. Since multiple input media are simultaneously available in system, WoZ is also extended to support multimodal system operations. The increase of system modalities brought two challenges that could cause severe operation inconsistency. First, if multiple evaluators were used, the synchronisation of cooperative operation between evaluators would be a problem; second, if single evaluator was used, the system would take account of the coordination of system components. Both challenges provided evaluator more change to generate inconsistent operations.

- Evaluator's interpretation on subject's activities may generate considerable inconsistent operations. The interpretation is seen one of most vulnerable parts in the operation, due to most misrecognitions and improvisations happened in this process. Due to the difference of background knowledge between evaluators and subjects, misunderstanding took place and led to inconsistent operations. In addition, previous issues such as unspecific training and schemas may also cause similar consequences of inconsistent operation, by affecting the interpretation of operation.

- Will different evaluators, even if they are all trained, bring alternative operations into a system? The difference between novice and experienced evaluators were rarely reported across previous studies, due to most evaluators were experienced members of research team, or system designers. However, individual preferences to applications and operations were not highlighted in these studies. This could cause light inconsistency of operation. Using experienced researchers could maintain high quality consistent operation; on the one hand, it could bring more inconsistency on the other hand, when another research member took the place. In addition, research member employment could hinder WoZ to be a reliable and low-maintenance method.

In summary the evaluator's operation was the paramount concern amongst these issues. Three factors have been identified as most possible causes of inconsistent operations, including the schema, the operation interpretation and the evaluator differences. Correspondingly, these three factors point to respective levels of operation inconsistency. Therefore the research gaps were shaped as follows:

- Inconsistency issues related to schemas
- Inconsistency issues related to evaluator 's interpretations (and system operations)
- And, inconsistency issues related to different evaluators and evaluators' differences in operation preferences

Based on the three levels of inconsistency issues, the next chapter aims to build a concrete research agenda to address these variables respectively.

Chapter 3

Specifications of WoZ Platform for Proposed Studies

3.1 Introduction of WoZ Platform

In literature review variables that were relevant with the reliability and validity of WoZ were summarised in three research questions. These pointed to different levels of inconsistent system facilitation. This chapter provides a closer look at the platform construction for WoZ studies, in order to build a practical grounding for future studies. Technical specifications for WoZ platform design, implement and evaluation were considered. There include technological challenges of WoZ platform system development, WoZ study preconditions and other methodological configurations such as subjects and evaluation criteria. Lastly ethical concerns of WoZ study are described, as all proposed WoZ studies are concerned with pseudo system operations.

This chapter is organised as follows. Firstly, based on findings in **Section 2.4.1**, **Section 3.2** provides an overview of implications for future WoZ study. Specific study objectives are described in this section. Secondly, **Section 3.3** provides a brief review of technologies that are essential to intelligent system development. The review leads to a short discussion of challenges in integrating these technologies in WoZ system design. Thirdly, **Section 3.4** moves on to describe proposed study method, including preconditions, subjects, evaluation criteria, and ethical issue. And finally **Section 3.5** summarises these specification for the first study.

3.2 Refinement of WoZ Study Objectives

The literature review in **Chapter 2** has summarised research questions and their relationships with evaluator system facilitation. Especially the reliability and validity of WoZ method was identified with a number of variables that affected evaluator's study management and control. Based on this understanding this section provides refinements of proposed study objectives.

Firstly, findings in literature review noted that the schema was a primary variable due to its crucial influence on evaluator's operation. The organisation of schemas was in various forms, such as informal verbal schemas that were used in pre-study training and incrementally-documented schemas (Molin 2004). All these schemas were generally in two categories according to the constraint to system operations. One was rigorous schemas which planned operation rules for every system interaction, and the other type was general schemas that only covered broad operation principles. Both types of schemas existed commonly in previous studies, but understanding of these schemas was primitive. In Bickmore (2002)'s opinion rigorous schemas provided inflexible operation rules and were too strict for unexpected system interactions. Similarly, such schemas might be over complicated for evaluator to memorise and execute in real-time natural-language dialogue system operation (Fraser and Gilbert 1991). On the other hand, general schemas fitted well with simple tasks such as data collection. Due to such schemas had great flexibility of interpretation, there were high risks of improvisational operation. In a very few cases both types of schemas were combined for system operation (Bickmore 2002), but consequences of the combination were not explicitly described.

Secondly, a number of variables were identified relevant with the reliability and validity of WoZ, as these were able to affect evaluator's operation consistency. Amongst these variables, control panel design and evaluator's interpretation were particularly highlighted due to their direct influence on system operation.

The literature review surveyed a wide variety of control panel designs, from simple 'Y/N' buttons (Tsukahara and Ward 2001) to complex tick-box interfaces (Whittaker et al. 2002). The simple design introduced high constraints to evaluator's operation, and generated predictive operations. However, this was not capable with complex operations, particularly with intelligent systems that had multiple communication channels

(Forbes-Riley and Litman 2009). On the other hand, complicated control panel design might lower the learnability and slow down evaluator's operation speed.

Evaluator's interpretations, such as interpretations of schemas and subject activities, were given heavy emphasises in previous review due to the dynamic nature of interpretation. Variables relevant with schema interpretation were considered an important reason for variable operation consistency, as these produced different schema understanding and led to diverse system operation.

In terms of subject's activity interpretation, more variables were involved such as the knowledge differences and subject's intention anticipation. Some measurements were adopted to improve interpretation accuracy across previous studies, such as direct user observations (Carbini et al. 2006) and strict experiment settings (Bradley et al. 2009). However, these measurements focused on interpretations of planned activities. Little attention was paid to system interactions that were not anticipated in schemas. These unplanned activities were likely to entice the evaluator to generate improvisational operations.

This suggested more endeavours to improve evaluator's interpretation capabilities, especially in terms of unplanned activities. In other words, study plans needed to draw strategies to help evaluator anticipate subject's activities, and thus preparing for operations in advance. The aim of this study plan was to improve the reliability and validity of WoZ through producing consistent operations. To achieve this, the research needed to compare different subject activity interpretation results provided by evaluator and study observers.

And finally, the research was aimed at looking into different evaluators and their individual differences of operation. Most WoZ studies employed expert-level evaluators. Such high-standard evaluator employment had an advantage of maintaining high-quality operations. However, this was limited by evaluators' individual operation preferences. The reliability and validity of WoZ could be affected by individual experiences if different evaluators were used in a study. When the expert evaluator remained throughout the study, the change of system application might also cause risks of operation inconsistency due to the difference of evaluator's adoption to new system application. Therefore, study plans should be proposed to evaluate different but trained evaluators' operations in terms of operation consistency levels. This would involve

different evaluators as dependent variables (both experienced and novice evaluators were employed and given training).

In summary, this section presented several implications that were useful in shaping the direction of future research. These implications focused on evaluator's operation variables and suggested four important variables as key research objectives. These included schema design, control panel design, subject's activity interpretation and multiple evaluators. Particularly, schema design was considered as the primary variable that needed to be addressed due to all following variable study concerned with it. In next section the discussion moves on to examine technical challenges and preconditions for WoZ study.

3.3 Specifications of WoZ Platform Construction

3.3.1 Requirements for Intelligent System Development

According to the concept of intelligent system defined in this research, the intelligence refers to system abilities that '*proactively, but sensibly, support people in their daily lives* (Augusto and McCullagh 2007)'. Some techniques are particularly highlighted by such intelligent system in domestic environments and scenarios, such as speech recognition, emotion recognition, natural language input, vision input, and mixed realities (Ramos et al. 2008). Due to some cutting-edge techniques may be too sensitive for domestic communication such as mind reading (Realo et al. 2003), or too difficult for evaluators to mimic such as vision input, a small amount of techniques was selected and developed in this work of research.

The first difficulty of WoZ system development for intelligent system operation was natural-language speech recognition. Natural-language speech recognition was considered a key technique due to its importance to system's sense of intelligence. To date speech recognition systems were designed with a small set of vocabularies (Bretan et al. 1995)[1]. However, these vocabularies consisted of specific system operation commands. For natural-language speech recognition there needed to integrate a large number of vocabularies to cover these vocabularies used in system interaction. In addition, it was challengeable to maintain evaluator's speech recognition performance at a consistent level.

Computing vision, more specifically gesture recognition in this research, was another difficulty for operation system development. Gesture recognition was an important technique for intelligent systems due to its advantages of naturalness of interaction. Gestures naturally performed by humans make much powerful interactions than normal interactions (Carbini et al. 2006). However, the state-of-the-art in continuous

[1] By the time of submitting this thesis, applications such as Apple® SiRi and Google® Voice have realised robust natural speech recognition covering a wide range of speech vocabularies.

gesture recognition did not demonstrate mature and massive-scale applications for domestic communication use, except in some pioneering projects in laboratory-based smart homes (Lee et al. 2011). Thereby, the challenge was to present high level natural gesture recognition while also maintains technology faithfulness in terms of recognition accuracy.

Augmented reality is a technique that the research intended to integrate into intelligent system. Although current augmented reality is not a dominated technique in the home, its attraction is justified for two main reasons. The first is the seamless connections between virtual and physical worlds as information enhanced environments, in which accurate tracking techniques such as speech and gesture recognition could be integrated naturally. The Second is the augmented reality could make a significant improvement in domestic communication, due to its capabilities of contextual information interpretation and re-modelling (Ramos et al. 2008).

The aim of this section is not to improve these technologies for intelligent system development. Instead it judges what techniques WoZ operation system fits with, and how suitable are these techniques for intelligent system development.

3.3.2 Requirements for Conversational System Development

Speech recognition was achieved by a human evaluator in this research, but this faced some barriers. The speech used by subjects had a great amount of dynamics such as genders, accents and emotion states. These dynamics could cause unintentional inconsistent operation, as evaluators may not respond accurately to unplanned speech. There was another challenge that evaluators need to be aware. If the evaluator used text-typing to respond to the speech, then the typing speed might cause noticeable delays that cause subjects' awareness. Salber and Coutaz (1993) used predefined replies to reduce such delay. Buisine et al. (2005) noticed the delay caused by speech recognition simulation, and pointed that the semantic understanding would be slightly affected by that. In addition the 'paralinguistic information' (Yang and Lugger 2010) was also influenced, as the evaluator needed to overcome human-human communication styles in speech segmentation and recognition, and instead to mimic computer styles in word segmentations that only

extract key words. Therefore the evaluator needed to adjust attentions to speech recognition from semantic sentences to key words. Consequently mechanisms were required to assist the evaluator to achieve computer-style speech recognition.

Gesture recognition faced a similar challenge as it needed to identify intentional movements from others. Subjects generated continuous limb movements and that made gestures such as 'endpoint localisation' difficult to recognise (Li and Greenspan 2011). This could become worse when the evaluator was provided limited observations, for example, observation from a hidden camera.

Extra attention was paid to augmented reality development in this research. Firstly, the research needed to identify a suitable surface to project virtual realities, as this provided different interactions from desktop monitors. The surface affected forms of representing combined virtual-physical realities. Secondly, the devices to represent combined virtual-physical worlds also mattered. For example, head-mounted displays differentiated from projector-based displays in terms of user perceptions to a system (Voida et al. 2005). These concerns needed to be properly addressed before applying the technique for domestic communication system developments.

In summary these challenges needed to be addressed prior to WoZ studies. These considerations were not limited in terms of technology development, but with the integration of evaluator's operations. In this regard experimental WoZ operation system development needs to consider a way for evaluator to participate and play part of these techniques.

3.3.3 A Trial Configuration of WoZ Platform

Before the start of practical WoZ platform development, a preliminary study was carried out for a trial configuration of WoZ study platform, in order to understand how technologies were applicable for smart device design in the laboratory environments and how the key system components should be configured. The findings of this study provided a practical grounding for the following technological development work. To make an easy understanding of these findings, a concise description of study was presented here. Part of this study was reported in other publications with different emphasises in (Bonner et al. 2009).

The study developed an augmented reality-based control panel for domestic central heating control. This prototype was firstly designed as a virtual control panel (see **Figure 3.1**), and then it evolved with a colour detection function using which to map dynamically control panel interfaces to a moving cube. By simulating a smart control panel that displayed anticipated gas and electricity consumption in pounds (see **Figure 3.2**), the study provided empirical experiences for computer vision-based smart device design. The usability of these designs was assessed in this study by conducting task-accomplishing evaluations. The preliminary findings demonstrated the applicability of integrating systems with domestic furniture such as coffee table. In addition the study indicated the usefulness of improving conventional interactions.

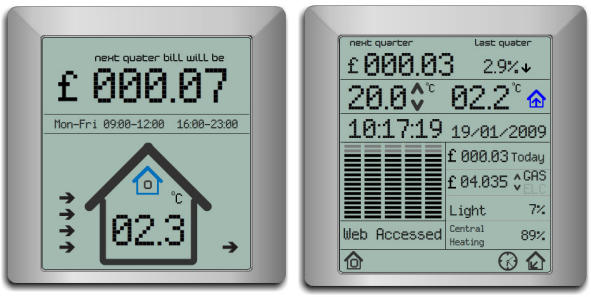

Figure 3.1: *Virtual control panel of domestic central heating*

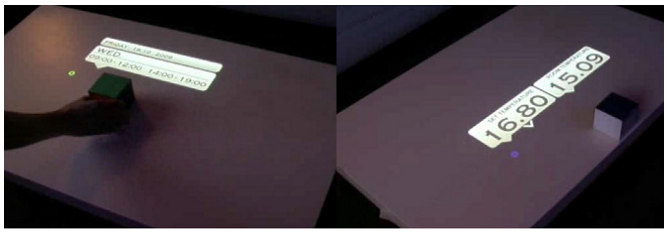

Figure 3.2: *Control panel interfaces following the colour cube*

3.4 Specifications of Proposed WoZ Study

3.4.1 Conditions of Effective WoZ Study

This section discusses the conditions that lead to effective WoZ studies in simulated laboratory environments. Based on Wooffitt et al. (1997)'s framework, the discussion consisted of three main aspects, including availability, faithfulness, and maintainability.

Section 3.2 described some inappropriate techniques for WoZ simulations, these were either impossible for evaluator to mimic or unavailable for evaluator to simulate. For example, it was anticipated that future domestic communication would have considerable connections and such connections have not yet been fully realised, but there is little sense in setting up simulations for this. This did not imply that evaluators should only mimic systems falling within their capabilities. This might indicate that, by mimicking system components, evaluators should pay more attentions to the availability of simulations. In other words, evaluators should not aim to mimic system components that computers could do far much better than they do. Fraser and Gilbert (1991)'s example of visual object discrimination makes good pointer to the availability condition.

The second condition is that the mimics provided by evaluators need to be faithful to current technology performances. There are a number of WoZ studies that applied additional mechanisms to achieve this. In Wooffitt et al. (1997)'s study a few unintentional speech recognition mistakes were used to excuse technical faults. While it seems unlikely to carry on such error-making strategies, as technologies make fast progress and the intelligent system learns autonomously. The dilemma was that the operation was proposed to mimic advance technologies while the contents of operation need to align with current technology levels, to make it thus faithful to existing technologies.

The third condition for the useful WoZ simulation is that the pseudo system intelligence should be convincingly maintained as if a real smart system. Most of current WoZ studies hid evaluators invisibly and used slight 'deceptions' for that reason. Lower information richness (see **Section 2.2**), such as texts and

disguised human voices, was used to maintain the illusion. This problem was highlighted in controlling evaluators' interactions with users. The extent to how severe this condition was to operation inconsistency levels was likely relevant to users' gullibility (Wooffitt et al. 1997), which varied across different groups of users.

Given these conditions this section draws additional and critical constraints to study plans. It was indicated that the system components facilitated by the evaluator should be identified as such that a human can naturally do much better than computers. It also showed that the system component the evaluator was going to facilitate should be supported with some extra mechanisms, thus to continuously maintain the intelligent system illusion, and to address the dilemma of technology novelty and faithfulness.

3.4.2 Subjects (Participants)

Considered thoughts were given about how many subjects were needed and what qualifications these subjects needed to meet. This had crucial influence on WoZ operation system development in terms of schema planning and control panel construction.

In **Section 2.4** studies were concisely described in terms of subject employment and qualification. One crucial concern was how many subjects should be involved in study. Previous studies intended to employ a small amount of subjects in iterative evaluations. One reason was that a small amount of subjects were configurable in design and evaluation study. In some studies which used considerable subjects, such as the in-service telephone system study (Wirén et al. 2007), a large amount of overlaps were noted in study results. Previous studies used to employ an average number of subjects from 4 to 12 (Bickmore 2002; Hudson et al. 2003). One extreme example of study used 32 subjects in studies, but these subjects were used in separate experiments (Hauptmann 1989; Hoysniemi et al. 2004). An attractive way was to employ a few carefully-selected subjects and put them in iterative studies, such as studies in (Bretan et al. 1995; Dahlback et al. 1993; Dow et al. 2010).

With some caution, the interpretation taken in these studies was twofold. The first approach was that the systems facilitated by evaluator had specific users, and this constrained subject selection. The second, the

approach was that the operation purposes affected subject employment. For example a data collection WoZ study needed more subjects, but an interaction understanding study required specific subjects.

3.4.3 Evaluation Criteria for Evaluator Operation

A set of evaluation criteria for operation inconsistency level assessment in WoZ studies were described. Due to that different system applications would be developed in the proposed studies, the criteria derived in this section focused on a general framework of operation assessment that may be suitable for all proposed studies.

A number of criteria were used across WoZ studies, including the richness of experiment data (Oviatt et al. 2004), the system respond speed (Carbini et al. 2006), and the adherence to schemas (Maulsby et al. 1993). These criteria were effective in assessing specific aspects of WoZ study such as system usability. However in proposed studies a wide range of aspects was concerned from schema interpretations to control panel operation, these were not well covered by previous evaluation criteria.

In terms of schema design the proposed criteria needed to support evaluations of the usability, learnability and accuracy of evaluator's schema interpretations. In making accurate schema understanding and generating consistent operations, evaluator's improvisations were evaluated using a range of validity criteria. The criteria for validity were defined as the extent to which the evaluator could make steady schema interpretations, thus leading to consistent operations. These aspects needed to be assessed comprehensively since there were strong relationships between schemas and operations.

Making the criteria suitable to assessing dynamic operation factors was therefore essential. Crucial criteria offered accurate dimensions to cover influence variables. Generally, the criteria should be derived to reflect the accuracy of evaluator's interpretation on subject activities, the usability of control panel, the effectiveness of control panel operation and the appropriateness of improvisational operations. It was important that the criteria could provide a comprehensive and complete scale with which the operation inconsistency level could be independently reflected.

The proposed criteria needed to support assessment of reliability of different evaluators' operation. Reliability criteria were interpreted as the degree of consistency to which different evaluators, including experienced and novice evaluators, could make the same interpretations on schemas and subject activities and generate equivalent operations. As the experiment data would be generated with different evaluators, the reliability needed to be examined cross evaluators. It was important that the criteria could be suitable with cross-evaluators assessments and provide an independent and efficient benchmark to reflect operation inconsistency levels.

To summarise these requirements for the evaluation of the reliability and validity of WoZ, a set of evaluation criteria were derived as follows.

- *Stability of evaluator's operations* – This is defined as the level of stability that evaluators have consistent reactions to schemas, subject activity interpretations and control panel manipulations.

- *Predictability of participants' interactions* – This is defined as the extent to which the evaluator is able to anticipate the next subject activities via observations and control panel operation preparations.

- *Appropriateness of evaluator's operations to schemas* – This is defined as the degree of appropriateness of evaluator's operations. This could be affected by the following factors, including the accuracy of selecting suitable schemas and the accuracy of control panel manipulations.

- *Effectiveness of evaluator's operations* – This is interpreted as how fast the evaluator could generate the operations that are suitable for subject activities.

3.4.4 Ethical Issues

Ethical issues in WoZ studies were admitted and unavoidable due to methodological approach in principle operations. Dahlback et al. (1993) claimed that the ethical issues could be overcome. On ethical grounds researchers such as Fraser and Gilbert (1991) were against serious deception of subjects by using pseudo system.

There are reasons for carrying out WoZ study using pseudo systems. As pointed out by Dahlback et al. (1993) there were differences of human interactions with computers and humans. And more important, to the respect of system operation, especially such that subjects had little conscious awareness such as typing rates, it was important to maintain the illusion by light-weight deception. On the other hand, if subjects expected to understand the limitations of intelligent system such as natural language understanding, it was allowed to reveal the system operation by demonstrating the technology limitations.

For the work of research related to intelligent system operations, the author claimed that if one subject participated experiment practices and semi-formal interviews afterwards, explicit explanations would be provided which introduced what the experiment was carried out for, what data was collected, and in what way the data was analysed and used. And furthermore consent forms were provided to subjects, according to which the data collected from experiments would be dedicated to limited laboratory use only, and at any time subjects were granted to request data destroy.

As the ethical issue was sensitive, and was commonly concerned in the proposed studies, an application was made to university ethical committee (see **Appendix 3.1**), and was approved to carry out proposed experiments.

3.5 Summary

The discussion of proposed study objectives and technical specifications provided more specific research objectives for WoZ studies. It refined the research gaps to practical considerations on WoZ operation system development and implementation. Four main system operation variables were identified as primary research objectives, including schema, control panel, interpretation of schema and subject activity and multiple evaluators. Each variable corresponded to a dedicated group of WoZ studies as distributed in the following chapters.

Chapter 4

Schema and the Impact in WoZ Study

4.1 Introduction of Study Objectives and Scope

This chapter reports an empirical WoZ study relevant with schema in WoZ study. A primary objective of this study was to develop holistic understanding of schema variable and its impact on evaluator's operation consistency. This was achieved through comparing experimental data across three schemas including a rigorous schema design, a general schema design and a combined schema design. These schemas were designed with different levels of rigorousness in terms of operation guidance and restriction. The study scope focused on the impact of schema design which differentiated it from previous WoZ studies that emphasised on specific trainings and operation systems (based on findings in **Section 2.3.3**).

Findings in this chapter, based on the practical consideration reported in **Chapter 3**, provided an empirical grounding to future development work. The work described in this chapter was partially reported in other publications with different emphasises (Li and Bonner 2010).

This chapter is organised as following. **Section 4.2** reports the study method, including selecting subjects, data collection, and data analysis. **Section 4.3** provides an overview of the study results which are further analysed and discussed from **Section 4.4 to 4.7**. And finally, **Section 4.8** discusses the main findings from the chapter and summarises the conclusions as well as future research directions. **Section 4.9** summarised the conclusions lastly.

4.1.1 Study Objectives

This study's objectives were as follows:

- To understand how schema-related variables affected evaluator's operation consistency and thus caused variable reliability and validity – **Section 2.4.1** noted a question that schemas in WoZ study were various, but little attention was paid to the impact of these schema variables. The primary aim of this study was to step forward to understand the impact.

- To understand requirements for operation system and smart device design for domestic communication study – A secondary commitment of this research was to explore the ways to designing reliable WoZ operation system, and to evaluate smart systems for domestic communication.

- To deepen the understanding of relevant research problems – Since WoZ concerned with design and evaluation variables, a strong need was proposed for the understanding of relevant research questions, such as control panel operation.

4.1.2. Study Scope

The primary goal of this study was to improve the reliability and validity of WoZ in terms of schema design. To offset the complexity of experimental design and implementation, the scope of this study was defined as follows:

- The study focused on activities within the domain of domestic communication – This was constrained to the interactions between conversational computers and people in the home. More specifically, it was constrained to natural-language speech dialogues with computers. Thus the extra complexity of manipulation applications, such as multimedia operation and physical device development, was avoided.

- The study focused on evaluator's interactions with three levels of schema design – Although the evaluator had heavy interactions with subjects in WoZ study, the emphasis of study was on the side of

schemas' impact on evaluator's operation performances. A further focus was aimed in schema design as described in **Section 4.2.**

- The study was a short-term study – As described in **Chapter 3,** conversations with smart system lasted shortly and conversational contents were largely limited by the purpose of conversations. Therefore, this study was carried out and based on one-off experiment procedures.

- The study focused on single evaluator rather than multiple evaluators in this stage – As noted in **Chapter 3**, multiple evaluators and their differences of operation would be investigated in later studies. Single evaluator study also avoided issues related to paralleling system design.

4.2 Study 1 - Method of Proposed Study

This section describes how the study was conducted through following a study structure which was used throughout the following studies. Firstly, technical development described the rationales and design of the system. Secondly, experimental design detailed the parameters of study such as study structure, subjects and data collection method. And finally the study procedure was provided to demonstrate the flow of experiment.

4.2.1 System Development

A natural language-based application was proposed with natural-language interactions. Speech recognition was realised through the evaluator's system operation. The preliminary study in **Section 3.3.3** indicated the applicability of computer vision-based application while it also highlighted the dynamic nature of colour detection in complicated illumination environments. Thus in this study a cube with black-and-white patterns was used for higher recognition accuracy (see **Figure 4.1**). The study defined the patterns as specifically designed illustrations that were easy to recognise while representing different functions.

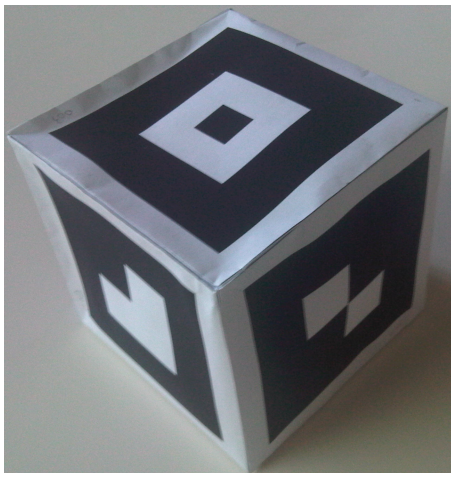

Figure 4.1: *White and black patterns on cube surfaces*

System components consisted of front-end applications and a control panel for evaluator operation (see **Figure 4.2, Figure 4.3**). These components were connected via intranet, as demonstrated in **Figure 4.3**. Front-end applications retrieved operation commands from the control panel in every 100 milliseconds, which was fast enough to catch the evaluator's operations.

Figure 4.2*: Front-end application interface*

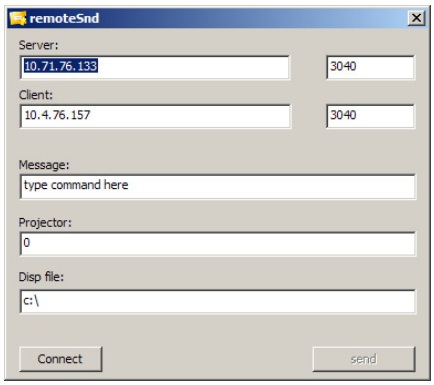

Figure 4.3*: Control panel*

Due to the fact that the evaluator was located in a different room, a surveillance channel was set up for the evaluator to monitor subjects' activities. The channel broadcasted encrypted video stream, and the evaluator received the video via an internet explorer. The structure of the connection and video broadcasting was illustrated in **Figure 4.4**.

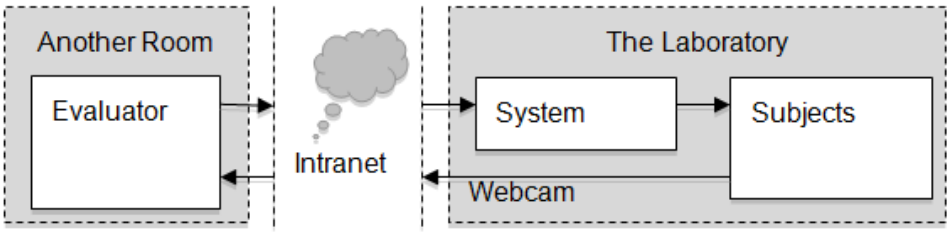

Figure 4.4*: WoZ operation system design*

4.2.2 Study Structure and Evaluation Variables

Study Variables

The system was divided into four main components as shown in **Figure 4.5**. Three schemas with different 'rigorousness levels' (see definition in **Appendix 1.1**) were set as dependent variables. Other system components, such as the evaluator, system, and subjects were seen as independent variables. This setting remained throughout the studies in this chapter.

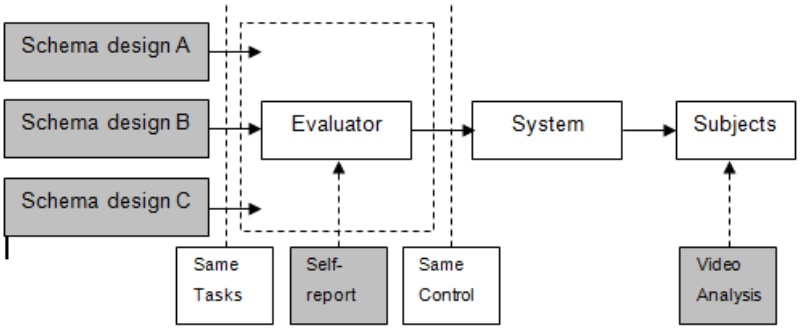

Figure 4.5*: Study structure and components of operation system*

Data Collection Method

There were two data capturing methods used in this study – video analysis and self-reporting. Video analysis aimed to record and observe subjects' system interactions, and self-reporting aimed to reflect evaluator's

operation experiences (Goldman et al. 2003). The use of these methods was justified for following reasons. Firstly, video analysis was a standard data capturing method adopted to enable repeated experimental analysis. It was successfully employed in a number of WoZ studies (Wooffitt et al. 1997). In this study it recorded subjects' activities and generated video footages for deep analysis. Secondly self-reporting was an efficient method to collect field study data from evaluator's point of view (Shami et al. 2008). In addition, the accuracy and validity of self-reporting were proved in a number of applications such as characterising instant messaging (Ellen et al. 2002) and social network activity (Moira et al. 2010).

Subjects

Three groups of subjects took part in this study. Each group consisted of 10 students. Of these subjects the females and males were equal. All subjects were students at undergraduate level and had years of computer experiences. In terms of interaction with novel system such as natural-language dialogue systems, subjects' experiences were commonly inadequate. These subjects showed strong interest in experiencing novel technologies and systems. The advantage of this was each group could generate averagely consistent interactions. And that also avoided risks of individual subject's increasing familiarity of system tasks.

It was noted that the selection of subjects did not cover a representative sample of the main population of users, and was therefore less meaningful statistically. On the other hand, these subjects met study's requirements that 1) subjects should have little experiences of spoken intelligent system interactions, 2) subjects should not know the truth of pseudo system and, 3) subjects should be active to explore system functions. Based on these reasons, study results should be taken account of as empirical suggestions (such as schema design principles) rather than statistical conclusions.

No subject was informed about the existence of evaluator and the pseudo system prior to the study. Measurements were applied to replace privacy-related information in video footage. For example, the personal information was confidentially kept in another document, through which the author could track back the original contents.

The study introduced an instructor due to it was envisaged that the presence of an expert instructor would be helpful to establish a trust between the pseudo system and subjects, leading to the fulfilling WoZ simulations' preconditions (Fraser and Gilbert 1991).

4.2.3 Study Procedure

The study started from instructor's brief introduction of the system. The introduction included system functions and a simple demonstration of spoken command. Before subjects used the system, they were given a list to train the system. The list consisted of several frequently used words and a set of open questions. Both the words and questions were extracted from frequently-used mundane conversations. The training list is included in **Appendix 4.1**. Subjects needed to say a word to the system and the system returned the text of the word. Speech recognition errors were required to be corrected during the training. After the training, subjects could ask the system questions. The system facilitated by the evaluator would answer these questions according to schemas. These schemas are included in **Appendix 4.2**. Each study, including system training and using, took approximately 30 minutes. And the video recording covered whole procedures from instructor's introduction to the end of study.

Video footages were transcribed manually into scripts. These examined and logged subjects' activities in accordance with time stamps in video sequences. The template of video scripts is included in

Appendix 4.3. These scripts were concerned with some privacy information such as subjects' names and personal preferences. The author changed privacy-specific information thus the scripts would not reveal such information. Due to the video camera was mounted on the ceiling subjects' face images thus were not captured.

Self-reporting captured system operation experiences from perspectives of evaluator. The validity of self-reporting depended on evaluator's '*knowledge of the relevant information, ability to recall it, and willingness to report it*' (Goldman et al. 2003). The evaluator in the study was a member of research team, and was familiar with the objectives the study had. As a result, the evaluator was experienced, reliable and motivated to report credible experiences of system operation. The contents of self-reporting included verbal descriptions of operation feelings and experiences.

In this study subjects were not asked for feedbacks. This was justified for two reasons. Firstly it was difficult to capture systematic feedbacks, as subjects interacted with the system simultaneously but they only experienced partial system functions. And secondly, recorded video footage provided rich data to reflect how subjects interacted with the system.

4.3 Study Results

The video footage collected 60 spoken dialogue examples across three studies. Throughout these videos all subjects were observed to have suitable computer knowledge to use the system. For instance they showed some experiences relevant with desktop computer operation, such like 'SYSTEM START' AND 'NEW USER'. Even though these were spoken commands, there was no noticeable barrier observed in the study. Similar data was captured when subjects were given tasks to train the system. Subjects intended to wait for the system's responses before they continued the training. It was also observed that subjects were likely to trial own words and questions. The video analysis showed a number of subjects made enjoyable expressions such as joyful laughs, even in some scene that the system made errors such as mistaking the word 'THUNDER' to 'HONDA'. Some high-level activities were also observed that subjects were highly motivated to have more trials of the system. In addition, it was noticed that not only could the subjects follow the training words, but also have short pauses in their natural speech, especially in long sentences, for instance the question of 'WILL I BE RICH WHEN I AM OLDER' would be intentionally divided into phrases and have different stresses.

Evaluator's description of system operation reported some facts. Recognising the subjects' training word was fast while the message-typing took more time. The average time of these training words took up to 2.0 seconds, and its performance became worse when using long messages. Recognising the subjects' random questions were not complex to the evaluator while anticipating the subjects' intentions and behaviours was not easy. Lastly some inconsistent operations were observed in the evaluator's operations with the general schema design. The evaluator generated some improvisational operation to the subjects' open questions in this stage. For instance different messages were presented when different subjects asked a same question.

No subject was aware of simulations provided by the pseudo system. This was justified for two reasons. The first, the system made some errors as if it was a preliminary intelligent system; and the second, the evaluator was located remotely, but this had drawbacks of unstable connections to the control panel, for example, frozen video streams and operation command delay were observed occasionally.

4.4 Analysis: Evaluator's Operation and Impact

The next four sections compare evaluator's operations in terms of operation's stability, predictability, appropriateness, and effectiveness. The analysis was summarised in the last section.

Stability

The stability in this study was defined as the extent to which the evaluator could carry out same operations with the schemas. Rigorous schema design offered strict operations leading to reliable operations, and this made the evaluator's judgement less improvisational. The rigorousness of schema design constrained the possibility of generating random operations. In this mode the evaluator played as a 'translator' who interpreted subjects' speech, matched with the schema and returned suitable operations. More importantly it reduced the evaluator's obligations of generating proper operations with unexpected activities. In contrast with general schema designs, the evaluator felt more obligations to providing suitable operations. On the other hand the rigorousness constrained the schema's guidance. The study observed a number of subjects' questions were not well addressed due to the limitation of schema.

The general schema design had opposite effects as observed. Although with this schema the evaluator was able to cope with some open questions, a number of improvisations were observed cross simple word recognitions and question answers. This poor operation consistency was justified for two reasons. Firstly, fewer references of operation were given to the evaluator for system operation, therefore major behaviours of evaluator were based on real-time responses. And secondly, the evaluator's individual preferences of operation had negative effects to maintain operation consistency. Without 'obligations' of strictly following schemas, the evaluator was given more flexibility to change operation styles and contents such as messages' tones.

The combined schema appeared more credible, due to that a number of planned operations were constrained by the schema, and some unexpected activities were facilitated by the evaluator's own judgement. The evaluator's report showed anticipated activities, for instance word trainings fitted better with rigorous schemas. In contrast, unplanned activities were not suitable for such schema due to the poor

coverage would generate severe improvisations, as described in the experiment using general schema. Unlike the general schema study, the combined schema study was observed to be the most reliable cross three experiments. It integrated low-level improvisations with unplanned activities as did in the general schema design study. The video analysis showed that subjects were more convinced when the system responded to unplanned questions in the manners of human-like intelligence.

Predictability

The predictability was defined as the level of advance preparations of operation gained on the basis of observation on subjects. In the rigorous schema study the evaluator reported that predictability was rarely used due to the strict matching-returning process. In this stage the evaluator's operations seemed not be necessary to request predictability. Subjects were observed that they would try alternative commands when the system could not respond as expected. In the general schema study the evaluator reported the heavy load of anticipation and operation. It was mentioned that, using general schema to operation activities was to some extent like human-human communication, in terms of understanding each other and making proper echoes. And in the combined study, the requirements of predictability of operation were observed to be constrained at a low level. In the part of activities planned in the schema, the operations were like those in rigorous schema study. In the unplanned part of activities, the limited anticipations presented low-level predictability.

Appropriateness

The appropriateness of operation aimed to reflect the extent to which the evaluator presented suitable operations according to the schema. In the rigorous schema study, the evaluator was given explicit rules indicating which participant activities should be matched with which schema rules. The unplanned activities were classified as unrecognised ones. In this stage, the evaluator had rigorous constrains to present operations other than those planned in the schema. In the general schema study, the evaluator reported a great deal of ambiguity of judgement. One participant activity could be interpreted in multiple ways with multiple operations. The video analysis of subjects' interactions also indicated that they used to receive general answers in longer time. In the combined schema study, the video analysis showed the fastest

system response to the subjects. In terms of the evaluator's self-reporting, some positive comments were found. These indicated that the rigorous schema rules provided suitable constrains to give appropriate operations, as did in the rigorous schema study. Furthermore, the flexible part allowed autonomous operations, which clearly helped the evaluator manipulate the operations.

Effectiveness

The effectiveness of operation was not the primary target in this study, as other chapters (see **Chapter 5** and **Chapter 6**) would further investigate the effectiveness's impact on operation consistency levels. However in this study it was adopted to gain preliminary understanding of how the schema change might affect the evaluator's operation effectiveness. In the rigorous schema study, the video analysis showed that subjects received fast system responses. Comparing the average system response time through three experiments, it was found that the rigorous schema study had a noticeable leading, followed by combined schema study and lastly the general schema study. Based on the evaluator's description of operation effectiveness, it was reported that the general schema study took extraordinary time to judge which and how operation should be given. In contrast, the rigorous schema study saved considerable time on the judgement.

Although other affective factors related to operation effectiveness such as control panel designs (as described in **Chapter 2**) were not considered in this section, the evidences drawn from the study had conservatively contributed to an understanding that the effectiveness of operation went up as the evaluator's time spent on judging activities was shortened.

Summary

Previous sections analysed how the three schemas affected the evaluator's operations in four aspects. The results were summarised as below.

	Rigorous Schema	General Schema	Combined Schema
Reliability	credible, can be reproduced, can generate consistent operations, less judgement obligations, narrow schema coverage of activities	not reliable, may not reappear, easy to be affected by evaluator's individual preferences, many improvisations wide coverage of activities strong activity judgement obligations	credible in rigorous part a few improvisations in the flexible part can generate most of consistent operations wide schema coverage of activities weak judgement obligations
Predictability	operations are predictable	hard to predict all activities	predictable within the schema's activities unpredictable in unplanned activities
Appropriateness	all operations are appropriate	various levels of operation appropriateness	most operations are appropriate according to the schema
Effectiveness	fast judgements fast operations of operation	slow judgements fast operations of operation	fast judgements fast operations of operation

***Table 4.1**: Summary comparison of observations in four aspects*

Although in some ways, the combined schema study was similar with other two studies to some extent, the results point to differences between the schema designs. The analysis indicated strong relationships between operation reliability and the rigorousness of schema design. Rigorous schema constrained the level of improvisation, and it reduced evaluator's obligation of making suitable activity judgements. However it also suggested that high level schema rigorousness may only be useful in a study with all actions anticipated. In terms of effectiveness of operation, the findings showed that the judgement of subjects' activities took most of operation time; neither did the schema interpretation nor the operations. Other constrains such as operation operations were improved by control panel redesign, while the judgement of operation was the bottleneck and difficult to improve. In other two aspects, the analysis did not extract strong evidence that

could make empirical suggestions. However one confirmation was made in the study that the improvisations were mostly generated from operations with general schema designs.

The findings reported in this section provided an empirical foundation for the rest work in **Chapter 5** to **Chapter 7**. Consequently new understanding towards traditional schema-evaluator relationships was proposed which aimed to clarify schema's direct influence on evaluator's operation consistency.

4.5 Analysis: Subject Activity and Impact

Last section (see **Section 4.4**) provided empirical comparisons between schemas and the evaluator's operations. In this section a detail comparison of subjects' reactions to the evaluator's subjects was presented. Since last section discussed the connection between the schema designs and the evaluator's operations, this section focused on the relationships between the evaluator's operations and the subjects' reactions. The comparison classified the analysis of subjects' reactions in three aspects, as follows. The analysis was summarised in the last section of discussion.

Subjects' Acceptance of System

The definition of subjects' acceptance to the system operations was mentioned in **Section 4.2.5**. Since little assessment of subjects' acceptance levels was proposed in previous WoZ studies, the research developed one from scratch based on an wide-spread technology acceptance model (TAM) (Davis 1989). This model provided a theoretical basis (perceived usefulness and perceived ease of use) to interpret users' degree of acceptance of a new system (Hong et al. 2011).

Subjects were monitored from a ceiling-mounted camera in the laboratory, thereby all interaction activities, particularly speech and body gestures, were observed from a top view. The findings indicated subjects had few barriers to use the system cross all experiments. During the training session, a number of activities were observed indicating that subjects used to apply desktop computer habits to the new system. Although a few system operation errors were observed, these errors mostly focused on unrecognised speech and used to be corrected via speech repeating. These evidences indicated a preliminary suggestion that the subjects could learn and use the spoken commands efficiently, which indicated the good level of ease of use.

However, since it was merely a simple natural language-based interactive system rather than a specific practical application, no hard evidence was gained from the study to reflect the exact level of usefulness of the application. Only empirical suggestions could be drawn that current intelligent system had high efficiency in conversing with humans using simple speech.

Subjects' Engagement of Interaction

This study attempted to understand subjects' engagement levels, thus to assess the relationship of evaluator's operations and subjects' involvement of interactions. By interpreting subjects' expressions and speech meanings in video scripts, study findings showed that subjects in the study expressed strong interests in using this system. All three experiments witnessed subjects asked more questions other than those listed in the training list. In addition, the subjects also showed interests in the system's capability of speech interpreting and answer producing. In the video analysis the instructor was asked by subjects five times about this, although system introductions were given at the beginning of the study.

However, this was only a partial success. These examples included limited subjects who experienced all system functions. Although rich data was captured from observations and analysis of subject interactions, better understanding of interaction engagement was expected from other subjects.

Subjects' Speech Utterances

The study attempted to collect subjects' speech utterances, thus to gain implications for natural-language system development. The statistics of subjects' speech utterances were made on the base of video scripts.

The statistic of speech utterances consisted of several aspects, including the average number of speech in each operation, the style (such as computer-like commands or natural languages) of these utterances, and the stress of utterance. Through the video analysis, 58 valid spoken utterances were extracted. Except the training words given by the training list, subjects' speech was observed to adopt two main types of utterances: half computer-like and fully human-like. Half-computer-like utterances such as 'SYSTEM ERROR, REPEAT' were often used to give command to the system, while the full-natural utterances were chiefly used to ask questions, or, asking for information. In addition, little evidence in video analysis was produced to suggest utterance differences between the evaluator's different operations.

Since this was not a longitudinal study, it did not assess whether and how these utterances evolved in long term use. However, the preliminary evidence provided some clues. The findings showed that subjects' utterances had some shadows of desktop computing manipulations, particularly in terms of giving

commands to control the system. In addition, when subjects were trying to gain answers or information from the system, the observation showed more human-human communication style utterances were used.

Summary

This results presented in the previous aspects illustrate an overall picture of subjects' interaction status. Although there were limitations in the study such as the samples and length, there was nevertheless interesting understanding gained on the base of the evidence.

An overall suggestion from the analysis showed that the subjects were not sensitively aware the changes of system operation. However, the highlight on acceptance suggested that current system was efficient in learnability, and such system would involve subjects' previous computer manipulation experiences which made the spoken system manipulation easier. In terms of engagement, the preliminary findings confirmed subjects' engagement, while the exact levels of engagement still required more evidence. And lastly the statistic of subjects' utterances drew some implications for spoken system design: subjects used different speech strategies to interact with the system. In terms of controlling the system they tended to use conventional computer-like commands, while in terms of asking for information subjects would prefer utterances in the manners of natural languages.

4.6 Analysis: Control Panel and Impact

This section reports implications of system design from the analysis of self-reporting and video scripts. **Section 4.4** and **Section 4.5** presented the analysis in the evaluator's operations and the subjects' reactions. However, the focus of this section was on the operation system design and the spoken intelligent system design. As a first step towards gaining the implications to operation system design, the section looks at the interactions between the evaluator and the operation system (or the control panel). Then, the section reviews subjects' use of spoken intelligent system and draws implications.

Control Panel Design

To some extent the evaluator's operations relied on the control panel, through which the system connected to the front-end application and presented messages. Thereby, the control panel was significant to the operation system design. In this study the findings relevant with the control panel were drawn on the base of evaluator's self-reporting. **Chapter 5** provides details of study focusing on the control panel's impact on operation consistency level.

In the evaluator's descriptions some comments related to the operation system were noticed, such as the slow typing, and the unstable connections. The statistics showed that typing a training word might take up to 2.0 seconds. The situation was worse when the evaluator was trying to compose a long message, which could take as long as 5.0 to 6.0 seconds. Although the findings did not show noticeable impatience of subjects, the long waiting was believed to harm the system usability. Additionally, the unstable connections happened occasionally during the study, which was caused by the bursting with data flow in the intranet. This seemed unavoidable when the study was lasting for a long period of a day.

Taking together these findings it was preliminarily suggested that the control panel should adopt some new elements that allowed efficient message output. In addition, the unstable intranet connection may also need further improvements, which will involve re-distributions of experiment components.

Spoken Intelligent System Design

Due to the system was designed as a conversational agent who provided simple natural language-based interactions; the findings had few suggestions to specific system utility design. Some empirical implications about spoken intelligent system design were drawn.

The findings showed that the subjects could learn to use the system by applying previous experiences. In addition, the system training process also helped the subjects to learn the system. Some manipulation commands were learnt from this process. In this stage, it is suggestive that presenting explicit commands, or vocabularies recognised by the system, may be helpful to improve the system such as the learnability and usability.

Summary

The data indicated both the operation system and the spoken intelligent system had some space to improve. **Chapter 5**, **Chapter 6**, and **Chapter 7** realised and assessed these improvements respectively.

The implications drawn from the operation system design were twofold. Firstly, suggestions were summarised to change the control panel's message typing mode. Current textbox had identified some drawbacks such as low effectiveness and high typing error rate. Select-box was an alternative choice, whereas it limited the flexibility of typing customised message. A preset message list might also help to address the issue, however, same limits existed. Another choice was to combine the textbox and message list, and gave different operation tasks to these, i.e. using message list to cope with planned activities and the textbox for those unplanned. **Chapter 5** designed and evaluated this type of combination. Secondly, the system connections were suggested to change. The data showed remote connections' stability was not guaranteed, although it helped to hide the wizard. To address the instability the system needed to consider the relocation of the wizards and the structure of system components, such as direct application-internal connections between the control panel and the application. **Chapter 5** composed and tested new types of connections.

In terms of spoken intelligent system design, some empirical implications contributed to the understanding. These implications included clear system vocabulary design, efficient training process for system learning, and explicit indication of system status.

4.7 Analysis: Study Strategy and Impact

The study provided much evidence to identify the schema design's impact on operation consistency levels. However, since the spoken intelligent system was an ongoing project and often had technical improvements, it appeared that a piece of new development would have a different impact on ongoing evaluator operations as well as subjects reactions (such as an improved control panel). A wide range of concerns were raised by the author relating to the overall study strategies. Furthermore, issues varied subtly between experiments. This section highlights some of these concerns that were relevant to subsequent work in the thesis. The concerns identified included:

Subject-related Concerns

Since the study had subjects in groups, it reduced subjects' average duration of system use. As a result, only a small number of subjects were given the chance to have comprehensive system use, which severely constrained the scale of samples. Thereby, this made contributions to the understanding that WoZ studies fitted better with single user mode in terms of sample coverage.

Data Collection Method Concerns

Other concerns were of a methodological nature, and were manifested in all experiments for subjects. These concerns were realised after the analysis of study data, as some study expectations were not satisfied well. Examples included difficulties in validating the findings drawn from the video analysis, and observing the longitudinal change of impact. Since subjects were in groups and their system use was crossed with each other's, thus it made the video analysis difficult to recover some system use. The concerns also suggested other useful data collection methods such as interviews.

Summary

Previous sections illustrated several study strategy problems that mainly involve subjects and data collection methods. Participant-related concerns were highlighted since these constrained the scale of samples, and

other method-related concerns were also highlighted. Such concerns suggested that the a large group of subjects could not guarantee the sample coverage as the WoZ study copes better with one single user at a time. Additionally, other useful methods are suggested.

4.8 Discussion of Study Findings

This section discusses this chapter's main findings and implications, and connects these to the work carried out in subsequent chapters. Firstly, this section highlights methodological issues which should be taken into account when interpreting experiment data and results. Secondly, the section moves on to discuss the schema designs' impact on the evaluator's operations, as well as subjects' reactions. Some empirical suggestions are presented to illustrate the contributions to the understanding. Thirdly the section considers implications for the work in subsequence chapters. And lastly, the section reviews whether the study achieved the objectives described in **Section 4.1** and summarises contributions.

Methodological Limitation

The wide variety of the evaluator's operations observed in the study highlighted the highly sensitive nature of schema designs. Note that this was only the case that adopted narrow participant groups and mostly they were students. It was envisaged that a broader study strategy would be expected to cover wider population of real users of the future domestic communication system. Due to the limitations of sample scale, it should admit that the results presented in this chapter were proposed to be indicative rather than statistically important over the whole user population.

Meanwhile, it was acknowledged that the study left some variables not further investigated, such as the control panel design. However it was argued that most of these factors had been separated as independent variables, and only left the schema designs as dependent variables to the evaluator's operation consistency levels. Furthermore, the average performance of subjects remained stable, although subjects varied cross three experiments.

Schema and Impact

Section 4.4 highlights that the operation effectiveness was affected by the speed and quality of the evaluator's judgement of operations, rather than the rigorousness levels of schema designs. As far as the author is aware, this is the first study revealing the relationships between operation effectiveness and

evaluator's judgement of operation, on the base of systematic investigations. Previous studies have noted variables related to operation training for wizards such as (Fraser and Gilbert 1991). However, the findings described in **Section 4.4** made clearer indications that one key point of maintaining high level operation consistency was to improve the evaluator's judgement of operations, which was more difficult to improve than the evaluator training and control panel designs.

The cross-schema data indicated that the level of schema designs' rigorousness constrains the evaluator's improvisation levels. It was argued that the rise of schema design rigorousness could lower the rate of improvisational operations made by the wizard. This argument was supported on the base of findings indicating that to some extent the rigorous schema design helped the evaluator judge subjects' activities easier. Clear rules of schema design could provide the evaluator understandable references, which took less time of the evaluator to fulfil the 'obligations' of selecting a suitable operation.

Limitations of highly rigorous schema design were also mentioned in **Section 4.4**; as such schema designs can only fit well with a study that has anticipated most participant activities. Previous studies mentioned this type of schema design as a tool to collecting data such as children's favourite body gestures for game control (Hoysniemi et al. 2004). This argument posed a cap over the levels of schema design rigorousness. Due to the varieties of subjects' background knowledge and vocabularies, comprehensive anticipation of their speech was challengeable (Fraser and Gilbert 1991). Thereby, high level rigorousness schema designs cannot fit well with interactive spoken systems like the study's system, and the combination of flexible schema designs should be integrated.

Section 4.5 attempted to explore subjects' behaviours in more detail to take account of the schema designs' impact. Little evidence was found in the study to reveal the impact of schema designs on subjects. However, implications were drawn and indicated some features of subjects' interactions with the spoken intelligent system. The efficiency of the system's learnability was observed. However due to this was only a general conversational system, the efficiency was less meaningful to other specific utilities.

The findings confirmed that the subjects applied previous experiences in the novel system. Previous studies noticed similar 'knowledge transferring' phenomenon between human-human communications and

interactions with conversational agents (Edlund et al. 2008) – they were common in speech use. However this study highlighted the 'experience transfer' that subjects recalled desktop computer operations and used these in spoken system. Furthermore, the study findings indicated that the subjects selected different utterances to achieve different tasks. Combined-computer-like speech used to be used for system manipulation, while natural languages used to be used for conversations. The significance of this indication was that future WoZ operation design should take account of speech differences – designing different styles of speech for different types of system manipulations.

Implications for the Following Research Work

The inconsistency problem of WoZ method was suggested to be a worthy research aim, as described in **Chapter 2**, but with little empirical support. It was argued that a study looking deep into operation-related factors could provide an empirical foundation for reliable WoZ use by highlighting: 1) evaluator's judgement was the bottleneck to improve the operation effectiveness, 2) improving the rigorousness of schema design could reduce the level of improvisational operations, and 3) high rigorous schema design could not guarantee high reliability.

Section 4.4 also highlights some issues including unknown impact of the control panel, experiment tasks, and the judgement of operation. Before the author investigated the further impact of the evaluator's judgement (referred as 'interpretation' in later chapters), it needed to first address other variables raised in this chapter such as the impact of the control panel design, and the impact of experimental tasks. Thereby, **Chapter 5** looks into these concerns and aims to find out how the variables affect these and what strategies are needed to address the inconsistency caused by these. On the base of understanding of these variables, **Chapter 6** investigates the impact of judgement of operation.

Section 4.5 indicates observations on subjects' interactions and points to some features found in spoken intelligent system interaction. Most observations focused on two features, suggesting the cross-communication-medium experience transfers and the use of utterances for different interaction purposes. This suggests that future operation system, and spoken intelligent system, should take account of speech

command differences. Subsequence work over **Chapter 5** and **Chapter 6** were planned to adopt this suggestion for system design.

In addition, some minor issues such as instable connections and the top webcam's narrow view were planned to be improved in **Chapter 5**.

Fulfilling Research Objectives

The study reported in this chapter was argued to be successful in fulfilling its objectives. A wide range of findings were presented in a manner of empirical suggestions. In addition, a number of practical implications also contributed to the study to help the author gain a foundational understanding towards the operation design and future spoken intelligent system design. In summary key contributions made in the study include improved understanding towards the relationships between the schema designs and the evaluator's operations, as well as the system design implications described above.

Furthermore, the study provides an empirical foundation for later chapters. Firstly, it provided practical motivations for the improvement work in **Chapter 5**. Particularly, the implications of operation system design led to the change of system development, including the way to designing the control panel. The methodological implications, such as the data collection, motivated the need for the longitudinal study reported in **Chapter 6**.

Contribution

The following methodological and pragmatic contributions were offered in this chapter:

- The study highlighted evaluator's operation consistency differences that were relevant with three schema designs. It showed that the combined schema design had relatively more consistent operations due to it applied strict constraints to these anticipated operations meanwhile provided general principles for unexpected interactions. Neither rigorous nor general schema design achieved same operation consistency.

- The study showed evaluator's operation effectiveness differences. The combined schema design had worse performance than the rigorous one, as the rigorous schema provided narrow but accurate guidance.

- The study gained empirical implications for reliable WoZ operation system design for natural-language spoken system manipulation. The result confirmed the necessity of filtering evaluator's speech typing, thus to make consistent output texts.

4.9 Conclusion

The chapter describes a series of experiments to compare the impact of different schemas on the evaluator's operations and the subjects' reactions. The findings showed that the rigorousness levels of schema design can help to constrain the improvisation levels in the evaluator's operations; but highly rigorous schema can only be applied with a study that has comprehensive anticipations of subjects' activities. The cross-schema rigorousness levels study also indicated that the most difficult factor related to operation effectiveness was the evaluator's judgement of operation, rather than other factors such as control panels and operation trainings. In addition, in terms of subjects' reactions the study suggested a finding that subjects using spoken intelligent systems used to adopt different styles of utterances to accomplish the interactions.

Some other considerations were also taken into account in the chapter, such as study strategies and methodological issues. On the base of these findings and implications, the chapter has produced a explicit guide to subsequent work in this thesis.

Chapter 5

Control Panel, Subject Activity Interpretation and the Impact in WoZ Study

5.1 Introduction of Study Objectives and Chapter Structure

This chapter reports on two groups of Wizard-of-Oz (WoZ) studies which focused respectively on variables of control panel design and subject interpretation. These variables, as suggested in **Chapter 4**, were claimed to have strong dependencies with evaluator's operation consistency levels.

A description was given in literature review which noted the importance of these variables and their impact on the reliability and validity of WoZ. The work in this chapter made progress in understanding these variables and their impact over previous studies. Firstly, few WoZ studies systematically evaluated these variables. Secondly, as well as being one of the few empirical studies of control panel and subject interpretation, this work was the only study that investigated these variables and impact within the context of domestic communication. The work reported in this chapter was partially reported in other publications with different emphasise in (Li and Bonner 2011).

5.1.1 Study Objectives

The design work in this chapter had two high-level objectives, each relating to a limitation of WoZ operation system development described in **Chapter 3**:

- To propose a control panel design that could flexibly support the operations of high-level intelligent

systems – **Chapter 3** criticised previous control panel designs for the lack of such support. To avoid similar criticisms the author set out to ground design efforts in the findings summarised in **Section 4.8**.

- To implement design with an aim of performing natural and intelligent interactions – This was concerned with a second limitation of previous systems that in most cases the interactions supported by the intelligent system were rarely designed natural enough. Therefore, the intelligent systems needed to be natural and smart to maintain a high level sense of intelligence as future interactive devices.

5.1.2 Chapter Structure

This chapter is structured as follows. The first step of the work was to develop new system applications which differentiated from the system components in **Chapter 4**. **Section 5.2** detailed the incremental design. Following that, **Section 5.3** reported a case study in which iterative studies were carried out to investigate the variables relevant with control panel design. **Section 5.4** presented another case study that focused on evaluator's interpretation to subject activities. Further discussion of study results and findings were presented in **Section 5.5**, and that was followed by the final summary in **Section 5.6**.

5.2 Approach for WoZ System Development

This section describes incremental designs of system application. This was motivated by two technical challenges described in **Section 3.3**. These designs, as explained in **Section 5.1.2**, were based on pragmatic adjustments to previous system components.

The design work was deliberately incremental, and aimed at extending rather than overthrowing previous system designs. Although a number of novel techniques were described previously (see **section 2.2.2**), radical design for domestic communication does not necessarily entail a strong contribution to HCI knowledge, particularly before pragmatic assessments being carried out. This was backed up by researchers (Carroll 2000; Newman and Lamming 1995). Therefore in the work of this thesis, the adoption of incremental design progress was not unreasonable on the above grounding.

Firstly, the types of system applications were taken in account. In **Chapter 4** a natural-language dialogue was designed and implemented, and it was well accepted by subjects. But the use of the dialogue was limited due to insufficiency of considerations on practical requirements. Applications for domestic communication should be frequently used, close to everyday lives and broad enough to cover wide mundane routines. After a short survey of potential devices for domestic communication, two everyday products – the calendar and television – were selected, and improved to fit with the study contexts. Meanwhile the dialogue was remained to support natural language interactions.

Secondly, the distribution of system applications was considered. In **Chapter 4** the connections between evaluator and system were achieved via intranet, which was later demonstrated to be not very reliable for remote system operations. In this study direct connections were adopted instead, and the front-end applications and background control panels were integrated into one united programme, thus to avoid any extra connection interfaces.

And finally, the application displays were also changed from the wall to the coffee table in the laboratory, since the coffee table seemed to be a site more likely to hold conversations. Furthermore, the application interfaces were displayed on the coffee table by a ceiling-mounted projector, which helped the programmes

to detect physical objects easier. The main object used on the table was a small paper cube with numbers on its surfaces. The cube adoption was based on two main reasons. One was that the programmes could recognise the numbered cube accurately, and the other was that the physical cube was likely to be extent to other mundane objects which helped to enhance the sense of ubiquitous computing and ambient intelligence.

Additional design details were described in respective case studies, such as in **Section 5.3.1** and **Section 5.4.1**.

5.3 Study 2 – Consistency of Control Panel Operation

This study looked into how different control panel designs affected evaluator's system operation in natural-language dialogue system. It aimed to identify variables that were responsible for inconsistent system operation. Based on these variables, the study summarised several principles to guide reliable control panel operation.

5.3.1 System Development and Design of Proposed Studies

The study was based on system applications that were developed to understand domestic communications. The focal variable in the study was the control panel and its impact on the evaluator's system facilitation consistency, which was identified as an important matter of the reliability and validity of WoZ. Table 5.1 gives a summary of key study components as below, and the following sections describe in more detail these components.

Study objectives	To understand how different control panels affected the consistency of system facilitation, and based on this To suggest measurements to address the inconsistency of system facilitation
WoZ system for study	Cube-based multimedia manipulation system Natural language-based calendar, and natural language-based dialogue system
Study variables	Different control panels
Study method	Task accomplishing
Data collection method	Video analysis, interview, and self-reporting

***Table 5.1**: Summary of study 2's key components*

5.3.1.1 System Development

The work of development was based on C++, it integrated a database to organise calendar appointments. The work of development as well involved the pattern recognitions supported by an open source library called ARToolkit. The library enabled accurate pattern recognition that could be trained in advance. To make patterns more semantic the cube surfaces were printed with numbers (see **Figure 5.1**).

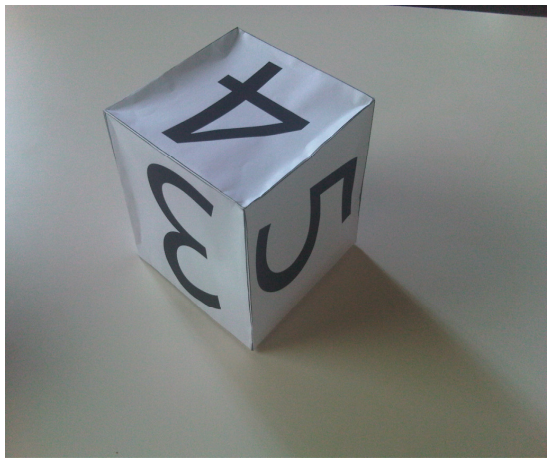

Figure 5.1*: The cube with numbers on its surfaces*

Based on the design in **Section 5.2** the new front-end system interfaces were designed to fit with the domestic contexts, and integrated with coffee tables (see **Figure 5.2, right**). Both the calendar and multimedia application were displayed on the coffee table, via a ceiling-mounted projector (see **Figure 5.2, left**). The cube as a controller was as well placed on the table. Flips and movements of top surface were associated with functions such as video programme change.

Figure 5.2*: The projector (left) rendered interfaces on the coffee table (right)*

A compact control panel was developed firstly (see **Figure 5.3**). The control panel integrated multiple functions in an all-in-one panel. Some functions used in the previous control panel designs were remained such as the IP and ports connections. The intention was to observe the evaluator's use of control panel from functions that might be useless for the operations. New features were also added to the control panel, such as calendar's date picker. These features were designed for the evaluator to facilitate the calendar and multimedia applications.

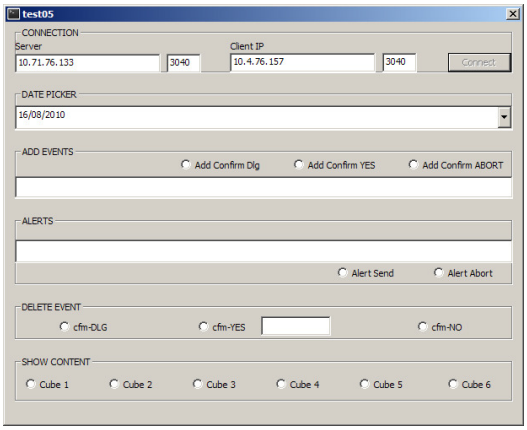

Figure 5.3*: Compact control panel design*

Alternative control panel designs were also presented. The study proposed another control panel designed with less operational functions but more layout flexibility, and more accessible interface elements (see

Figure 5.4). This design allowed the evaluator to re-arrange the control panel layouts, but with less optional functions.

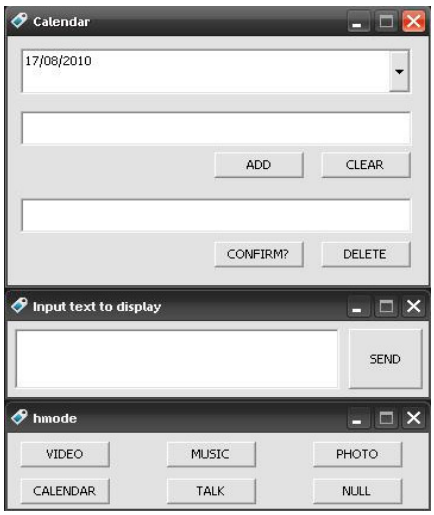

Figure 5.4*: Second control panel design*

A third control panel design was presented based on the second one. The design was on the grounding of a hypothesis that preset functions for anticipated subject activities, such as system's responding speech, could potentially improve the efficiency of control panel use, as the evaluator did not need to follow full operations and thus save time. Only the natural language dialogue and multimedia management were involved with the new design, as **Figure 5.5** illustrates.

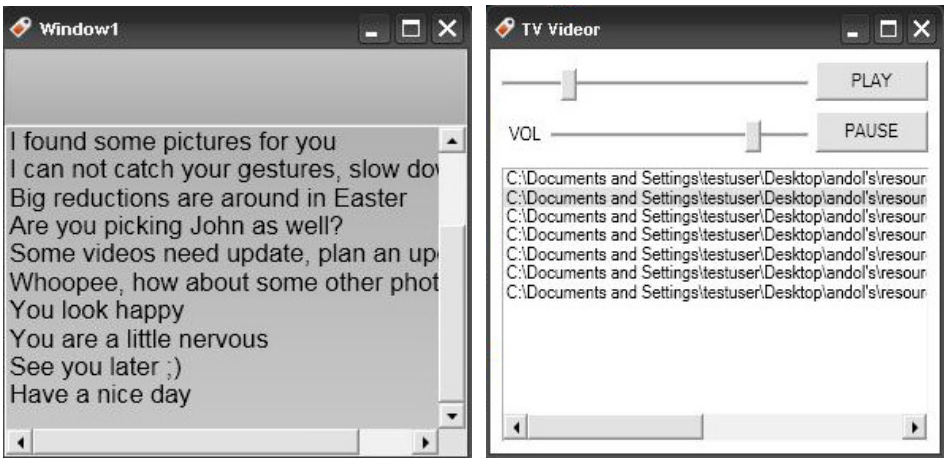

Figure 5.5*: The dialogue (left) and multimedia control (right) control panel design with preset messages*

5.3.1.2 Study Variables and System Components

Study Variables

To enable multimodal interaction with subjects the evaluator was given schemas that consisted of both rigorous and flexible scripts. This was based on the understanding in last study, and to guide evaluator to facilitate a wide range of subject activities using these control panels.

To assess the impact of control panels the study aimed to isolate the schema, wizard, tasks and subjects independently, thus to compare the operation inconsistency related to control panels. The study required an operator who was very well trained and familiar with all versions of control panel designs. Motivated by this, the study employed the control panel designer as the wizard. Meanwhile, the study employed two respective participants who were heavy users of the calendar and multimedia devices in the home and work place. A strict task list was used with experimental scenarios (see scenarios in **Appendix 5.1**). The benefits of using scenarios in evaluator-participatory studies included controllable flexibility and constraints to interactions. These tasks were compulsory but could be accomplished in random orders. That was used for a reason that the evaluator could have unpredictable user interactions despite the evaluator was familiar with the tasks.

A paragraph was illustrated to demonstrate the structure of study, as shown in **Figure 5.6**.

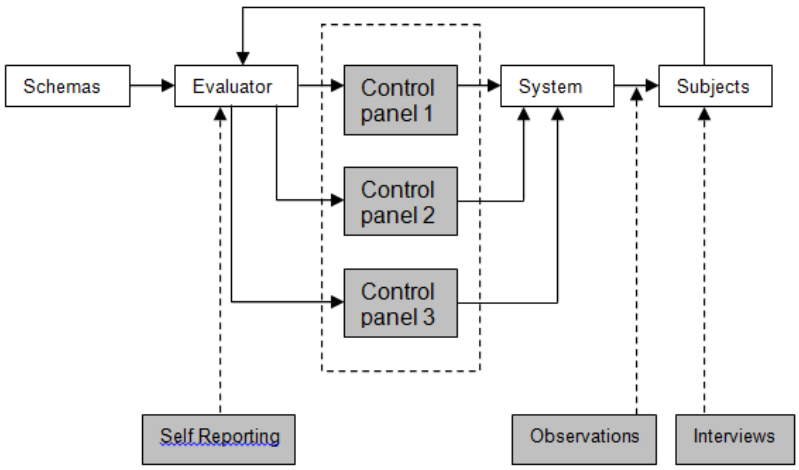

Figure 5.6*: Study variables*

Data collection method

Previous data collection methods – self-reporting and video recording – remained in this study, as these methods captured rich materials for study analysis. However, since last study had groups of subjects the interview method was not applicable. As this study employed fewer subjects, semi-formal interview was used to collect subjects' feelings and experiences of system interaction. This data could be useful to support video analysis, due to subjects' comments validated researcher's interpretation of system interaction from subject's point of view.

The semi-formal interview was carried out by the instructor, who also provided observations on whole system interactions. The advantage of doing so was that the instructor could raise specific questions relevant with subjects' system interaction, as the instructor was familiar with study objective in advance. The data collected via the interview was in forms of video footages. These footages were transcribed into scripts afterwards.

The aim of this study was to investigate control panel design's impact on system operation consistency. It was essential to assess the operation difference across three control panel designs. In this regard, the

evaluation criteria used in last study were still applicable to assess evaluator's operation consistency in this study.

Subjects

As illustrated in the graph of **Figure 5.6**, this study set subjects as independent variables. To lower the risks of subject's individual biases on system interactions, the study employed three subjects who interacted respectively with three control panel studies. All subjects were constrained by study tasks and speech commands, their interactions were maintained at a steady level. Meanwhile, subjects also produced some low-level randomness of system interaction. In this regard, the evaluator faced unpredictable system interactions to facilitate.

5.3.1.3 Study Procedure

The study started from the instructor's introduction of system tasks that subjects needed to accomplish afterwards. The list of tasks can be found in **Appendix 5.2**. Meanwhile, a list of speech command formats was also introduced to subjects, who needed to use their speech in accordance with these formats. The list of speech command formats can be found in **Appendix 5.3**.

After the task learning and speech command format introduction, subjects were allowed to start these tasks. The video camera started to recording system interactions from then. During the task progress, the instructor provided little interfere in subject's interactions, unless the system crashed down or subjects were trapped in specific tasks. After all tasks were accomplished successfully, subjects needed to take a semi-formal interview that was hosted by the instructor. The interview was also video recorded, as subjects' comments were transcribed in scripts for analysis.

5.3.2 Study Results

The data generated by the studies was rich in terms of evaluator's operation and system interactions. But the recorded videos and interviews provided variable scripts for data analysis. Subjects' awareness of the

pseudo system and different control panels was negligible, thus leading to good validity of data. On the other hand, the analysis had to refer to the data of subjects' interactions to identify the differences caused by control panels. The studies provoked rich discussions about designing natural and intelligent systems. This was understandable due to subjects were interested in using such novel systems for domestic communication. The subjects provided rich feedbacks to system use within domestic communication contexts, but only in empirical and individual manners rather than in statistic ones.

Each study lasted around 30 minutes, during which the task and command learning as well as the post-study interviews were included. Three video clips were collected, and the durations varied from 20 to 25 minutes. The videos included subjects' interactions and their feedbacks on system performances. Evaluator's self-reporting presented verbal descriptions of control panel operations. This mainly included a range of dimensions that could reflect evaluator's use of control panels, such as navigation, terminology, design and layout, user control, and match with tasks (Lindgaard 1994).

5.3.3 Analysis: Difference in Evaluator's Operation

Next three sections focused on the difference of evaluator's operations, subjects' interactions, and based on which the impact on operation inconsistency levels. The analysis presented in these sections was summarised lastly.

The overall reliability of the control panels was high, as the operations given by evaluator were generally consistent in presenting system application functions. So were the robustness and stability of the control panels. But some variable differences of evaluator's control panel operations were noticed in the results, and reported below.

Stability

The level of stability that the evaluator had consistent control panel operations was basically high, due to the studies adopted strict scenarios and tasks. The focus of the inconsistency was on these incontrollable errors

that were made with the control panels by the wizard. These subtle errors were critical to highlight the differences across the control panel use.

In the first control panel study, some occasional mistakes made by the evaluator were observed. These mistakes were reported as the results of incorrect operations. For example, the evaluator picked incorrect dates or clicked unexpected radio buttons. Such interactions caused confusions for subjects but who used to repeat their commands to correct the system outputs. This is illustrated in the extract (1). Although these errors were incontrollable and unpredictable, these were likely be avoided due to these were mainly relevant with misunderstanding or misuse of control panel functions.

(1) The mistaken operations (underlined) and subjects' repeating commands

W – wizard, S – subject

1	*[02:26]*	*S:*	*Today*
2	*[02:27]*	*W:*	*(selecting date today – but selecting the date of Tomorrow by accident)*
3	*[02:42]*	*S:*	*(after a short looking around)* *add (the) appointment* *(after give the speech, looking at the table and waiting for system response)*
4	*[02:43]*	*S:*	*(Suddenly noticed the incorrect date) Today*
5	*[02:47]*	*S:*	*go swimming*
6	*[02:48]*	*W:*	*typing 'go swimming' in the text area and sent out*
7	*[02:48]*	*S:*	*confirm*
8	*[02:49]*	*W:*	*clicking the 'Add confirm dlg' instead of 'Adding confirm YES', so the system popped an alert window to ask for input confirmation*
9	*[02:50]*	*S:*	*(ye) confirm*

Incorrect operations were reported less in the study using the second control panel. The most noticeably inconsistent operations captured in this study were the speech typing errors. This type of errors happened frequently in two occasions. The first was in some commonly-used words; the second was in unrecognised speech. Since the control panel provided an open text area for message typing, the texts were sent without check. It was precisely the control panels that affected the accuracy of operations. In most of these cases subjects were observed to ignore the incorrect spellings, as illustrated in the extract (2). At this point this type of errors was negligible to operation inconsistency levels, however, little evidence was provided in this study to state the reason for subjects' tolerances to such system errors.

(2) The spelling errors and subjects' reactions

1	*[00:35]*	*S:*	*Friday (diary)*
2	*[00:36]*	*W:*	*(selecting the date,)*
			and displaying relevant appointments
3	*[00:35-00:55]*		*(looking at the task list for a short while)*
4	*[00:56-01:40]*	*S:*	*Saturday, add appointment, 9am,*
			pick up friends to go to [*] park
5	*[00:59]*	*W:*	*((cannot recognise the word [*] park))*
			selecting the date 'Saturday',
			typing new appointment '9am,
			picking up friends to go to'
6	*[01:45]*	*S:*	*confirm*

In the third study similar typing errors were observed as well, especially in the responses that used long messages – including subjects' long appointment contents and evaluator's long response messages. In that case the evaluator typed the messages within inconsistent orders and expression styles. The extract (3) demonstrates a scene like that. In addition, since the third control panel was integrated with preset

messages (see **Figure 5.5**), it was understandable that fewer errors like that were found. But there was no big decrease of incorrect long message typing in operations.

(3) The inconsistent long sentence typing by the wizard

1	*[10:32]*	*S:*	*tomorrow (waiting for system's response)*
2	*[10:34]*	*W:*	*selecting the date of tomorrow in the calendar*
3	*[10:39]*	*S:*	*13th April, (er) pick Tom, from university, at 5:30*
4	*[10:45]*	*W:*	*starting to type the message*
			'13th April, 5:13, picking Tom, from university'
5	*[10:56]*	*S:*	*(no) five-thirty*
6	*[10:58]*	*W:*	*typing the right time in new appointment '5:30*
			13th April, picking Tom, from university'
7	*[11:08]*	*S:*	*OK*

Similar typing errors and inconsistent messages compositions were mainly found in the use of calendar and natural dialogue applications. In contrast, the multimedia manipulations with the cube presented highly consistent operations. But the cube numbers were not consistently facilitated with the fixed video programmes. Little evidence was provided across the three modules for multimedia control to prove obvious differences between the control panels. But from evaluator's point of view, the third control panel provided better user experiences due to that design could facilitate almost all video interactions by a single click. There were few significant differences between multimedia manipulations, but the 'one-click' advantage might indicate a new principle of control panel design for other applications.

Predictability

The extent to which the evaluator was able to anticipate subjects' actions and supported advance operations was reported to be different across the control panels in the studies. As the questions described in **Section 2.3.2** and **Section 3.2**, the focus of the studies was on two concerns: (1) how the control panels should be

designed functionally, and (2) how the control panel should be designed to enable constraints to control panel operations. Three features were incrementally designed in the control panels, including function complexity, operation constraints and organisation flexibility (*the level of adaptation to which the evaluator could customise the control panel to fit with operation preferences, (Lindgaard 1994)*). The operation differences caused by that were analysed according to the dimensions described previously.

In the first study the control panel provided high function complexity, low operation constraints and low organisation flexibility. Evaluator's self-reporting stated that the radio buttons provided good navigations to move around the control panel between modules, but they had some obscure abbreviated labels that could easily cause incorrect operations, like the example demonstrated in extract (1) line 8. The self-reporting also indicated that extra efforts were needed to adapt to the control panel design – especially the layouts of operation functions. Even the evaluator was meanwhile the control panel designer, the adoption to the function layouts was still sensible. As a result, the sense of being in control was low in the beginning of control panel operation, but the match with tasks was generally high due to the control panel was specifically designed.

The second control panel design divided the control panel into three independent modules, and combined some functions. This improved organisation flexibility by layout customisations; meanwhile it reduced as well the function complexity. The most differences stated in the self-report focused on the enhanced sense of user control that might lead to higher operation effectiveness, but relevant evidence of that was rarely provided in this study.

The third control panel, however, did not improve the predictability very much, although a number of preset messages were integrated. Both the video analysis of system response speed and the evaluator's self-report showed similar prediction levels. The similar results were justified for two possible reasons. First was that most operations were formatted, in other words the strict tasks produced predictable subject interactions. Second was that the preset messages had a narrow coverage that only fitted with a small number of interactions. An exception was that, the improvement of unrecognised speech operations was reported due to the preset message could deal well with the unrecognised speech.

Appropriateness

Since the focus of the studies was on control panel operations and their impact on evaluator operations, the subjects' tasks were based on scenarios and the evaluator's operations were limited in specific tasks. Although low-level improvisations were observed occasionally in the study, these were seen negligible. Thereby, the appropriateness of evaluator's operations was not reasonably applicable to assess the control panels' impact on operation inconsistency levels.

Effectiveness

So far the evidence had shown the efficiency of control panel operations in responding to subjects' speech and cube actions, but some sensitive effectiveness differences were concerned. The time for operation judgement was particularly highlighted.

In the first study the evaluator's response speed was generally high; most of control panel operations were accomplished in less than one second in the responding speed statistics. The distribution of the response time, however, was reported with slight differences. Comparing to the time spent on carrying out the operation, the analysis of video scripts showed that more time was spent on selecting proper control panel functions. The evaluator used to take far more time before the practical operations, as demonstrated in the extract (4). Compared to explicit subject inputs, messed speech commands could take much longer for the evaluator to select proper function to response. Although the extract (4) could not reflect how the operation time was exactly distributed around the operations, based on the evaluator's self-reporting the extra length of evaluator operation was justified for reasons that: (1) an operation that used two functions distributed in distance in the control panel could take more time to locate the functions such as date selecting functions and appointment deleting functions, and (2) the evaluator needed to identify the suitable function amongst a cluster of elements.

(4) The time distribution between function selecting and acting

1	*[00:20]*	*S:*	*meeting*
2	*[00:21-00:28]*	*W:*	*(firstly needed to check if it is today,)*

			secondly needed to catch up the appoint
			content typing
3	*[00:29]*	*S:*	*confirm*
4	*[00:30-00:35]*	*W:*	*inserting the contents into the database*
5	*[00:36]*	*S:*	*delete appointment*
6	*[00:37-00:41]*	*W:*	*(waiting for subjects' selection of*
			appointment to delete)
7	*[00:42]*	*S:*	*today*
8	*[00:43-00:47]*	*W:*	*(waiting for the confirmation of*
			date of appointment delete)
9	*[00:48]*	*S:*	*no*
10	*[00:55]*	*W:*	*cancelling the appointment deleting operation,*
			going to prepare the next possible operations

By adding flexibility in control panel designs such as the layouts, the second study showed shorter durations of operations. There were other reasons for this reduction. The control panel presented less function options, and this might make the evaluator easier to select proper functions. The control panel meanwhile proposed more constraints of functions, through which the evaluator was given less flexibility for improvisations. Therefore to facilitate same tasks the evaluator could save a little time than he did in the first study. The extract (5) was presented for the comparison to the extract (4)'s operation durations.

(5) The durations for the operations of appointment adding and deleting

1	*[02:44]*	*S:*	*add appointment, 9am,*
			for half an hour with John, confirm
2	*[02:45-02:53]*	*W:*	*selecting the date of today (no date specified*
			by subjects), typing the appointment contents,
			and confirming the input
3	*[03:21-03:49]*	*S:*	*delete appointment, Saturday, 3pm,*

			at [Boundry] Park*
4	*[03:50]*	*W:*	*clicking and outputting the message 'speech not recognised, please try again'*
5	*[04:09-04:14]*	*S:*	*the [* Boundry] Park*
6	*[04:17]*	*W:*	*to delete the appointment specified by the subject*

The last study, however, did not show an increase of operation effectiveness on the base of the second control panel, in contrast the effectiveness had slight decrease. The only exception was the responses to unrecognised speech – by clicking the unrecognised message, as demonstrated in the extract (5) line 4. The evaluator's self-reporting indicated some reasons. First was that the evaluator could not memorise all messages and thus the looking-up time was long, and second was that the evaluator used to check through the message list before carrying out operations to see if there was shortcut operations. The dwelling operation effectiveness was as well reflected in the video analysis. Although some fluctuations of response time were reported, the average operation durations were generally equal as that in the second study.

5.3.4 Analysis: Difference in Subjects' Reaction

The most common interactions of subject consisted of two types of responses across the three studies. In the forms of natural language interactions, the organisation of these interactions had the following characters.

Subjects' speech utterances, except the formatted speech commands, were reported with noticeable differences across the studies. Three main dimensions were considered important factors contributing to the subjects' reaction differences, including the accuracy of system responses, the speed of system processes, and the sense of intelligence presented by system operations. All of these were direct dimensions that the studies could reliably assess.

In the first study most of subjects' speech utterances were observed in means of computer-like. Separated words were commonly used in the appointment content inputs, but far less natural and long messages were given to system interactions. While in the second study the speech utterances became more natural, long

appointment inputs were given to the system but in slow speaking. In contrast in the third study such long speech inputs were also reported, but short and separated words were used when the system prompted the 'unrecognised' message.

These utterance differences reported across the studies might be interpreted for two main reasons. The first was that the speed of system responses might influence the flow of speech communication, which accumulated the sense of intelligence of natural interactions. It was commented by subjects in their interviews that, maintaining a fluent interaction flow that the system received correct speech and returned accurate and rapid responses made them put more attention to the interaction contents rather than the system itself. While disruptive interactions, for example, caused by unrecognised speech, could have more distractions. The second was that the difference of system response accuracy could influence subjects' tolerances of system errors and willingness of repeating the speech. Some errors made during the operations seemed more likely to be corrected than others, like between the extract (2) and (3). In some cases the subjects were observed to use another interaction ways rather than repeating the same speech slowly as requested. The interviews provided a little evidence about the reason for such different subject reactions. One reason was that the subjects intent to interact with the system in a means of machine-understandable manipulations.

From the side of subjects' interactions, these were more likely to be used to validate the understanding of evaluator's control panel operations, rather than independently used for subject interaction assessments. On the other hand, the subjects' interaction differences observed from the studies as well contributed to the understanding towards the intelligent system design for domestic communication, especially in terms of system acceptance and usability issues.

5.3.5 Discussion of Study Findings

Given the analysis that was taken previously, this section identified the relationships between these control panel designs and operation inconsistency levels. Three main questions related to control panel design were addressed as follows.

- How control panels should be designed to reduce system operation errors? The control panels were considered an important factor causing these incontrollable operation errors. Other factors, such as the evaluator's personal experience, were negligible in this stage due to the evaluator's familiarisations were independently maintained. Throughout the three studies there were two main factors contributing to the decrease of operation errors, these included the function designs and layout flexibility. The studies had identified three main factors that might cause operation errors. These included the longer time to locate the function, the easier layouts to misuse, and the heavier operation load. Thereby, to reduce the operation errors the control panels needed to be designed with as compact functions for system operations, but with far more flexibility that could enable evaluator to shape a preferable operation way.

- How the control panels affect the operation effectiveness? The studies suggested two ways via which the control panels exerted influence on operation effectiveness. In the first way the control panels had impact on the evaluator's function selections. From the evaluator's self-reporting and the video analysis of subjects' interactions, the study results showed that the evaluator spent more time on selecting proper functions rather than on conducting the operations. This explained why the preset messages were not helpful in improving operation effectiveness as expected, but the second control panel, the improved and simpler version, had relatively higher effectiveness than the other two. In the second way, the studies showed that the function elements, such as text areas and message lists, exerted as well important impact on operation effectiveness. As suggested in **Section 5.3.4**, the 'one-click' advantage in intense operations could have significant influence on the effectiveness.

- How the control panels apply constraints to the improvisational operations? Little evidence was produced to support complex control panel designs, but a basic suggestion was abstracted from the results – the control panels designed with constraints to improvisational operations had better performances in operation consistency. Such constraints were incrementally applied to the control panels, thus the evaluator received decreasingly freedom of making improvisational operations. The pros of the constraints could be justified for two reasons. One was that the constrained function designs provided limited operation capabilities, which rarely allowed improvisational operations. The other was that the control panels provided more specific functions for task operations.

The studies as well contributed to understanding to intelligent system design for domestic communication. The studies provided empirical results that indicated the subjects' interaction states; the interviews as well produced useful information in terms of subjects' feelings to the pseudo intelligent system. The summary of these indications was twofold. On the one hand the studies collected a great deal of practical interaction speech which might be useful for future intelligent system design. That included the ways the subjects used the speech commands as well as natural languages. It also showed some pointers to the corporations of speech and physical object manipulations such as the cube in a preliminary ambient intelligence environment. The studies, on the other hand, indicated some essential requirements in such interactions. For example, the subjects were observed to look for useful system process indications for further interactions.

Some unexpected findings were collected in the studies, including the subjects using natural languages as system commands and the subjects sought supports from the instructor, but these were negligible due to these happened occasionally and were properly facilitated by the evaluator according to the schemas.

The limitations of the studies were worth to mention. The most criticisms would be given to the lack of longitudinal studies – this could contribute to the understanding of the control panel's impact change over the periods. Since WoZ used to be adopted for short term studies, the limitation did not severely harm the study's validity and reliability. The other tactic limitation was that the preset message's coverage of control panel operations was poor, despite that the improvement of message coverage would produce additional evidence to support the conclusions summarised above. And finally, the studies assumed the evaluator to follow strictly the schemas, although there was no severer improvisational operation reported across the studies. The assumption worked well with planned operation tasks, while in the future studies that adopted natural interactions such as natural speech for system manipulations, there were more considerations required to extend this narrow assumption. This would be specifically investigated in the next section.

5.4 Study 3 – Consistency in Subject Activity Interpretation

This study aimed to address one of the limitations described in **Section 3.2** – Evaluator's variable interpretation of subject activities and relevant impact on system operation consistency. In last study the tasks used for system interaction were strictly planned due to the study needed to produce consistent subject activity interpretations. That was also justified as an important reason for using speech command formats in last study, as evaluator's interpretation to subject speech had to be understandable and controllable. This study, however, extended this constraint to using natural speech for intelligent system interaction.

5.4.1 Review of System Development

The situation became more challengeable when incorporating previous system applications for natural language interaction. The system applications adopted in last studies required additional examinations of applicability to reflect how these applications were suitable for new study structures and natural speech interactions. Since the system applications in last studies had shown some features that were supportive to natural language interactions, it seemed efficient to develop new system applications on the base of these applications.

The developments of new system applications were based on the understanding of control panel design in study 2. Two implications were applied in this development, including the separated control panels and the dynamically-incremental present message list. This control panel design, as demonstrated in last study, helped evaluator maintain high operation effectiveness. The preset message list designed for fast and accurate speech responses were developed with a new function that enabled dynamic message collecting from evaluator's typing. For example, the evaluator typed a message which – if it was not existed in message list yet – would be added to the list automatically. This could reduce the evaluator's unfamiliarity of a long message list. Meanwhile, it enhanced the message list's coverage on subjects' speech interactions.

The overall aim of these enhancements was to ensure the evaluator could efficiently facilitate the operations and share more efforts in subject action interpretations. The system applications developed for this study were a hybrid of the previous applications, based on which the positive features were particularly strengthened in this study. As a result the new system applications comprised a calendar, a natural dialogue system and a multimedia controller.

The final system applications' control panels are illustrated in **Figure 5.7**.

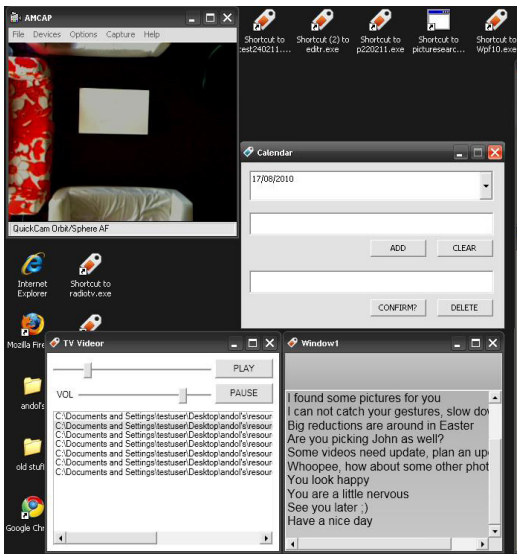

Figure 5.7*: Control panels for the system applications*

5.4.2 Method of Proposed Studies

Prior to the details of study objectives, WoZ system, study procedure and data capture method, a summary of key study components was presented in **Table 5.2** as below.

Study objectives	To investigate how the interpretation (and prediction) of subject activity poses impact on system facilitation consistency, and also

	To suggest measurements to address the variable interpretation
WoZ system for study	Cube-based multimedia manipulation system Natural language-based calendar, and natural language-based dialogue system
Study variables	Dependent variable – the evaluator's interpretations (see Figure 5.8) Independent variables – subjects' self interpretations and the experimenter's interpretation through video analysis
Study method	Task accomplishing based on certain rules
Data collection method	Video analysis, interview, and self-reporting

***Table 5.2**: Summary of study 3's key components*

5.4.2.1 System Development

The main improvement of the system in this study was the control panel design, as described in **Section 5.4.1**. The intention of reusing these system applications was to save the evaluator training. Meanwhile, as this study used improved control panel design, the influence relevant with evaluator's control panel operation was largely restricted.

5.4.2.2 Study Structure and Evaluation Variables

Study Variables

To highlight the difference of subject action interpretation, this study set evaluator and interpretations as dependent variables. Subject's activities were firstly interpreted by evaluator during the study, and these were taken another interpretation by researcher. Additionally subjects were also required to recall and explain their activities. These progresses were reflected in study structure in **Figure 5.8**. This study set other

variables such as the schema and control panel as independent variables. The intention of such experimental setting was to compare subject's activity interpretation from multiple sources.

Considerations were taken on how schema and control panel design produced reliable outputs for the interpretation. To achieve that, the understanding of schema design in **Chapter 4** was applied. The schemas in this study were designed with a careful combination of rigorous and general rules. The control panels were also designed according to the findings in last study.

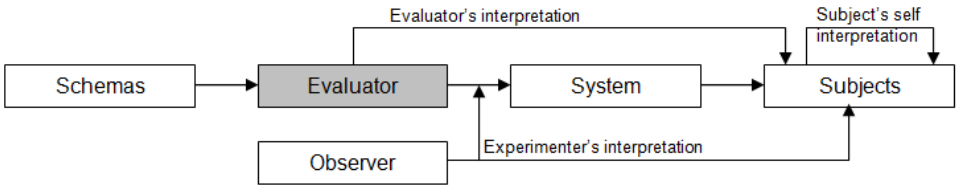

Figure 5.8*: Study variables*

The dynamic dependent variable was subjects' activities – these were variable and unpredictable but should be maintained at similar levels so that the studies could have referable assessments. To achieve that, the studies continued to use the scenarios and tasks to constrain subjects' interaction, while these were concise versions – their aim focused on giving flexible spaces for system interactions but constrained these at controllable levels. This generated natural interactions unpredictable to the wizard, but mainly stayed at a foreseeable range of task-related interactions.

Data Collection Methods

As described earlier three methods were adopted to collect study data. The video recording and analysis was the predominating method, as it was more objective compared to self-reporting and semi-formal interviews. These methods had different emphasises: the self-reporting focused on how subject's activities were understood; the video analysis focused on how the interpretation procedures were carried out, and; the interviews focused on how the interactions were motivated and what system responses were expected.

The four criteria described in **Chapter 3** were carried on to reflect the extent to which the evaluator's interpretations could affect the operation inconsistency levels. Amongst these criteria the stability and predictability of the evaluator's operations were particularly weighted in this study due to these two criteria had direct dependencies with subject activity interpretations.

Subjects

Three subjects were employed in the study, and each followed the same experiment scenarios and tasks. The subjects selected for this study were heavy users of interactive devices such as mobile phones, and more importantly they also played suitable roles in terms of domestic communication. Two of the subjects were male and the other was female. Their representative roles in the domestic communication were different. These roles, as validated in the later interviews, were not predominating in their life in the home but were well experienced.

The evaluator was remained throughout the study, as longitudinal interpretations could be captured across subjects within same scenarios and tasks. The evaluator who produced the interpretations in the studies was limited to an operator who was experienced at system operations, thus the focus would be on the interpretation with little interferes from the system operations. In the studies the system application designer filled the role, as the above requirements were well fulfilled.

5.4.2.3 Study Procedure

All three studies followed a same flow. The subjects were given the tasks within the scenarios' contexts, and they were encouraged to use their own ways to interact with the system to accomplish the tasks. The evaluator remained to facilitate the subjects' natural speech interactions throughout the studies and was not exposed until the end of the studies. The scenarios and tasks can be found in the **Appendix 5.4**, and the schemas for evaluator's operations can be found in **Appendix 5.5**. In all groups, the instructor – as described in last study – only interfered with subjects' interactions when the subjects were trapped unexpectedly. The whole procedures were video recorded.

The average durations of each study were approximate 30 minutes, excluding the pre-experiment learning and post-study subject interviews. All studies were video recorded from the starting of system interactions. The videos were manually transcribed into scripts according to the time stamps, thus to support further analysis.

5.4.3 Study Results

The data generated from all studies was divided into three categories. Firstly from the evaluator's self-reporting, the data provided an overview of activity interpretations – including the actions of receiving, understanding and judging of subject actions. The data was crucial to understand the mechanisms of how the judgement decisions were made. But the data source of evaluator's self-reporting was subjective, and thus not able to provide high reliability of evaluation data. Thereby, the data generated from all studies was divided into three general categories. The data from the evaluator's self-reporting provided an overview of the evaluator's interpretations to subject activities – including the actions of receiving, understanding and judging of subjects' interactions. This data was crucial to understand the mechanisms of how the judgement decisions were made. But the data from the self-reporting was too subjective to reflect the original meanings of subjects' activities.

Secondly, from another source of video analysis, the data provided probes to examine evaluator's interpretations' stability, predictability, appropriateness and effectiveness. Especially it validated the data from self-reporting by evaluating the direct observations on system actions.

And finally, the data from the subjects' interviews could contribute to the study analysis in terms of system usability, operation appropriateness levels and effectiveness. In addition the data could reflect the differences between the system interpretations and subjects' expectations that the system could understand their natural actions.

All data sources produced rich data that covered a wide range of variables in the system operations, as well as some unexpected data. The detail analysis of the study data was presented as follows.

5.4.4 Analysis: Evaluator's Operation Judgement and Impact

The aim of study analysis sections was twofold. Firstly **Section 5.4.4** focused on evaluator's interpretation activities, and based on which to identify the influence variables relevant with inconsistent operations. Secondly **Section 5.4.5** moved on to discuss the impact and measurements that concerned with the inconsistent operation caused by interpretation difference. Evidence and answers to the questions that were posed in the beginning of this study were provided in **Section 5.4.6**. In addition, an overview of understanding to the relationships between evaluator's interpretation and the operation inconsistency was presented.

Stability

The reliability of the evaluator's interpretations to subject activities was generally high on the one hand, since the major operations fell in the subjects' expectations to system intelligence, as described in subjects' interviews. On the other hand, some differences of operation reliability across the studies were observed in terms of speech recognition accuracy. The speech recognition accuracy was defined as the level of faithfulness to the schemas which were followed by the evaluator to generate accordant interpretations to similar subject activities. The interpretations were considered accurate when the evaluator's interpretations were consistent with those extracted from video analysis and subject interviews.

The similar subject activities, as referred in this case study, included the actions that had same system interaction motivations but in different speech utterances. These were primary objects that the evaluator facilitated with, as the video analysis has commonly observed various speech utterances with similar operation motivations. The difficulty of the utterance variety was that, this was not explicitly covered by the schemas, in other words, the evaluator needed to translate the operation goals from the utterances and match that with specific schema rules. The extract (6), (7) and (8) demonstrated an appointment-adding operation in different speech utterances across three studies. Similar phenomenon could be found as well in other system interactions, such as deleting appointments and playing videos.

(6)The utterances used in the first study for appointment-adding

W – Wizard, S – Subject

1	*[00:08]*	*S:*	*create an event*
2	*[00:09]*	*W:*	*(to interpret if the subject involved any specific date;) no date specified, then picking the date of today*
3	*[00:19]*	*S:*	*event for tomorrow and 2 o'clock*
4	*[00:20]*	*W:*	*((to interpret the specified date and time;)) re-picking the date, and preparing for event contents input*
5	*[00:23]*	*S:*	*I have a meeting*
6	*[00:24]*	*W:*	*inputting the event contents as 'I have a meeting'; (considering whether to present a proactive prompt related to 'meeting location?')*

(7) The utterances used in the second study for appointment-adding

1	*[02:20-02:22]*	*S:*	*am I free on Friday?*
2	*[02:23]*	*W:*	*to interpret the specific date of Friday; picking the date from the date picker; the calendar showing 'no events'; (considering to prompt 'adding new event?')*
3	*[02:44]*	*S:*	*book an appointment on Saturday, 3pm, for 90 minutes, at Boundary Park*
4	*[03:26]*	*W:*	*specifying the date of Saturday first; picking the date from the calendar; typing event contents '3pm, for 90 minutes, Boundary Park'*

5	*[03:49]*	*S:*	*that's it*

(8)The utterances used in the third study for appointment-adding

1	*[09:45]*	*S:*	*calendar please*
2	*[09:47]*	*W:*	*loading the calendar application*
3	*[10:39]*	*S:*	*(a long pause of waiting for calendar launching)*
			13th April, er- pick Tom, from university, at 5:30
4	*[10:45]*	*W:*	*to interpret the date of 13th April;*
			picking the date from the calendar;
			typing the contents of 'pick Tom from
			university at 5:30'; and confirming the input

In addition the analysis focused on the results that were relevant with the speech recognition accuracy, as which could have critical impact on the operation inconsistency levels. In self-reporting a few cases had been described that the evaluator was not able to interpret all possible natural speech, especially those related to abbreviations, terminologies and unfamiliar names of places such as the appointment location in the extract (7) line 4. This was understandable due to the knowledge differences between the evaluator and subjects. Furthermore, the recognition accuracy issue that concerned with subjects' motivation interpretations was also reported in the video analysis, which would be in depth analysed within the predictability criteria.

Predictability

The analysis generally divided the predictability of subjects' interactions into three levels, according to the explicitness of the interaction motivations. The intuitive speech i.e. the extract (6) line 1 implied clear motivations and operations, and was well guided by the schemas. That was frequently encountered in the studies, however, it used to be followed by the complicated speech that had clear interactions but equivocal operation commands such as the extract (7) line 1 and the extract (8) line 3. This type of interaction cost the evaluator extra efforts to translate the interaction goals into concrete operations, and then to match schemas with these operations. Furthermore, subjects' interactions could be with equivocal goals and operation

commands. In the chain of *interpreting goals – translating operations – matching schemas*, as described in the evaluator's self-reporting the goal interpretation was the most difficult process to maintain consistent interpretations. As the example demonstrated in the extract (6) line 6, the interpretations to such equivocal activities were highly contextual and isolated. The similar speech could bear various motivations and operations in different interaction scenarios.

Thereby, some predictive prompts were presented in the studies to cope with equivocal interaction intentions. In the extract (6) line 6 and extract (7) line 2 the evaluator had attempted to pop up proactive prompts (see definition in **Appendix 1.1**) based on context understanding. The extract (9) gave a positive example of making correct predictions of subjects' intentions of interactions. From the subject interviews this was highly accepted as a high sense of system intelligence. On the other hand, incorrect predictions could destroy such impressions, such as the extract (10). The subjects commented such negative predictions as 'do not think it (the system) gets what I mean'.

(9)The proactive prompts and system interactions

1	*[06:39]*	*W:*	*when adding cat-related appointments,*
			prompting a message of cat check –
			'book an annual cat check in Petvet? '
2	*[06:45]*	*S:*	*Oh- -, it's just remind that- -! Oh it's funny!*
			Coz it's about the cat put down here.
3	*[06:54]*	*S:*	*So interesting! So book this as an open diary*
4	*[07:00]*	*W:*	*adding the event 'cat check at PetCheck'*

(10)The incorrect predictive prompts and system interactions

1	*[10:26]*	*S:*	*want to book a cinema visit for tomorrow*
2	*[10:42]*	*S:*	*book, visit cinema, Friday, 11th November,*
			7:30 pm (a short pause afterwards)
3	*[11:36]*	*W:*	*adding the event; prompting a message 'fancy*

to watch some new cinema movie previews?'

4 [11:53] S: Er, no, cinema

Appropriateness

The appropriateness of evaluator's operation measured the consistency of how the evaluator's interpretations met the subjects' expectations of system understanding. In other words it was defined as the accordance of interpretations from different sources. In designing the study structures, it was intended to form multiple data sources, as a single source – either from the evaluator or the subjects – was difficult to comprehensively understand how appropriate the evaluator's interpretations fulfilled subjects' expectations.

Throughout all three studies the level of appropriateness of evaluator's interpretations was relatively high. The operations extracted from the video analysis were accordingly validated by asking the evaluator to recall the interpretations. Some interpretations commented by subjects would take another validation to the video analysis and evaluator's self-reporting. Thus the reliability of appropriateness was evidently high, and each interpretation was assessed from at least two data sources.

The most inappropriate interpretations reported in the studies were such that mentioned in the predictability evaluations. A very few cases were observed to have conflicts between data sources. Presenting incorrect predictive interpretations was one of these. In addition some others were also reported inappropriate when the evaluator thought the interpretations were appropriate but the subjects did not perceive that in the same way, as in the extract (11).

(11)The conflicts between the evaluator's interpretations and subjects' perceptions

1 [06:00] S: shaking the cube

2 [06:01] W: changing the video programmes to play

3 [06:02] S: shaking the cube again

4 [06:03] W: changing the video programmes again

5 [06:06] S: shaking the cube again, (even longer)

6 [06:07] W: changing the video programmes again

In the extract (11) line 1, the subjects in the post-study interviews explained the motivations of the first shaking as to change video programmes, which was interpreted by the evaluator correctly. The third shaking had changed its meanings as resetting the programme changes while the evaluator was not aware of that.

Effectiveness

Although the effectiveness of interpretations was significant to system operation, it was not the primary goal pursued in this study. But some critical implications related to responding speed and accuracy were extracted. From the last studies it was presumed that the evaluator's operation speed was crucial to the sense of system intelligence. Throughout the studies the results of subject interviews showed that the operation speed indeed mattered to system usability, but not to the sense of system intelligence. The extract (9) and (10) had very long response durations than those in the previous studies. But the subjects' feedbacks stated an increase of the sense of system intelligence. This unexpected result was justified for two reasons. Firstly that was understandable to subjects that an intelligent system would take long time for algorithm calculations. Secondly the sense of system intelligence rather relied on the capabilities of understanding subjects' interactions and returning proper interactions in turn.

Summary

In summary this section overviewed a series of variables that affected the evaluator's interpretations. Three levels of interpretations were presented in the section, along with the variables at respective levels, such as speech receiving, speech recognising, speech judging, motivation predicting, and schema matching. The quality of speech recognitions and motivation predictions, throughout the above analysis, were regarded as the primary variables that would cause inconsistent operations. But further analysis was still needed to understand the impact of these variables, and to draw measurements to address the impact.

5.4.5 Analysis: Subject Activity Interpretation and Impact

In the last section two variables were highlighted as the primary factors affecting the subject activity interpretations, including the speech recognitions and subject activity predictions. These two variables had been evidently reflected throughout the studies. But still there was insufficiency of understanding in terms of the impact. Thereby in this section the aim was to analyse what impact the variables exerted on operation inconsistency levels.

Speech Recognition and Impact

There was rich evidence to suggest that the speech recognition was the core of the interpretations to subject activities. It connected to the schema interpretations on the one hand, and on the other hand it produced direct guidance to control panel operations. In addition it was as well relevant to the predictions of subject activities, which was described in depth below.

Two types of speech recognition reported in the result analysis were considered harmful to the operations. Firstly, the inaccurate speech recognition would enlarge the gap between the evaluator's understanding and subject expectations, as demonstrated in the extract (11). That could lower the reliability of study results due to the interactions poorly reflected subjects' true intentions. The constraints posed by the schemas might be as well damaged when the evaluator's interpretations were based on misunderstanding.

Secondly, inconsistent speech recognition could have similar harms to the consistency of operations. Inconsistent interpretations, as commented in subjects' interviews, would lower the usability of the intelligent system by making subjects confused with unpredictable system responses.

Predictions of Subjects' Activities

The operations based on appropriate subject action predictions greatly enhanced the sense of system intelligence such as in the extract (9), but incorrect predictions could have opposite effects. The predictions were heavily weighted in subject activity interpretations due to their enormous usefulness to intelligent

system operations, but there were three main questions to take account of before applying these in practical operations.

Firstly, *when to make the predictions* – This related to the judgements of the time gaps for predictive operations. All predictive operations across the studies, including the extract (6) and (7), were presented shortly after indicative operations. This was supported for two reasons: it appeared natural to prompt relevant messages, and the prior interactions had produced useful contents.

Secondly, *how to make the predictions* – This related to the control panel's functions. Due to the limit the predictions were presented through text messages. These provided intuitive expressions for subject-system interactions, but were also criticised for their lack of naturalness.

And finally, *what predictions to make* – This related to the use of interpretations. The experimental practices suggested an acceptable way to deliver the predictions – extracting the key words and combining these with anticipated intentions. It was likely to be non-interruptive amongst the interactions, and tightly related to ongoing interactions.

5.4.6 Discussion of Study Findings

Admittedly, the overall interpretations given by the evaluator were accurate and consistent, as considerable operations were accordant between three data sources. But the focus of the studies was on a small number of variables that generated negative impacts on operation consistency. The impact of these variables had been discussed in last two sections, which were summarised in two points. Firstly the inaccurate and inconsistent speech recognition harmed the evaluation data reliability, and weakened the schema and control panel's constraints to improvisations; secondly the incorrect activity predictions massively decreased the sense of system intelligence, and could deepen evaluator’s interpretation misunderstanding.

Some measurements were implied in the discussion. Clear vocabularies were suggested to define the boundaries of system's speech recognition capabilities. This was also mentioned by the subjects due to they were not familiar with the level of intelligence and thus felt too general to interact with the system. Intensive

evaluator training was also suggested to improve the evaluator's ability of interpreting equivocal activities. It could not assume that, however, the improvement of specific interpretation skills from the training would overcome the differences of individual operation preferences. This problem would be more exaggerated when different evaluators were involved in a group of WoZ studies. Thereby at this point more endeavours were expected to clarify such individual differences and the impact. This was described in **Chapter 6**.

During the studies, a wide range of data related to natural speech utterances for domestic communication was collected. The studies reported in the **Chapter 4** and the first case study in this chapter were novel in terms of exploring reliable WoZ system constructions for consistent operations, the research described in this case study was regarded one of the few empirical studies of operation systems and evaluator behaviours carried out to date. To the best knowledge of the author, it was the only study that investigated cross-control panels and subject action interpretations.

But some methodological limitations were acknowledged, which affected the generality of the evaluation results. Firstly, like previous WoZ studies, the subjects were limitedly representative due to a small sample was adopted. In addition the subjects were all known to the author and instructor. Inspired by this, the next study will be scheduled with more subjects from different backgrounds.

5.5 Contribution

Three main contributions to the thesis were made in this chapter as follows:

- Understanding of control panel design variables and impact was provided in this study – This revealed that control panel's function and layout design could affect evaluator's operation consistency levels, by applying different levels of system operation constrains and guidance. The combination of limited function and flexible layout was claimed an effective way to maintain stable reliability and validity.

- Understanding of evaluator's interpretation of subject activities was provided in this study – This showed that most interpretation errors occurred in the judgement of subject's intention anticipation, as context-based subject intention prediction only worked with events planned in the calendar. Also, the understanding noted that the match was variable between subject interpretation and schemas, due to some subject activities were equivocal to match with specific schemas.

- Empirical understanding of smart device design for domestic communication study was provided – This included the requirements for much clearer system progress indication. Understanding of WoZ operation system development was also provided, including the default distribution of control panels.

Chapter 6

Multiple Evaluators and the Impact in WoZ Study

6.1 Introduction of Objectives and Chapter Structure

This chapter reports the implementation and evaluation of two trained evaluators' system operation difference. The difference of multiple evaluators' system operation, as suggested in findings in **Chapter 5**, was recognised as a crucial concern relevant with the reliability and validity of WoZ. In this chapter three empirical studies were proposed, through which the research aimed to fulfil two methodological objectives as follows.

- To understand the study management and control difference between two trained evaluators, including relevant variables between multiple evaluators and the impact of these variable.

- To find out solutions to improving the reliability and validity of WoZ in multi-evaluator studies, thus to support non-specialist evaluators' use of WoZ for domestic communication study.

Following sections described these two objectives in depth, along with pragmatic objectives of intelligent system design. The contributions made over previous work were also described, as well as the structures of this study.

6.1.1 Study Objectives

The findings in **Chapter 5** indicated the difficulties of addressing evaluator's background knowledge for consistent subject action interpretation. In this regard the difference would be more incontrollable when multiple evaluators were involved in WoZ study. This meant more variables would be engaged between different evaluators, and the difficulties would increase in terms of synchronising the operation between evaluators. From evaluator's point of view, such difference could be difficult to cope with, because some variables were unlikely to be synchronised such as personal feelings, interpretation misunderstanding and responding expressions. But from system designer's point of view, such differences could be possibly reduced to acceptable levels. Little evidence relevant with multiple evaluators' operation difference was provided in **Chapter 5**. Thus, more efforts were proposed in this chapter to support further understanding of relevant variables and impact.

Objective 1: To Identify Variables and Impact

The implementation and evaluation of multiple evaluators' operation were essential to the reliability and validity of WoZ, since an experienced but highly personal evaluator could not on its own constitute a reliable WoZ study. **Chapter 2** and **Chapter 3** noted that previous WoZ studies used to adopt one experienced evaluator and much of these studies was criticised for the insufficient considerations on evaluation reliability and validity, as using other evaluators might lead to inconsistent results. **Chapter 4** and **Chapter 5** highlighted the variables causing inconsistent operations, but that was still limited in one evaluator.

A start point of this study was to extend such limitations from one evaluator to multi-evaluators. The aspects compared in this study covered a broad range of variables from the schema interpretation to control panel operation and to subjects' action interpretation. The 'synchronisations' between two evaluators were a complicated mechanism, which required variables to be initialised equally. An incremental evaluation approach was taken to enable step-forward assessments of operations consistency between evaluators.

Chapter 5 concluded with studies cross several operation sessions, but with little evidence to the understanding of operation reliability and validity with multiple evaluators. Based on the understanding

gained from current evidence, it was not unreasonable to pursue a more in-depth evaluation with this objective. Two sub objectives were proposed, as follows.

- To observe the operation difference between two evaluators who received incremental training after each study.

- To assess the longitudinal operation differences across all groups of studies – This was to extract the understanding of overall effects of incremental training.

Both sub objectives were relevant with assessment of variables causing operation differences and further operations after further improvements. The specific aspects to be assessed included the follows:

- The constraints given by schemas – Whether the schema guidance would be consistently followed by evaluators.

- The practical operations related to the control panel – Whether and how the control panel contributing to the evaluators' operation difference.

- Evaluators' subject interpretation accuracy, appropriateness and consistency – Whether the evaluators having consistent knowledge and capabilities to understand subjects' motivations and interactions.

Objective 2: To Conclude Operation Strategy for Stable Reliability and Validity

As well as identifying the influence variables and impact, the studies also aimed to gain empirical understanding of multi-evaluators' operation strategy for reliable WoZ. The proposed studies offered good opportunities to gain such understanding. The literature review examined previous WoZ studies in terms of multimodal operations, and noted heavy dependencies with an experienced evaluator's individual capabilities. One exception was Salber and Coutaz (1993)'s study in which multiple evaluators were employed with different roles. But still, little attention was given to the reliability and validity of individual evaluators' operation when they were switched with others, or, if the speech evaluator was exchanged to the mouse evaluator, the operation reliability and validity would be a critical concern in that case. In the context

of this thesis, this work offered a chance to build on the exploratory studies carried out in **Chapter 4** and **Chapter 5**.

6.1.2 Chapter Structure

This chapter is organised as following. **Section 6.2** described the first group of studies; including study method, study results and data analysis. Based on the findings in **Section 6.2**, the second group of studies was presented in **Section 6.3**. In **Section 6.4** the third group of studies was presented. **Section 6.5** summarised all findings gained from previous studies, and made a longitudinal discussion of these findings. And finally **Section 6.6** concluded all works and understanding in this chapter.

6.2 Study 4 – Identifying Variables in Multiple Evaluators WoZ Studies

This study aimed to identify influence variables that generally affected multiple evaluators' system operation consistency. As well, the impact of these variables was also investigated. Particularly, this study highlighted the variables that were relevant with three aspects as schema interpretation, control panel operation and subject prediction, as the variables in these aspects were pragmatically evaluated in previous studies and provided empirical understanding of the reliability and validity of WoZ.

This study presented a comprehensive comparison of system operation between two evaluators (one less-experienced and one experienced). The comparison was achieved by collecting and analysing subjects' feedbacks, evaluators' experiences and overall system interactions. In addition two evaluators' difference in control panel preference and proactive system operation was also compared. The output of this study included the main influence variables, impact and implications for next study.

A summary of this study's key components was presented in **Table 6.1** as below.

Study objectives	*To observe individual evaluator's system facilitation* *To investigate what variables are responsible for the difference of system facilitation in multiple evaluators, and* *To identify in what aspects such difference can be addressed through practical measurements*
WoZ system for study	*Cube-based multimedia manipulation system* *Natural language-based calendar, and natural language-based dialogue system* *Web-based information retrieving system*
Study variables	*Dependent variable – individual evaluator's system facilitation and activities in the study (see Figure 6.4)* *Independent variables – WoZ system, user tasks*

	and other system configurations
Study method	*Task accomplishing based on certain rules*
Data collection method	*Video analysis, interview, and self-reporting*

***Table 6.1**: Summary of study 4's key components*

6.2.1 Method of Proposed Studies

6.2.1.1 System Development

This section justifies the rationales behind the developments of the system applications in the studies. The incremental application improvements were in line with the first objective described above, that of investigating the evaluators' operation differences with a high level intelligent system for domestic communication study. There were two main concerns considered in terms of the technical developments as following.

Firstly, an additional function was added to current system components. The intention was that the real-time information retrieving application could provide sufficient contents for system interactions. The new function was a speech-driven application that enabled real-time information retrieving from the Internet. The retrieved information comprised pictures and websites, both of which were based on keywords search via the search engine API of Bing. Parts of source codes of the API can be found in **Appendix 6.1**. When the keywords were captured by evaluators and typed in the text area in the control panel (see **Figure 6.1**), the function automatically started to search and retrieve most relevant pictures or websites from the search results. A check-box was designed for the evaluators to distinguish what type of information to display.

The requested pictures would be loaded and displayed on the table immediately (see **Figure 6.2**). The pictures displayed on the table had similar display layouts as those in the control panel. Therefore, the evaluators could monitor which picture was pointed by the subjects via the webcam mounted on the top, and make specific reactions such like zooming in and out of the pictures.

Figure 6.1*: The control panel for the new function*

Figure 6.2*: Displaying pictures on the coffee table*

Secondly, a set of control panels were confirmed to facilitate the applications. The control panel layouts were not fixed in previous studies, due to these studies aimed to investigate control panel variables. Based on the finding in study 2 the control panel followed a similar way of design. An overview of the final control panel design was illustrated in **Figure 6.3**.

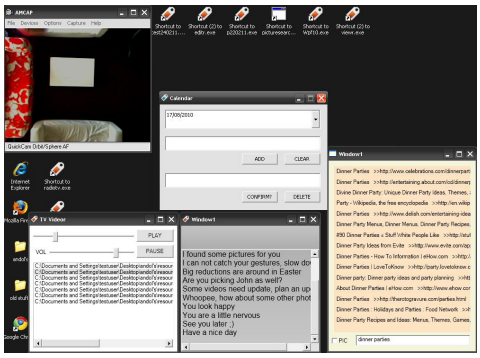

Figure 6.3*: Control panel design*

In addition some technical challenges of system application design were concerned. Evaluators' control panel operation areas were restricted, as evaluators might expose when they moved control panels and mouse pointer to front-end applications. Calibrations were needed to project the retrieved pictures on the table. The size of picture was specified to fit with the table surface, and the position of selected picture display was also specified. The synchronisation of two video cameras (one for evaluators and the other for subjects) were achieved through a noticeable starting operation.

6.2.1.2 Study Structure and Evaluation Variables

Study Variables

The overall study consisted of three incremental studies, and each study compared the operation difference between two evaluators (see **Figure 6.4**). To highlight the influence relevant with evaluators' operation, system components such as schemas, subject tasks, experiment scenarios and control panels were set as independent variables. The main variables throughout three studies were evaluators' individual interpretations and operations.

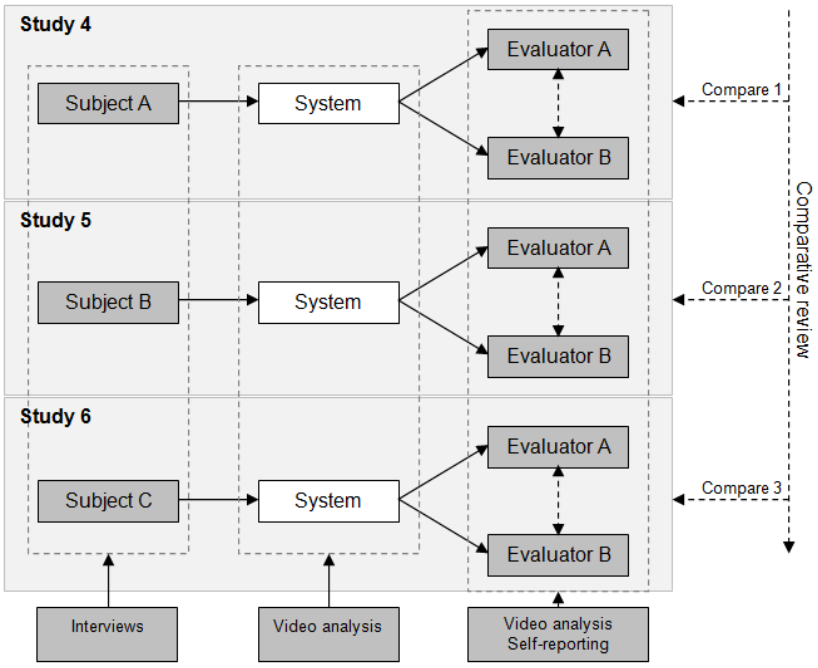

Figure 6.4*: Study variables*

Four comparisons were planned throughout the studies. As illustrated in **Figure 6.4,** in each study all system operations were compared between two evaluators, and finally a longitudinal comparison across all three groups of study was carried out lastly. Such structure of study setting was justified for two reasons as following.

- Firstly, three subjects were employed to ensure the reliability of study results. In **Chapter 5** the data generated from natural language interactions was variable, especially in terms of speech utterances and system error tolerances, for example. Using more subjects could broaden the interaction coverage, and improve the reliability of study results for the following comparisons. But due to the constraints of experiment scenarios and tasks, considerable overlaps of subjects' interactions were anticipated, as observed in **Chapter 5**, which made the large sample subjects less meaningful.

- Secondly, such study structures helped to highlight the incremental progresses of two evaluators' operation synchronisations. The first group of studies were based on previous understanding, and incremental improvements were made after the evaluations. The second group was on the base of these

improvements, and thus generating the improvements for the third group of studies. In such structure, each groups highlighted specific operation differences, and after careful improvements the longitudinal comparisons across these groups highlighted the tendencies of operation consistency. The incremental comparisons provided an overall picture of two evaluators' operation changes.

Data collection method

Video recordings of evaluators' operations were added in the studies, as suggested in **Chapter 5**. These videos were recorded to examine the influence of control panels. Thus, if the new evaluator could not use the control panels properly then the specific operations were not considered in the analysis. The video recordings of subjects remained focusing on their interactions with the system, by which to validate the evaluators' self-reporting and subjects' comments.

The semi-formal interviews were carried out by the instructor. The main themes of interview focused on subjects' feelings of the system performances in terms of the capabilities of recognition and the overall sense of intelligence. Comparisons were carried out between two of subjects' interviews.

In addition, the evaluators' self-reporting was important to the studies, to highlight their feelings of the operation system. These feelings were seemed as potential causes of operation differences.

To evaluate the levels of operation inconsistency, the criteria in four dimensions were adopted in the studies due to these criteria had been proved useful in assessing the operation qualities across previous studies.

Subjects

Three colleagues of the author were selected to take part in the studies, and all had different backgrounds in research. The selection of colleagues as subjects had two advantages. Firstly, it was hoped to leverage the existing trust between the subjects and the author, thus making the system more convincing and avoiding the ethical issues. Secondly, all subjects were technically sensitive in terms of their backgrounds, including artificial intelligence, computer vision and liquid engineering. An additional reason was that these subjects were easy to get more post-study data such as informal discussions.

On the other hand, two problems were concerned with the selection of subjects. Firstly, extra efforts were needed to control subjects' potential biases due to the familiarisation with the author. In this regard, the studies employed the instructor to interview the subjects. The instructor was not known to the subjects, and thus could neutrally encourage both positive and negative user comments. The data received from the studies showed that this had been achieved, as demonstrated in the following data analysing sections. Secondly, it was not the aim of the study to cover a comprehensive population of subjects, thus the three subjects in the studies were likely to generate different results based on their high sensitivity of system operations.

Two evaluators were involved in the studies, one had been employed in previous studies, and the other was a new evaluator who rarely had any experiences of facilitating an intelligent system. The employment of the new evaluator was in line with the first main objective described in **Section 6.1.1**.

The new evaluator was a research student who had backgrounds of computer visions, but had rare experiences of intelligent system operations. Extra trainings were given to the new evaluator before the formal operations. The intention was justified for two main reasons. Firstly, the training could help to investigate whether a novice evaluator could be trained as skilful as the experienced one. The experienced evaluator was seemed as the representatives of the evaluators who might be the designers or members of research group. Secondly, the trainings enabled similar familiarisation of system facilitates such as schemas and control panels, thus reducing the influence of system facilitates.

6.2.1.3 Study Procedure

All three groups of studies followed similar procedural flows. The subject was given a list of tasks to accomplish, within the contexts of the scenarios provided (see the tasks in **Appendix 6.2** and scenarios in **Appendix 6.3**). Few constraints were applied to the ways of accomplishing the tasks, and the subject could use natural languages freely to interact with the system, thus the prior system learning was not necessary. The subject was not aware of the existence of evaluators, nor of the switches of evaluators. This generated reliable results for the later comparisons. The instructor setting in the laboratory gave brief system function

introductions at the first stage of the studies. In the following interactions the instructor would not give hints or help the subject to accomplish the tasks. And in terms of the evaluators, both evaluators facilitated with the same subject, by following the schemas given before the studies (see schemas in **Appendix 6.4**).

The experienced evaluator firstly facilitated with the subject, and on the next day the new evaluator did same operations. The switch of evaluators was unknown to the subject. After the first study, the subject was interviewed about the feelings of the intelligent system, while after the second study the focus of interview was on the changes of the system operations. From the task learning to interviews, the whole procedures were video recorded, including the subject's interactions and the evaluators' control panel operations. These videos, together with evaluators' self-reporting and subject's interviews were transcribed into scripts for the further analysis.

6.2.2 Study Results

The average durations of each study lasted around 30 minutes, excluding the prior task learning and post-study interviews. This provided rich data in forms of verbal descriptions, video scripts and interview comments. Due to the subject's unawareness of the evaluators, the validity of the results was generally high.

The verbal descriptions consisted of three types of data, including the two evaluators' personal feedbacks of schema guidance, control panel use, and subject interpretations. In addition, other data such as complaints of poor monitoring video were also collected, but this collection was considered less important due to these facilitates were equally provided to both evaluators.

The data collected by the video recordings focused on two aspects of interactions. Two evaluators' operations were video recorded, as these were considered useful to validate the data from self-reporting. Meanwhile, these video recordings were essential to the analysis of evaluators' operation preferences, such as the control panel layouts, preset command selections and text typing styles. The other aspect of interactions, the subject's interactions with the intelligent system, was considered important to analyse the subject's action differences. Both aspects of video recordings produced objective data for further operation difference analysis.

The data generated by the semi-formal interviews consisted of three main dimensions. The feelings of the system performances were firstly concerned, as the subject was interviewed by the instructor to tell the differences of the system operations across the two studies. Another dimension of data was the subject's comments to the sense of system intelligence. Some sub dimensions were included in this dimension, including the system's interpretation capability fluctuations and the responding speed. And the other dimension of data provided was the consistency of system performances – this included an overall comparison across two studies.

6.2.3 Analysis: Evaluators' System Operation Differences and Impact

There were some variable differences observed across the studies, in terms of schemas' interpretation and executions, control panels' operations and the subject's action interpretations. Within each of these aspects, the differences and relevant motivations were highlighted in the following analysis in turn.

Firstly, the schema interpretations and executions were different in the operations. Although significant differences were not observed in the video analysis, the evaluators had reported some distinctive feelings to the schemas. These feelings were summarised as the differences of evaluators' capabilities to memorise and execute the schema rules. In the new evaluator's self-reporting, it was commented that some schema rules were relatively easier to execute while some of others were not.

Although the new evaluator had been trained and qualified to do the operations, the new evaluator had admittedly applied some personal understanding or preferences upon the schemas. From the experienced evaluator's point of view, similar differences were also admitted, as some of the schemas were relatively more understandable and memorable to execute. Both evaluators stated that such discriminations had little influence on the overall operation consistency levels, but some minor differences might be relevant in the evaluators' responding performances, which were described in the third aspects. One reason was hypothesised to explain such discriminations. Since the two evaluators had different backgrounds, it was likely for them to unconsciously discriminate those they were familiar with from those they were not. But in this stage more evidence was required to validate the hypothesis.

Secondly, more differences were found in the control panel operations. One most noticeable difference was that the two evaluators used different layouts of control panels (see **Figure 6.5**). As stated in the verbal descriptions, similarly the evaluators seemed the layouts as parts of operation preferences that could help them to better operate the control panels. Since other variables described below were also involved in the studies, in this stage the influence of the layout differences was insufficient to judge.

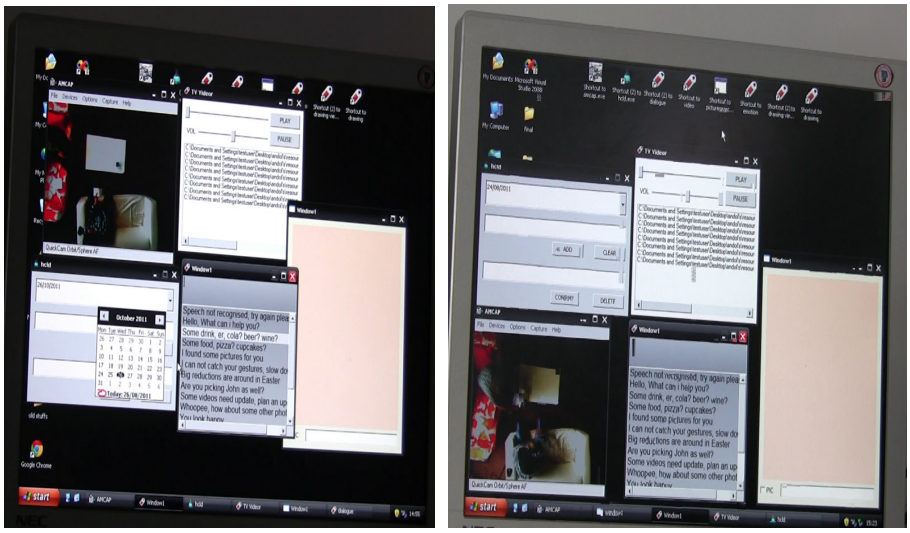

Figure 6.5*: Control panel layouts. Left – new evaluator's, right – experienced evaluator's*

Other differences were also observed in video recordings of evaluators' operations. In terms of the use of preset messages, the new evaluator was more likely to present 'unknown speech, please repeat again, while the experienced evaluator was observed to make predictions to the equivocal speech more often. Extract (1) and (2) demonstrated such difference when two evaluators gave different operations to ambiguous subject actions. The self-reporting provided little evidence to identify such differences, but the subject's comments validated the differences, as the change of system's operation styles was sensed. Such differences might be justified for two reasons, one was that the evaluators had different capabilities of recognising subject's speech, and the other was that this was influenced by the personal operation styles. In addition, this might also be influenced by the factor of confidence, as the new evaluators might be less confident thus leading to conservative operation strategy. The evidence generated from current studies was insufficient to prove that

this was relevant with individual confidence, therefore the next groups of studies would continue to investigate these hypothesises.

(1) The experienced evaluator's operations

W – evaluator, S – subject

1	*[29:05-29:14]*	*S:*	*So it would be something like dry food,*
			like pizza, snacks [hangle (unrecognised speech)].
			So that people don't miss around.
2	*[29:13]*	*W:*	*starting to add the food shopping appointment,*
			typing the contents as 'dry food – pizza,
			snacks', then directly confirmed the input
3	*[29:30-29:43]*	*S:*	*so maybe like pizza, burgers,*
			and some [long pause] biscuits
4	*[29:42]*	*W:*	*typing the new food as 'pizza burgers snacks'*
5	*[29:45-29:50]*	*S:*	*er, pretty enough*

(2) The new evaluator's operations

1	*[04:42-04:52]*	*S:*	*so if you can (long pause) invite three of*
			my friends, name 'munir', (long pause) 'jack'
2	*[04:53]*	*W:*	*typing the confirmation message as*
			'invite friends - muni?'
3	*[05:36-05:40]*	*S:*	*no, its 'munir', m-u-n-i-r (in fast spelling pronouns)*
4	*[05:41]*	*W:*	*confirm 'munir?'*
5	*[05:52-05:55]*	*S:*	*yes, this is correct.*

Minor differences were also observed in terms of the system's operation speed, as in the second study some specific actions were facilitated much slower than those in the first one. Based on the interviews, the subject was particularly sensitive to such differences. The subject seemed such inconsistent operations as 'unstable' system performances.

Other differences reported in the video analysis of the evaluators included the evaluators' message utterances and message typing speed and accuracy. Some personal utterances were extracted from the video recordings. The new evaluator was observed to use more question-style utterances, as demonstrated in the extract (2) line 2 and 4. In contrast the experienced evaluator tended to use confirming style utterances, as shown in the extract (1) line 2. The data from self-reporting implied that the utterances were closely relevant with the identifications the evaluators. In other words, to a same subject action, the evaluators might identify themselves as different roles such as machines or humans, and thus generating diverse utterances. Such differences were likely to be caused by different interpretations to the subject actions, more specifically, to the roles of a human or an intelligent system.

Thirdly, the differences of subject action interpretations could be further divided into two sub aspects. From the verbal descriptions, it was further confirmed that two evaluators used different criteria to judge the subject interactions, excluding the schema rules. For example, in extract (3) the new evaluator responded in forms of machines, as he identified himself as a human instinctively. The extract (4) demonstrated another self-identification of evaluators as a predefined machine. Surprisingly, the evaluators' also commented that the unrecognised subject speech would affect their judgement criteria by changing the identifications between machines and humans.

(3) The new evaluator's actions

1	*[11:55-11:57]*	*S:*	*using the cube to click the pictures on the table*
2	*[11:56]*	*W:*	*clicking the selected picture, and zooming in on a new window*
3	*[11:58-12:11]*	*S:*	*ok, I like this kind of boots, my size is six, can you find me the website for this boots*

4 [12:10] W: disappearing the picture,

starting the website search

(4) The experienced evaluator's actions

1 [15:00] W: typing the new appointment's contents

2 [15:01] S: do I need to save it? Save it.

3 [15:03] W: giving the message 'undefined operation. Please try again'

4 [15:06] S: no, er, save them, oh, confirm, yes

In addition, differences were also reported in the operations of multimedia manipulation. In terms of video programme changing, the new evaluator used to pick a random video to play, while the experienced evaluator intended to follow the video files' order. In terms of natural dialogue interactions, the new evaluator tended to select best suitable operations, as he indicated in the self-reporting; and the experienced evaluator looked for relevant operations. And in terms of the internet searcher, the new evaluator used to type single keywords, while the experienced evaluator used more phrases. These differences were mostly relevant with evaluators' practical operations. These could be seen as the results of the differences mentioned in the last aspect, or, of the hypothesised assumptions of evaluators' individual operation differences.

6.2.4 Analysis: Evaluators' Personal Differences and Impact

The impact of these differences described above was reflected in the dimensions of the four evaluation criteria. The overall levels of operation inconsistency were low, as both evaluators' operations interpreted and responded to the subject's actions accurately. The main impact was generated by the differences related to the unconscious operation.

Reliability

The reliability of operation referred to the levels of stability to which the evaluators maintained in their operations. In terms of individual evaluators, the stability was rated high due to throughout the whole study the evaluator carried out the operations in personal styles. For example, the new evaluator used to prompt a confirm dialogue when the subject gave fast and long speech; and the experienced evaluator likely insisted to use predictive prompts. In terms of multiple evaluators, however, the stability of the operations was relatively low, and resulting in subject's awareness of system intelligence fluctuations. The subject sensed clearly the differences of system performance, also some distinctive feelings were given.

Three aspects were concluded due to their relationships with the cross-evaluators operation differences, including the schema interpretations, control panel use and subject understanding. Particularly, the schema interpretations and executions, as well as the subject understanding was reported as the primary sources generating significant differences. As observed in the video recordings, the differences related to the control panel operations were more likely to be overcome by additional trainings. In the contrast, the inconsistent schema interpretations and executions could severely lead to different operation strategy. In addition, the schema interpretations were difficult to be synchronised between evaluators. The validation results would be discussed in the next group of studies, after the additionally intense trainings. The evaluators' differences were also difficult to address when these were relevant with subject's action interpretations. Two difficulties were anticipated to achieve that, including the synchronisation of speech recognition capabilities, and the synchronisation of criteria of judging how intelligent should the proactive prompts be.

Predictability

Some differences were reported in the video analysis that the evaluators adopted different strategy to anticipate the subject's actions. For example, the evaluators had different prepares when the subject said *'I need to look up the calendar*'. The new evaluator moved the mouse pointer to natural dialogue and prepared to type messages; while the experienced evaluator focused on the calendar data picker. Similar differences were mainly observed when the subject gave equivocal speech commands, such as 'I need to check the equipment' and 'show some photos'. The predictability differences also concerned with the subject's

interaction contexts understanding. The new evaluator tended to ask for more details when given equivocal speech, while the experienced evaluator tended to integrate relevant contents from previous interaction contexts.

The impact of these differences on operation predictability negatively aggregated the operation inconsistency, in terms of the sense of system intelligence and system usability. From the subject's point of view, such inconsistent operations led to unpredictable operations, which made the subject difficult to find the clues of further interactions.

Appropriateness

The impact of evaluators' inconsistent operations on the appropriateness of operations was twofold, including the individual evaluator and the group of evaluators.

In terms of the individual evaluator, the operation appropriateness levels were generally high, due to the evaluator presented the operations in forms of consistent styles. Although such styles were individually personal, these were considered system's properties by the subject. But in terms of whole group of evaluators' operations, the appropriateness levels were noticeably variable. Two aspects were concerned with the low levels of overall operation appropriateness. Firstly, it was believed that the fluctuations of system's speech recognition capability led to inappropriate operations. For example, in extract (2) the subject gave a name 'Munir' for new appointment, the experienced evaluator spelled it correctly in the first time, but the new evaluator was not able to give the correct spell in the beginning, which directly lowered the subject's feelings of the sense of system intelligence.

Effectiveness

There was sufficient evidence showing that the operation differences also included the effectiveness disparities. From the subject's point of view, the effectiveness difference was mainly represented in forms of system responding speed. In other words, the system had the operations in inconsistent speeds – rather than a computer-like stably performing system. In addition the difference of effectiveness between two

evaluators was also presented as the various numbers of proactive prompts, as the new evaluator was observed to use more confirm prompts.

Summary

The overall effectiveness difference was sensible from both the subject and the evaluators' point of view. The video recordings of the evaluators' operations clearly showed the difference of average operation durations. However, that was not to say that the two evaluators had completely different in operation effectiveness. Instead, the difference of effectiveness extracted in these two studies were more like interesting indicators, which showed the consequences of these significant differences described above.

6.2.5 Discussion of Study Findings

There was rich evidence to suggest that the two evaluators had significant differences in terms of operation preferences and subject understanding. The methods adopted in the studies generated both positive and negative data related to the system operation differences. As well, the intelligent system in the studies was naturally used by the subject, and the evaluators' operations were based on these natural interactions.

However, the evidence was not firmly provided in the studies to imply whether these differences were likely to be addressed, since multiple variables were involved simultaneously. The author did not isolate these variables as a single factor, and then recognised its influence and solutions. This part would be done in the next group of studies, as it could help the author to better identify the relationships between these variables and operation inconsistency levels.

This group of studies did not aim to produce a statistical conclusion of the operation differences. Rather, the author carried out these studies in incremental steps towards the understanding of how such differences could be addressed to enhance the reliability of WoZ in multiple evaluator studies. Interesting data was generated from the studies and implied that, with the best evidence from current studies, some types of operation differences naturally existed between the evaluators whose operations were even well trained. Three main aspects were identified as the sources of such differences, and two of these were preliminarily

considered more important ones. The weakness of this group of studies was that, the studies generated overall evidence that clearly confirmed the differences between evaluators' operations, but more specific data was lacked to indicate whether the differences resulted from one of variables or it was a synchronised effect.

The levels of evaluators' operation differences might be dependent on the two primary aspects. That was to say, attention needed to be placed on these aspects in the following studies, to gain the understanding of specific variable influences.

Therefore, to identify these variables' impact on operation inconsistency levels, a second group of studies were planned. More constraints were planned to be applied to the variables described. The intention was that the evaluators would be facilitating the intelligent system with little influence from control panel operations. The potential variables related to control panels such as the layouts of control panels would be fixed and isolated as independent variables. In order to make the evaluators' operations be relevant with only schema interpretations and subject understanding, intensive trainings were provided to both evaluators. In addition, the communication between the two evaluators was also enhanced, by learning each other's operations. This intended to mostly improve the evaluators' skills of control panel operations, including the selection of control panel functions and the use of the controls, and thus controlled the influence generated by control panel use.

6.3 Study 5 – Improving Variables in Multiple Evaluators WoZ Studies

In contrast to the first group of studies, the evaluators were trained with more learning and practices, through which it was to minimise the control panel influence. To obtain specific data from the variables related to schemas and subject interpretations, the control panels' influence now was constrained at the minimum levels. This was essential to the group of studies to highlight the variables in the aspects of schema and subject understanding.

A summary of this study's key components was presented in **Table 6.2** as below.

Study objectives	*To observe individual evaluator's system facilitation* *To measure the effects of proposed measurements to address the difference between individual evaluators*
WoZ system for study	*Remained the same as previous study's*
Study variables	*Focal variable – individual evaluators' system facilitation differences*
Study method	*Remained the same as the previous study's*
Data collection method	*Remained the same as the previous study's*

Table 6.2: *Summary of study 5's key components*

6.3.1 Method of Proposed Studies

6.3.1.1 System Development

The main system amendments were made in terms of the control panels' layouts, as described in **Section 6.3.4**. These consisted of three technical concerns, including the new control panel layouts accepted by both evaluators, the layouts constraints applied using the programming, and the mechanisms of restoring the layouts when the control panels were crashed and re-launched again.

To find out the common control panel layouts, an initial comparison was carried out between the layouts used in the first group of studies. The most different of the layouts was the live video monitor window (see **Figure 6.5**), and in this group of studies the window was fixed above the calendar control panel. This layout had been used in the trainings and the practices showed both evaluators had been used to it. Secondly, more programming constraints were applied in the new layouts, which integrated automatic alignments of the control panels, thus the separated control panels would be in strict arrangements as planned. And finally, additional mechanisms were applied in the layouts arrangements, as the control panels would be restored to the previous positions in the screen when these crashed down and were re-launched by the evaluators.

6.3.1.2 Study Structure and Evaluation Variables

Study Variables

As illustrated in graph **Figure 6.4**, this group of studies followed similar experimental settings as study 4. Since intense training was provided and the control panel design was fixed, the variables relevant with evaluators' control panel operation were restricted as independent variables. Thus, this group of studies aimed to highlight two variables in terms of schema interpretation and individual system operation preferences.

Data Collection Methods

Since this group of studies aimed to collect similar data for evaluators' system operation difference analysis, the data collection methods used in study 4 remained. As well, due to this group of studies aimed to evaluate other variables that were highlighted in study 4, the evaluation criteria also remained to reflect how the restrictions of control panel operation helped improve system operation consistency.

Subjects

A new subject was used in this group of studies, since the previous subject was likely to compare new system operations to the first operations before intense training and control panel restriction. For this reason,

the focus of new subject was on comparisons of system operations with improved control panels. Two evaluators in this group of studies were remained, but received additional training to ensure reliable control panel operation.

6.3.1.3 Study Procedure

This group of studies followed a similar procedural flow as study 4. The scenarios, tasks and schemas were all remained. Two evaluators were as well remained with additional operation trainings in terms of control panel operations. But the studies selected a new subject who would accomplish the tasks and interacted with the evaluators in turn via the operation system. The subject's unawareness of the evaluators was guaranteed throughout the studies.

6.3.2 Study Results

Each study lasted around 30 minutes similar as the previous studies, excluding the instructor's prior study introductions and the post-study interviews. Three methods remained to collect the data from the verbal descriptions, video recordings, and semi-formal interviews.

The data generated from the video recordings focused on the subject's reactions to the system operations. In terms of evaluator’s control panel operations, the video recordings were mainly used to validate the consequences of the training, by comparing two evaluators' operations. Given the trainings the data from the evaluators' self-reporting emphasised on the evaluators' experiences of interpreting the subject and the difficulties of making proper operations. In addition, in terms of the subject's interviews the focus of data collection was on the feedbacks of system's sense of intelligences. Although the variables such as the system responding time were also included in the interviews, these were not seemed as negligible factors.

6.3.3 Analysis: Differences in Schema and Subject Interpretation and Impact

The aspects of schema interpretations and subject understanding remained noticeable sources of operation differences. However, these differences found in the studies centred on the unplanned and predictive operations. Presenting consistent operations over all operations was at times difficult for the evaluators, especially when unspecified subject actions were given and these were uneasy to synchronise.

In terms of schema interpretations, the operations were generally consistent when dealing with the strict schema rules. For example both evaluators presented highly consistent responses in manipulating the calendar appointments. The flexible schema rules, however, showed some inconsistent operation. For example, in extract (5) and (6) the two evaluators presented sensitively incorrect predictions when facilitating with the calendar. More differences like these would be observed in the studies, if two evaluators received the same speech. Some speech that was not recognised by one evaluator was rarely used again in the next study by the subject. Also, there were some differences in judging the situations to make proactive actions. Both evaluators reported the difficulties of judging the right time and forms to make the proactive actions.

(5) The new evaluator's operations

1	*[01:45-01:46]*	*S:*	*appointment*
2	*[01:46-01:58]*	*W:*	*(preparing for new appointment inputting)*
3	*[02:01-02:05]*	*S:*	*today's appointments*
4	*[02:06-02:07]*	*W:*	*selecting the date of today, displaying the events*

(6) The experienced evaluator's operations

1	*[03:25-03:28]*	*S:*	*I want to make a party*
2	*[03:29-03:31]*	*W:*	*(preparing for the selecting of appointment date)*
3	*[03:32-03:35]*	*S:*	*appointment*
4	*[03:37-03:39]*	*W:*	*starting the appointment inputting*

5	*[03:42-03:55]*	*S:*	*er, need to buy some food,*
			3 kg beef … (more words continued)

In terms of the subject understanding, the speech recognition performances were relatively equal between the evaluators. However, two minor differences were captured in understanding the subject's motivations and interactions. The new evaluator was observed with lower tolerances of the subject's interaction mistakes. For example, when the subject in the studies made an equivocal pronoun of '30th', the new evaluator insisted to prompt two confirmations, until the subject repeated a very clear pronoun (see the extract (7)).

(7) The new evaluator's tolerance to subject's mistakes

1	*[08:05-08:07]*	*S:*	*September 30th (sounds like 13- - th)*
2	*[08:08-08:12]*	*W:*	*responding the confirmation as*
			'new appointment for SEPT 13- - th?'
3	*[08:14-08:19]*	*S:*	*er, delete*
4	*[08:23-08:26]*	*W:*	*disappearing the confirmation message*
5	*[08:34-08:37]*	*S:*	*new appointment*
6	*[08:37-08:39]*	*W:*	*preparing for new appointment input*
7	*[08:47-08:49]*	*S:*	*September 30th (still sounds like 13- - th)*
8	*[08:50-08:51]*	*W:*	*display confirmation message 'sept 13- - th?'*
9	*[08:52-08:56]*	*S:*	*September, thirty- -a- -s*
10	*[08:58-09:02]*	*W:*	*selecting the right data from the calendar,*
			Prepare for new appointment input

As demonstrated in the extract (7), the statistics of the evaluators' preferred operations were found with a few differences. Firstly, the different frequency of confirmation still existed between the evaluators, as the new evaluator had more of this. Secondly, it was further noticed that the confirmations used by the evaluators had different use. The new evaluator's confirmations connected more with the next operation,

such as 'confirm delete?' and 'show the 5th picture'. On the other hand the experienced evaluator's confirmations were more relevant with the contents, such as 'the bottom picture'. Such operation utterance differences were considered irrelevant with the control panels, nor with the schemas or subject's actions. Such that existed in both groups of studies was believed to have essential relationships with the evaluators itself. The variables related to these differences seemed not be easy to be changed by trainings, and these were probably integrated within the evaluators in unconscious way, like individual habits.

Some other minor differences related to schemas and subject interpretations were also captured, but were considered negligible due to the little impact on the subject's interactions, such like the overall speech recognition accuracy. The relationships between the operation differences and the evaluators' individual preferences had been confirmed in the studies. Providing additional trainings did have an effect on the reduction of operation differences, but the relevant variables seemed not likely to be affected in easy trainings.

6.3.4 Analysis: Differences in Control Panel Operation and Impact

Some operation differences in the control panel operations were worth to report here, since these had potential dependencies on the above aspects. The consequences of these differences, however, had little impact on the operations, due to these only being observed from the video analysis without the subject's validations.

Firstly, it was found that the evaluators had different preferences to do the predictive operations. This was also validated in the extract (1) and (2). The new evaluator tended to put the mouse pointer on the calendar dates, but the experienced evaluator put the pointer on the date picker. Secondly, the idle operations were different when the evaluators did not present operations. The new evaluator used to move the pointer around the screen, but the other evaluator tended to leave the mouse on the natural dialogue control panel.

6.3.5 Discussion of Study Findings

This section discussed the evaluation results which were described in **Section 6.4.4** and **Section 6.4.5**. Although some minor differences of evaluators' control panel operations were reported, it was argued that the studies had achieved the goals successfully. Firstly, the additional trainings prior to this groups of studies had good effects on synchronising the evaluators' practical control panel manipulations, thus the influence of control panels was minimum. In this regard, the potential variables related to the control panels were believed to be isolated in the studies. Secondly, a set of operation differences were reported across the studies, and then clearly confirmed the influence variables from the two primary aspects.

Three relevant variables were justified as the reasons for the main operation differences. As described in the prior sections, the studies consistently generated operation differences in terms of schema interpretations and subject understanding. Respective variables were summarised to be responsible for these differences.

In terms of schema interpretations, the differences were justified for two potential variables. The evaluators' judgement criteria against the schemas were an important factor generating inconsistent operations. These criteria centred on the proactive operations. Noticeable disparities could be observed between the evaluators towards whether a subject action should be included in the schemas. Such differences were difficult to address by trainings. Furthermore, the evaluators' operation routine planning could be also an important variable. Based on the evaluators' consistent judgements, the routines planned to execute the schemas might generate inconsistency. As the video analysis demonstrated, the evaluators would adopt different strategy to prepare for the next operations – one focused on calendar data picker and the other one focused on natural dialogue. From the evaluators' point of view, on the one hand such strategy were adopted naturally along with the schema interpretations. On the other hand, these individual strategy adoptions might lead to significant operation differences, from the experimenter's point of view.

The differences of operation strategy planning might be motivated by the evaluators' capability diversities. That was to say, the individual properties of evaluators such as background knowledge and personal perceptions, could drive the evaluators to adapt to different operation strategy, and bring some instinctive preferences of operation.

There was preliminary evidence suggesting that these variables were much tougher than the control panel operations to address. The difficulties were anticipated in two aspects. Firstly, it was believed hard to sketch schemas covering all possibilities of subject's natural language interactions. If that could be done, then the evaluators as humans might still not be capable to memorise and execute these schemas. Secondly, conventional trainings had little effect on these variables. As demonstrated in study 4 and 5, prior trainings likely worked on practical operation differences, but not on these individual variables.

To validate these hypothesised variables and to identify the solutions, the next group of studies were proposed with these improvements, as follows.

- A set of keywords were extracted to trigger the proactive operations – The intention was to provide references to avoid the differences related to evaluators' judgment criteria.

- More training was planned to assure the evaluators to be capable to capture the keywords accurately.

6.4 Study 6 – Controlling Multiple Evaluators' System Facilitation

This group of studies focused on two influence variables, including individual evaluators' background knowledge and recognition capability. These two variables were implied in findings of study 5 but without careful addressing. In this study, additional evaluator training was provided, and more constraints were also applied to evaluators' interpretation of schema and subject. The intention of improved constraints was to ensure evaluators to make consistent interpretation and operation routine plan. Since these measurements were applied in the studies, the studies were presumably based on equivalent judging criteria for proactive operations.

A summary of this study's key components was presented in ***Table 6.3*** *as below.*

Study objectives	*To observe individual evaluator's system facilitation* *To measure the further effects of proposed measurements to address the difference between individual evaluators* *To reflect the overall results of system facilitation after applying these measurements*
WoZ system for study	*Remained the same as previous study's*
Study variables	*Focal variable – individual evaluators' system facilitation differences*
Study method	*Remained the same as the previous study's*
Data collection method	*Remained the same as the previous study's*

Table 6.3: *Summary of study 6's key components*

6.4.1 Method of Proposed Studies

6.4.1.1 System Development

To make a consistent comparison throughout the studies, the system applications in this group of studies were remained. In addition the control panels for evaluator use remained, due to the improved control panel design in study 5 supported evaluators to produce consistent operations.

6.4.1.2 Study Structure and Evaluation Variables

Study Variables

This group of studies had similar structure of experimental setting as study 4 and 5. Since the variables such as control panel operation and subject interpretation were strictly constrained in this group of studies, the main dependent variables were evaluators' personal background knowledge and speech recognition capability.

Data collection methods

The three data collection methods remained, as this group of studies also required the data captured in study 4 and study 5. Meanwhile, as this study had similar aspects of system operation to evaluator, the evaluation criteria were remained as well.

Subjects

Similar as study 5, this study used another subject who compared system operation difference after system improvements. The evaluators in this study were remained as well, however their operations were required to follow new measurements.

6.4.1.3 Study Procedure

The studies followed a similar flow of study 4 and study 5, with the same evaluators receiving more supports to proactive judgment. According to the prior scenarios and tasks, the list of keywords was extracted with which the evaluators could make proactive operations. More specific constraints were also applied to the operations, such like the limited number of confirmation prompt use. The new subject followed the same procedures as precious subjects did. The intention of such improvements was to minimise the variable influence from equivocal judgment criteria.

6.4.2 Study Results

Each study took around 40 minutes on average. This was slightly longer than the previous studies due to the evaluators provided limited proactive assistances. Thus some interesting data was captured in terms of the subject's interactions, which was useful for the intelligent system design. Video analysis, verbal descriptions and interviews provided rich data in terms of operation differences. The focus of the data was on evaluators' differences of subject's motivation and interaction judgments, rather than the practical operations.

6.4.3 Analysis: Differences in Evaluators' Judgments and Impact

The operations of the evaluators remained a generally consistent procedure, with some minor inconsistency found about the proactive operations. But first of all the reliability of the judgment criteria support needed to be confirmed. From the video analysis of evaluators' operations, several examples of proactive operations were observed, with rapid responses to subject's actions. These responses, as transcribed from the videos, were triggered by specific keywords in the vocabularies such like 'hiking pictures'. For example, both evaluators autonomously started the picture searcher when the subject said 'hiking equipment'. At this point, the reliability of judgment criteria was successfully achieved, based on the observations of evaluators' operations.

Some previous operation differences were repeated in the studies. Firstly, few operation differences were found relevant with the schema interpretations. Due to the additional support of proactive operation trigger keywords, the evaluators in the studies had much less equivocal judgments to make. It was clearly the trigger keywords that to some extent supported the evaluators to judge consistently. Secondly some minimum control panel -related differences also reoccurred, such like the movement of mouse pointer. However, such differences were not observed to generate sensible operation differences.

Furthermore, the output of utterances was reported with noticeable inconsistency between the evaluators. In the studies, the differences included the tones, lengths and expressions of messages. These centred on the text typing. In contrast the multimedia operations were much more consistent in terms of output of responses. This was precisely relevant with the flexibility of text typing method, even more, such relevance was driven by evaluators' individual differences, such like the stability of self-role recognition and usual conversation tones and expressions. However it was out of the thesis's range to investigate the uniqueness and the impact of individuals' speech. The evidence generated from the studies, nevertheless, suggested the irrelevance between the operation trainings and response utterances.

Differences of operations were also reported that the evaluators had occasional misrecognitions. These cases seemed like randomly happened, since several misrecognitions were made by both evaluators but some were not. Based on current understanding, this was likely relevant with the evaluator individuals, in terms of perceptions.

6.4.4 Discussion of Study Findings

This section describes the findings of the studies, in which there was little evidence generated to further indicate the variables and solutions related to evaluators' knowledge and capabilities of operation. It was argued that the studies were a success on a few counts. That was to say, the trigger keywords did help to enhance evaluators' judgments on proactive operations, and thus the connections between evaluators' individual knowledge and capabilities of operation were confirmed. Furthermore, this was one of the few examples of evaluators' reliability evaluations performed to date. A key evaluation aim of the studies was to

explore the solutions to the unreliability of individual differences. Based on the range of feedback received – some new, some repeated – it was argued the studies achieved partial success.

The weaknesses were that the data generated form the study was limited, thus leading to the repetitions of evaluations. On the other hand, the studies validated the understanding gained from previous studies that these individual variables had significant impact on operation consistency, and were difficult to be changed via trainings. The weakness of the studies was acknowledged that more special speech such like abbreviations should be adopted to test the ultimate levels of recognition differences. Due to the limitations of the scenarios and tasks, the evaluators in this study had been very familiar with the possible speech used to interact with the system, and thus reducing the width of evaluations.

6.5 Comparative Review of Study 4, 5 and 6

This section looks through all three groups of studies carried out, and summarises the variables and solutions to the operation inconsistency in cross-evaluators operations.

Firstly the overall study operation consistency levels incrementally increased. In the first group of studies, the newly trained evaluators facilitated a subject, and produced a number of operation differences. These differences provided good materials for the later study evaluations to identify the influence variables. The following aspects may have summarised the main operation differences in cross-wizard operations.

The schema interpretations – as noted in **Section 6.3.2** and **Section 6.3.3**, it was difficult for evaluators to memorise and execute all schemas constantly, changes may be of the improvisational natural of flexible schemas. It was acknowledged that the new evaluators, who was newly trained, stressed the needs for more support of schema following. This emphasised the importance of carrying out the incremental studies to allow the progresses of evaluator. Despite this, the schema interpretation differences were implied that there were deep motivation variables.

The control panel operations – as indicated in **Section 6.3.4**, control panels caused a number of differences between evaluators, but these might also be the easy part to improve. This meant that the evaluators should have more opportunities to exercise the control panels, and be given less constraints to the operations. In **Section 6.4.3**, the additional training of control panel operations was demonstrated to be effective. It also confirmed that the rest two aspects were relatively uneasy to improve.

The subject understanding – a range of variables were justified as the reasons for differences of subject understanding between evaluators. In **Section 6.5.3**, it was indicated that the variables such like evaluators' individual knowledge and recognition capability were closely relevant with evaluators itself, and were likely the most difficult variables to change.

The incremental studies continued to narrow down the range of variables that might cause operations differences. In the first group of studies three general influence aspects were abstracted. Then in the second

group of studies the two primary aspects were confirmed, and these also identified the common variables behind the schema and subject action understanding. And the third group of studies validated these specific variables. All studies were incrementally based on previous studies. A key effect of this was that the evaluators' operation differences were observed in an evolutionary way.

The study results highlighted the benefits of evaluating over the incremental progresses. The variables related to control panel operations were improved, and showed to be able to improve further. Other variables related to understanding, including recognition capabilities, background knowledge insufficiency, perception preferences, were showed more difficult to be change in the later studies. Thus, this suggests that the reliability between two evaluators was clearly lower than the reliability of one evaluator. Furthermore, synchronisations between evaluators face variable difficulties, depending on the relationships to evaluators' individual properties. However it is envisioned that extra trainings could effectively enhance the reliability of multiple evaluator operations, since some of these variables could still be gradually improved.

6.6 Contribution

The contribution of the studies in this chapter was summarised as follows:

- Firstly, the study provided understanding of influence variables that affected evaluators' operation consistency. Three main variables were identified as evaluators' perception of control panel operation, evaluators' individual judgement criteria and evaluators' personal expertise and experiences.

- Secondly, by gradually controlling these variables, the incremental studies suggested out that variables relevant with control panel operation and interpretation criteria could be improved by applying careful constraints and assistance. However the difference of personal expertise and experiences could only be improved to an acceptable level.

- A secondary contribution made in the studies also included methodological recommendations for WoZ study. These included aspects such as multi-evaluator training and video recording.

6.7 Conclusion

This chapter presents three groups of studies to incrementally investigate the variables related to cross-wizard operations and the operation reliability. The evaluators received multiple training to facilitate different subjects. A number of findings from the studies have been discussed. Firstly the reliability in multiple evaluator operations was proved to be lower than in one evaluator operations. The studies also indicated that, synchronisations between evaluators face variable difficulties, depending on the relationships to evaluators' individual properties. However it is envisioned that extra training could effectively enhance the reliability of multiple evaluator operations, since some of these variables could still be gradually improved.

Next, **Chapter 7** moves on to have an overview of the findings from all previous studies, and discusses the overall contributions and conclusions to the thesis.

Chapter 7

Research Findings and Contribution

7.1 Introduction

This chapter discusses substantive findings from **Chapter 4**, **Chapter 5** and **Chapter 6**, and maps these findings to the questions from literature review. Previous studies have investigated major system components in WoZ study and identified several key variables that affected the reliability and validity of WoZ. Since these studies were incrementally conducted, the overall finding of these studies was not discussed comprehensively in previous chapters. Thus, in this chapter, these findings (study 1 in **Chapter 4**, study 2 and 3 in **Chapter 5**, and study 4, 5 and 6 in **Chapter 6**) were reviewed and summarised as the base of research contribution and future work.

This chapter is made up of the following sections. Firstly, **Section 7.2** presents an overview of how study findings addressed the questions from literature review. Then, **Section 7.3** presents a critical revisit of research framework and methodology of previous study. Research contribution is described in **Section 7.4**, and it is followed by an overall assessment of this thesis in terms of fulfilling research objectives, weaknesses, strengths and challenges. And finally **Section 7.5** reveals some directions of future work.

7.2 Research Questions and Findings

In the literature review in **Chapter 2**, several questions relevant with the reliability and validity of WoZ were described. In considerations of practical WoZ study in **Chapter 3**, these questions were developed to several key perspectives of study plan. Each perspective was motivated using findings from earlier studies, and was used to extent incrementally the understanding of the reliability and validity of WoZ. These perspectives also indicated routes for the studies in this research, as a series of empirical studies were carried out along with these perspectives and a rich set of experimental findings were offered.

7.2.1 Questions from Literature Review

The literature review examined how the reliability and validity of WoZ were affected by the evaluator. Together with considerations on technical requirements for domestic communication study, several questions were extracted based on the understanding of previous WoZ studies. These questions covered major components of WoZ study, and were summarised as follows.

- How the schemas were designed, and how evaluator's schema interpretation and execution were evaluated?

- How the control panel was designed in WoZ operation system, and how the control panel design affected evaluator's system operation and thus affected WoZ's reliability and validity?

- How subject's activities and intention were interpreted by evaluator, and how the interpretation progress affected the consistency of system operation?

- Whether multiple evaluators could produce consistent system operation in WoZ study, what variables hindered the synchronisation between evaluators, and how evaluators' system operation difference could be addressed?

Based on the findings in the research, there was a critical improvement to understanding of these questions.

- How the schemas were designed, and how evaluator's schema interpretation and execution were evaluated – This question was addressed by identifying a key variable of the rigorousness of schema design. In addition, the understanding of how the varieties of schema design affected evaluator's operation was also provided. The schema interpretation was addressed in the study with subject interpretation.

- How the control panel was designed in WoZ operation system, and how the control panel design affected evaluator's system operation and thus affected the reliability and validity of WoZ – This question was answered by a set of variables relevant with control panel design. In particular, two variables of control panel function design and layout design were highlighted along with the understanding of how these variables affected system's operation. Also, the understanding included guidance for reliable control panel design.

- How subject's activities and intention were interpreted by evaluator, and how the interpretation progress affected the consistency of system operation – This question was addressed by a group of studies comparing the difference of subject activity interpretation. These studies highlighted several key variables relevant with subject activity and intention interpretation. These variables comprised judgement of subject intention and the match between schema and subject interpretation. As well, the advantages and difficulties were provided in these studies.

- Whether multiple evaluators could produce consistent system operation in WoZ study, what variables hindered the synchronisation between evaluators, and how evaluators' system operation difference could be addressed – This question was addressed by carrying out a set of incremental studies. The studies identified influence variables causing interpretation and operation difference between evaluators. These variables included evaluators' interpretation of schemas and control panel, judgement criteria and personal expertise and experiences.

7.3 Revisiting the Research Framework

This research has been aimed at increasing on HCI knowledge with respect to the reliability and validity of WoZ. As described in literature review in **Chapter 2**, to date researchers encountered a need to update the understanding of design and evaluation methodologies. Consequently evaluator-participatory methodologies were particularly highlighted due to the dynamic nature of evaluator' system operation. This research focused on one specific evaluator-participatory methodology – WoZ. As discussed in the review of WoZ use, previous studies were limited in terms of improving the reliability and validity of WoZ. Although studies used WoZ as primary design and evaluation method, few had systematic considerations on the variables and impact. Therefore, there is a lack of foundation for systematic work aimed at implementing and evaluating these variables. Objectives of this research were summarised as follows.

- To understand how schema design affected the consistency of evaluator's system operation, and then affected WoZ's reliability and validity

- To understand how control panel design affected the consistency of evaluator's system operation, and then affected WoZ's reliability and validity

- To understand how the evaluator's interpretation of schemas and subjects affected the consistency of system operation, and thus affected WoZ's reliability and validity

- To understanding how multiple evaluators should be involved with consistent system operation output, and thus maintain high reliability and validity of WoZ

To achieve these aims the research was structured on a three-step progress of WoZ system development and evaluation. These steps included 1) schema design and evaluation, 2) control panel design and subject interpretation evaluation, 3) and multi-evaluator evaluation.

- Schema design – Schemas were concerned throughout the studies in this thesis, as each study required specific schemas to guide evaluator's system operation. The exploratory study, reported in study 1 in

Chapter 4, investigated the variables relevant with different schemas. This enabled following studies to be carried out based on reliable and effective schema guidance.

- Control panel and subject interpretation – Since the control panel was important to connect schemas and front-end system applications, it was concerned as a component that had significant influence on evaluator's operation and WoZ reliability and validity. The study 2 in **Chapter 5** investigated how different designs of control panel caused inconsistent operation, and revealed how the design of control panel function and layout helped improve system operation consistency level. Also, this understanding provided a practical grounding for the latter study of multiple evaluators. In addition, evaluator's interpretation of subject's activities and intention was investigated in study 3 in **Chapter 5**. This study addressed the difficulties that hindered evaluator to produce accurate subject interaction interpretation and anticipation.

- Multiple evaluators – **Chapter 6** reported the third step of the research, the multiple evaluators' operation in WoZ study. The evaluation facilitated the comparisons between two evaluators throughout three incremental studies. The variables identified in previous studies were tapered off gradually, and that highlighted the influence of evaluators' individual personalities and preferences on the consistency of system operation.

7.4 Contribution to Knowledge

Corresponding to the three-step progress of research above, the overall contribution made by this research consisted of several empirical understanding. These were structured in two main areas, including the contribution resulting from the evaluation of variables relevant with the reliability and validity of WoZ, and the contribution resulting from the design and evaluation of intelligent device for domestic communication study.

7.4.1 Improving the Reliability and Validity of WoZ Studies

This section describes the contributions made throughout this research in terms of methodological and practical understanding. To echo the research questions proposed in the beginning of research in **Section 2.3.3 "Research questions about the reliability and validity of WoZ"**, these contributions were described in the same order by following the studies of schema, control panel and subject activity interpretation, and multiple evaluators. Lastly as an important part of contribution, practical understanding was also included.

Firstly, based on the exploratory study in **Chapter 4**, understanding was provided about schema design's impact on evaluator's operation consistency. This contribution helped improve the organisation of schemas in various forms, and it provided reliable guidelines to pre-study training, as such did in Molin (2004)'s study. Also, this contribution provided guidelines to the design of schema rigorousness according to different system applications, which extended Bickmore (2002)'s opinion of how much flexible the operation should be. In general, this contribution provided the understanding of Fraser and Gilbert (1991)'s questions relevant with schema variables.

- It revealed that rigorous and general schema design did not support highly consistent system operation. Rigorous schema design provided efficient guidance to planned system operation, but it was not scalable for natural-language dialogues that used non-restricted speech commands. Also the weakness of such schema design included limited number of system operations, as a long list of such schemas cannot enhance evaluator's operation in terms of predictability and effectiveness.

- The general schema design, on the other hand, had high effectiveness but also lowered the reliability of system operation due to inconsistent operations.

- Combined schema design has more consistent output due to it applied strict guidance to anticipated system interactions and general guidance to unexpected interactions. Both were in the range of consistent operation.

Secondly, the study 2 in **Chapter 5** provided two further empirical contributions.

- Improving of control panel design variables and impact was achieved. The study revealed that control panel's function and layout design had important influence on evaluator's operation consistency levels. By applying different constraints to control panel, the combination of strict function and flexible layout has improved stable reliability and validity. With respect to control panel, the studies indicated an effective way of control panel use. This contribution provided crucial criteria for researchers to judge how applicable a control panel design was. Also, this contribution provided guidelines to resolve Forbes-Riley and Litman (2009)'s concern about reliable multi-communication channel operation.

- The strict function design provided fewer chances of improvisational operations, and flexible layout design assisted evaluator's to build up an efficient work space which fitted with personal operation preferences in control panel operation.

In addition, the study 3 in **Chapter 5** provided another empirical contribution.

- Understanding of evaluator's interpretation of subject activities was provided. The study revealed that most interpretation errors occurred in the stage of interpretation of subject intentions. The context-based subject intention prediction only worked with events that were planned in the calendar. Also, it has noted that the matching was variable between subject interpretation and schemas, due to some subject activities were ambiguous to follow the schema. With respect to subject interpretation in WoZ study, the contribution identified the evaluator's predictive interpretation as the key part of interpretation consistency. And based on this, this contribution extended current measurements of interpretation to planned subject activity in (Bradley et al. 2009; Carbini et al. 2006).

Thirdly, the study 4, 5, and 6 in **Chapter 6** investigated multiple evaluators' system operation in WoZ study. These studies compared evaluators' system operation difference by narrowing down variables that were identified in previous studies. Two further contributions in this area were made in these studies. This contribution provided an empirical guidance to multi-evaluator WoZ studies such as the study in (Salber and Coutaz 1993). Also, this contribution provided a practical grounding to maintain high reliability and validity in multi-evaluator and multimodal system studies, such as those in (Carbini et al. 2006; Edlund et al. 2008).

- The study provided understanding of influence variables that affected evaluators' operation consistency. Three main variables were identified as 1) evaluators' perception of control panel operation, 2) evaluators' individual judgement criteria and 3) evaluators' personal expertise and experiences.

- By gradually controlling these variables, the incremental studies suggested out that variables relevant with control panel operation and interpretation criteria could be improved by applying careful constraints and assistance. However the difference of personal expertise and experiences could only be improved to an acceptable level.

7.4.2 Improving Practical Design and Evaluation

The contribution in this part resulted from the practical work of WoZ operation system and smart device design, implementation and evaluation. These contributions provided empirical guidance for reliable WoZ operation system design and natural-language dialogue system development. This contribution provided further understanding of reliable WoZ system design based on previous WoZ studies such as (Carbini et al. 2006; Dahlback et al. 1993; Fraser and Gilbert 1991; Mavrikis and Gutierrez-Santos 2010). Also, this contribution made incremental understanding of smart system design which broadened the practices in studies of (DEMIRIS et al. 2006; Hsu et al. 2010).

Also, this contribution made progresses in understanding how smart devices should be designed, and this improved the understanding of

The first contribution was the WoZ operation system itself, as described throughout all studies in **Chapter 4**, **Chapter 5** and **Chapter 6**. The operation system was offered as a platform for WoZ operations. It enabled evaluator to mimic apparent system functions. As highlighted in study 2 in **Chapter 5**, the operation system design was concerned with most variables. The operation system in this research was an example of incremental design, and offered as a case-study of WoZ use for domestic communication study. The initial evaluation of the WoZ operation system, as described in each study's technical developments, resulted in several design improvements for more complicated smart device. These included the distribution of system front-end and back-end applications, the communication between control panel and system applications, and the presentation of real-time information retrieving.

The studies across **Chapter 4**, **Chapter 5** and **Chapter 6** also resulted in other practical contributions relevant with smart device design for domestic communication study. These studies indicated that smart devices for domestic communication needed much clearer indications of system capability and progress. The indication of system's speech understanding affected how the smart device was treated by subjects, such as human-like conversational agent or an intelligent speech recognition system. The indication of system progress offered similar impact, and also it was particularly useful when subjects used previous desktop computer system operation experiences in the natural-language spoken system.

7.5 Critical Review of Thesis

This research consisted of a series of dual-purpose empirical studies. In each study there were several objectives. To offer a critical review of this thesis, it was necessary to consider the overall success of the research in achieving these study objectives. Both the weakness and strength of the study, and design and evaluation methodologies in the research are concerned.

7.5.1 Fulfilling Research Objectives

The overall achievement of research objectives was considered successful, as the main studies described in **Chapter 4**, **Chapter 5** and **Chapter 6** reached generally the goal of identifying relevant variables and offering effective measurements. The respective achievement of study goals for each study is described as follows.

Study 1 in **Chapter 4** successfully gained practical understanding of schema-related variables and the requirements for operation system and smart device design. After a series of empirical studies, this chapter offered the understanding of how schema design's rigorousness affected operation consistency, and it also extracted implications for the design efficient natural-language dialogue system manipulation. However, few practical implications were offered in terms of WoZ operation system improvement in this study.

Study 2 and 3 in **Chapter 5** provided practical understanding of variables relevant with control panel design and subject interpretation. Two respective case studies were carried out, and found out important variables such as control panel's function and layout design and the criteria for subject activity judgement. With respect to control panel design, study 2 reached its goal by finding an effective way of combining control panel function and layout design. With respect to subject interpretation study 3 was considered generally successful due to the understanding gained.

And finally, study 4, 5 and 6 in **Chapter 6** aimed to investigate multiple evaluators' operations and the impact on the reliability and validity of WoZ. The findings of these studies did not identify new variables, but the

understanding was deepened in terms of how these variables affected multi-evaluator system operation consistency. In these studies, previous understanding was applied to taper off dependent variables. In this regard these studies actually evaluated previous understanding. In summary, it is claimed that this set of studies was meaningful in generating new understanding of multi-evaluator operation, although few new variables were identified.

7.5.2 Research Weakness and Strength

The focus of this research was on empirical studies of improving reliability and validity of WoZ. These studies yielded a rich data set (at the side of data sources) and offered systematic understanding of these variables and their impact. In particular on the reliability and validity of WoZ, a key achievement of this research was the incremental work of system operation evaluation, which resulted in systematic improvement of reliability and validity and operation system development. Furthermore, if taking all these studies together, this could be seen as comparative studies over long term, that was rarely carried out in previous WoZ research. The comparative review of study 4, 5 and 6 demonstrated routes for the improvement of the reliability and validity of WoZ.

However, a number of limitations were acknowledged regarding the empirical work of design and evaluation. The studies described in this thesis were exploratory, and it acknowledged that some guidance that was extracted from study understanding was specified to natural-language dialogue systems. For example, the guidance for schema design was limited in system operations of speech recognition. This was a natural consequence of performing interaction method-specified WoZ studies. However, the guidance may be also applicable to other interaction systems by re-planning evaluator's system operation.

7.5.3 Challenges

Although the findings provided good knowledge to address these questions, there were still some challenges remained in the research. These were described as follows.

Firstly, it was acknowledged in the research that evaluators had initial differences in terms of background knowledge and speech recognition capability. Previous studies assumed that these differences could be covered by intense training, while in the case studies presented in this thesis; it showed that even after several intense trainings there were still some noticeable differences between evaluators. The risks of inconsistent operation still existed when evaluators' had mis-recognition of speech.

Secondly, evaluators were not limited to provide real-time system operation, regardless of how the operation system was designed and how intense training was provided. The difference of operation effectiveness was observed across system applications. The multimedia application received most accurate and consistent system operation, as this application did not need speech recognition. In terms of speech recognition operation, system response latency was commonly observed throughout the research, although it did not affect the overall system interactions in this research.

7.6 Future Work

This section describes avenues for future research work, including outlines of promising directions for empirical work, the perspectives of the reliability and validity of WoZ, and the need for developing and improving methods for data collection and analysis for WoZ study. This research mainly concerned with two avenues for WoZ study, including methodological study of WoZ's reliability and validity and practical development work of WoZ operation system and smart devices for domestic communication. Since WoZ was a flexible, and highly dynamic design-evaluation combined methodology, there showed a clear need for following-up studies along both routes. Assuming sufficient supports of study resources, the 'wish list' for the following study is laid out as following.

Firstly, throughout the research, the main interactions were natural languages and the cube. It was claimed that natural interactions were essential for ambient intelligent, and that was also of great importance to domestic communication. The study would be carried out using more technologies to support natural interaction, such as gesture recognition. This could extent the scope of research to a broad HCI area. In this regard, more system applications that represent natural interaction would be developed.

Secondly, one general criticism that can be levelled at most WoZ studies to date, including this research, was that the study focused on one subject at a time. In real circumstances there was a huge possibility of multiple subjects using the system simultaneously. The author calls for the increase of attention to the need for multiple subjects – people who interact with the system via multiple communication channels simultaneously. it was envisaged that such simultaneous system use with multiple subjects will have different needs and problems to WoZ's reliability and validity, as well as to WoZ operation system design.

And thirdly, the author would reiterate the importance of data collection and analysis methods. Without these methods it was not clear whether the variables were important or not. In particular, the author aims to improve the video analysis method for more efficient video footage transcription and analysis. Such improvement is envisaged that researchers' video recording and analysis work could be achieved through semi-automatic progresses.

In summary, as in the area of WoZ studies, the design is never-ending. For example, a number of comments on smart device's functions were made by study subjects. The most common request for following design would be to make more widgets personalised for specific users. In terms of WoZ study, the author hypothesises that such 'personalised widgets' for domestic communication study may find some interesting further results.

Acknowledgements

I am grateful to my supervisor Dr John V. H. Bonner whose rigorous supervision helped me throughout the Ph.D. I have benefited a lot from that, and am indebted to him for the advice, support and reassurance.

Also, it is worth to mention my friends who participated in the research and provided valuable feedbacks, to name a few, Jing Wang, Qian Xu, Munir Naveed, Shahin Sha and T. Lee.

Lastly I appreciate the University of Huddersfield for providing the fee waiver bursary for my Ph.D study.

References

[1] Aarts, E. 2004. Ambient intelligence: a multimedia perspective. *Multimedia, IEEE 11*, 12-19.

[2] ACM 1993. Back to real world. *Special issue on computer-augmented enviroinments and communicatin 36*.

[3] Anderson, B., McWilliam, A., Lacohee, H., Clucas, E. and Gershuny, J. 1999. Family Life in the Digital Home - Domestic Telecommunications at the End of the 20th Century Kluwer Academic Publishers, 85-97.

[4] Andersson, G., Hook, K., Mourao, D., Paiva, A. and Costa, M. 2002. Using a Wizard of Oz study to inform the design of SenToy. In *Proceedings of the Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, London, England2002 ACM.

[5] Aoki, K. and Downes, E.J. 2003. An analysis of young people's use of and attitudes towards cell phones. *Telematics and Informatics 20*, 349-364.

[6] Arnold, M. 2004. The connected home: Probing the effects and affects of domesticated ITCs. In *The Eighth Biennial Participatory Design Conference* Adrian Bond, Toronto, Ontario Canada.

[7] Augusto, J.C. and McCullagh, P. 2007. Ambient intelligence: Concepts and applications. *International journal of Computer Science and Information Systems 4*, 1-28.

[8] Bailey, B.P., Biehl, J.T., Cook, D.J. and Metcalf, H.E. 2008. Adapting paper prototyping for designing user interfaces for multiple display environments Springer-Verlag, 269-277.

[9] Balbo, S., Coutaz, J. and Salber, D. 1993. Towards automatic evaluation of multimodal user interfaces. In *Proceedings of the Proceedings of the 1st international conference on Intelligent user interfaces*, Orlando, Florida, United States1993 ACM.

[10] Bardram, E. 2005. The trouble with login: on usability and computer security in ubiquitous computing Springer-Verlag, 357-367.

[11] Barliner, A., Fischer, K., Huber, R., J.Spilker and E.Noth 2003. How to find trouble in communication. *Speech Communication 40*, 117-143.

[12] Baud, F.L. and Denslow, W. 1900. *The wizard of oz*. Random House, New York.

[13] Bernhaupt, R., Obrist, M., Weiss, A., Beck, E. and Tscheligi, M. 2008. Trends in the living room and beyond: results from ethnographic studies using creative and playful probing ACM, 1-23.

[14] Bickham, D.S. 2006. Is television viewing associated with social isolation: roles of exposure time, viewing context, and voilent content. *Achieves of Pediatrics & Adolescent Medicine 160*, 387-392.

[15] Bickmore, T. 2002. Towards the design of multimodal interfaces for handheld conversational characters. In *Proceedings of the CHI '02 extended abstracts on Human factors in computing systems*, Minneapolis, Minnesota, USA2002 ACM.

[16] Billinghurst, M. and Kato, H. 1999. Coolaberative mixed reality. In *International Symposium on Mixed Reality(ISMR'99)*, Yokohama,Japan, 261-284.

[17] Blythe, M. and Monk, A. 2002. Notes towards an ethnography of domestic technology. In *Proceedings of the Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, London, England2002 ACM.

[18] Bohn, J., Coroama, V., Langheinrich, M. and Mattern, F. 2004. Disappearing computers everywhere - Living in a world of smart everyday objects. *Human and Ecological Risk Assessment 10*, 763-785.

[19] Bonner, J.V.H., Li, A.X. and Robinson, J. 2009. Designing and evaluating smart domestic technologies which use infrequent interaction. In *The Eighth International Conference on Pervasive Computing*, Nara,Japan.

[20] Bonnie, E.J. 2004. Beyond the UI: Product, process and passion. *NordiCHI 10*.

[21] Bonnie, J. 1995. Why GOMS? ACM, 80-89.

[22] Bowden, S. and Offer, A. 1994. Household appliances and the use of time: the United States and Britain since the 1920s. *Economic History Review 47*, 725-748.

[23] Bradley, J., Mival, O. and Benyon, D. 2009. Wizard of Oz experiments for companions. In *Proceedings of the Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, Cambridge, United Kingdom2009 British Computer Society.

[24] Bretan, I., Ereback, A.-L., MacDermid, C. and Waern, A. 1995. Simulation-based dialogue design for speech-controlled telephone services. In *Proceedings of the Conference companion on Human factors in computing systems*, Denver, Colorado, United States1995 ACM.

[25] Bruns, F.W. 2006. Ubiquitous computing and interaction. *Annual Reviews in Control 30*, 205-213.

[26] Buisine, S., Martin, J.-c. and Bernsen, N.O. 2005. Children's Gesture and Speech in Conversation with 3D Characters. In *The HCI International 2005*, Lasvegas, NV, USA.

[27] Bunn, F., Byrne, G. and Kendall, S. 2009. Telephone consultation and triage: effects on health care use and patient satisfation. *Cochrane Database of Systematic Reviews*.

[28] Cairns, P. and Cox, A.L. 2008. *Research methods for Human-Computer Interaction*. Cambridge University Press.

[29] Carbini, S., Delphin-Poulat, L., Perron, L. and Viallet, J.E. 2006. From a Wizard of Oz experiment to a real time speech and gesture multimodal interface. *Signal Processing 86*, 3559-3577.

[30] Card, S.K., Moran, T.P. and Newell, A. 1980. Computer text-editing: an information-processing analysis of a routine cogniti. *Cognitive Psychology 12*, 32-74.

[31] Carroll, J.M. 2000. *Making use. Scenario-based design of Human-Computer Interactions*. MIT Press.

[32] Carroll, J.M., Kellogg, W.A. and Rosson, M.B. 1991. *The Task-Artifact cycle, in Designing interaction: Psychology at the Human-Computer Interface*. Cambridge University Press, Cambridge.

[33] Chaparro, B.S., Hinkle, V.D. and Riley, S.K. 2008. The Usability of Computerized Card Sorting: A Comparison of Three Applications by Researchers and End Users. *Journal of Usability Studies 4*, 31-48.

[34] Clarizio, G., Mazzotta, I., Novielli, N. and de Rosis, F. 2006. Social Attitude Towards A Conversational Character. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, 2-7.

[35] Craft, B. and Cairns, P. 2009. Sketching sketching: outlines of a collaborative design method. In *Proceedings of the Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, Cambridge, United Kingdom2009 British Computer Society.

[36] Dahlback, N., Jonsson, A. and Ahrenberg, L. 1993. Wizard of Oz studies: why and how. In *Proceedings of the Proceedings of the 1st international conference on Intelligent user interfaces*, Orlando, Florida, United States1993 ACM.

[37] Davis, F.D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly 13*, 319-340.

[38] DEMIRIS, G., SKUBIC, M., KELLER, J., RANTZ, M.J., OLIVER, D.P., AUD, M.A., LEE, J., BURKS, K. and GREEN, N. 2006. Nurse Participation in the design of user interface for a smart home system. *Smart Home and Beyond*.

[39] Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J.D. and Gandy, M. 2005. Wizard of Oz Support throughout an Iterative Design Process IEEE Educational Activities Department, 18-26.

[40] Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J.D. and Gandy, M. 2005. Wizard of oz support throughout an iterative design process. In *IEEE Pervasive Computing*, 18-26.

[41] Dow, S.P., Mehta, M., MacIntyre, B. and Mateas, M. 2010. Eliza meets the wizard-of-oz: blending machine and human control of embodied characters. In *Proceedings of the Proceedings of the 28th international conference on Human factors in computing systems*, Atlanta, Georgia, USA2010 ACM.

[42] Edlund, J., Gustafson, J., Heldner, M. and Hjalmarsson, A. 2008. Towards human-like spoken dialogue systems. *Speech Communication 50*, 630-645.

[43] Edlund, J., Gustafson, J., Heldner, M. and Hjalmarsson, A. 2008. Towards human-like spoken dialogue systems. *Speech Communication 50*, 630-645.

[44] Edwards, W.K. and Grinter, R.E. 2001. At home with ubiquitous computing: Seven challenges. In *Ubicomp 2001: Ubiquitous Computing* Springer Berlin / Heidelberg, 256-272.

[45] Ellen, I., Candace, K., Diane, J.S., Alan, W. and Steve, W. 2002. Characterizing instant messaging from recorded logs. In *Proceedings of the CHI '02 extended abstracts on Human factors in computing systems*, Minneapolis, Minnesota, USA2002 ACM.

[46] Elliot, K., Neustaedter, C. and Greenberg, S. 2005. Time, ownership and awareness: The value of contextual locations in the home. In *UNICOMP 2005: Ubiquitous computing (the 7th International Conference on Ubiquitous Computing)*, M. BEIGL, S. INTILLE, J. REKIMOTO AND H. TOKUDA Eds. Springer, Tokyo, Japan.

[47] Firth, L. and Mellor, D. 2005. Broadband: benefits and problems. *Telecommunications Policy 29*, 223-236.

[48] Forbes-Riley, K. and Litman, D. 2009. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language In press*.

[49] Forbes-Riley, K. and Litman, D. 2011. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language 25*, 105-126.

[50] Fraser, N.M. and Gilbert, G.N. 1991. Simulating speech systems. *Computer Speech and Language 5*, 81-99.

[51] Friedewald, M. and Raabe, O. 2011. Ubiquitous computing: An overview of technology impacts. *Telematics and Informatics 28*, 55-65.

[52] Gamm, S., Haeb-Umbach, R. and Langmann, D. 1997. The development of a command-based speech interface for a telephone answering machine. *Speech Communication 23*, 161-171.

[53] Gaver, B., Dunne, T. and Pacenti, E. 1999. Design: Cultural probes ACM, 21-29.

[54] Goldman, N., Lin, I.-F., Weinstein, M. and Lin, Y.-H. 2003. Evaluating the quality of self-reports of hyperthension and diabetes. *Journal of Clinical Epidemiology 56*, 148-154.

[55] Gould, J.D., Conti, J. and Hovanyecz, T. 1983. Composing letters with a simulated listening typewriter ACM, 295-308.

[56] Grinter, R. and Eldridge, M. 2003. Wan2tlk?: everyday text messaging. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems*, Ft. Lauderdale, Florida, USA2003 ACM.

[57] Harper, R. 2003. *Inside the smart home*. Springer-Verlag, New York.

[58] Hauptmann, A.G. 1989. Speech and gestures for graphic image manipulation. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*1989 ACM.

[59] Hawes, N., Wyatt, J. and Sloman, A. 2009. Exploring design space for an integrated intelligent system. *Knowledge-Based Systems 22*, 509-515.

[60] Helander, M.G., Landauer, T.K. and Prabhu, P.V. 1997. Handbook of Human-Computer Interaction. In *Designing for quality in use* Elsevier Science, Eaglewood Cliffs, N.J.

[61] Henrysson, A., Billinghurst, M. and Ollila, M. 2005. Face to Face Collaborative AR on Mobile Phones. In *Proceedings of the Proceedings of the 4th IEEE/ACM International Symposium on Mixed and Augmented Reality*2005 IEEE Computer Society.

[62] Hewett, T., Baecher, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Srong, G. and Verplank, W. 1996. ACM SIGCHI curricula for human-computer interaction. ACM SIGCHI

[63] Hjorthol, R. and Gripsrud, M. 2009. Home as a communication hub: the domestic use of ICT. *Journal of Transport Geography 17*, 115-123.

[64] Hong, J.-C., Hwang, M.-Y., Hsu, H.-F., Wong, W.-T. and Chen, M.-Y. 2011. Applying the technology acceptance model in a study of the factors affecting usage of the Taiwan digital archieves system. *Computers & Education 54*, 2086-2094.

[65] Horrigan, J. 2009. Home broadband adoption 2009. Pew Internet & American Life Project

[66] Hoysniemi, J., Hamalainen, P. and Turkki, L. 2004. Wizard of Oz prototyping of computer vision based action games for children. In *Proceedings of the Proceedings of the 2004 conference on Interaction design and children: building a community*, Maryland2004 ACM.

[67] Hsu, C.-L., Yang, S.-Y. and Wu, W.-B. 2010. 3C intelligent home appliance control system - Example with refrigerator. *Expert Systems with Applications 37*, 4337-4349.

[68] Huart, J., Kolski, C. and Sagar, M. 2004. Evaluation of multimedia applications using inspection methods: the Cognitive Walkthrough case. *Interacting with Computers 16*, 183-215.

[69] Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. and Yang, J. 2003. Predicting human interruptibility with sensors: a Wizard of Oz feasibility study. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems*, Ft. Lauderdale, Florida, USA2003 ACM.

[70] Hudson, W. 2007. Old cards, new tricks: applied techniques in card sorting. In *Proceedings of the Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - Volume 2*, University of Lancaster, United Kingdom2007 British Computer Society.

[71] Hughes, J., Brien, J.O., Rodden, T. and Rouncefield, M. 1997. Ethnography, communication and support for design. *Workplace Studies*.

[72] Hughes, J., King, V., Rodden, T. and Andersen, H. 1994. Moving out from the control room: ethnography in system design. In *The 1994 ACM conference on Computer supported cooperative work* ACM, Chapel Hill, North Carolina, United States

[73] Hughes, J., O'Brien, J., Rodde, T., Rouncefield, M. and Viller, S. 2000. Patterns of home life: Informing design for domestic environments. *Personal and Ubiquitous Computing 4*, 25-38.

[74] Humberto T. Marques, N., Leonardo, C.D.R., Pedro, H.C.G., Jussara, M.A., Wagner Meira, Jr. and Virgilio, A.F.A. 2004. Characterizing broadband user behavior. In *Proceedings of the Proceedings of the 2004 ACM workshop on Next-generation residential broadband challenges*2004 ACM.

[75] Hutchinson, H., Mackay, W., Westerlund, B., B.Bederson, B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., Hansen, H., Roussel, N., Eiderback, B., Lindquist, S. and Sundblad, Y. 2003. Techonology probes: Inspiring design for and with families. *CHI*, 17-24.

[76] Irani, L., Jeffries, R. and knight, A. 2010. Rhythms and plasticity: television temporality at home. *Personal and Ubuquitous Computing 14*, 621-632.

[77] Kantner, L. and Rosenbaum, S. 1997. Usability studies of WWW sites: heuristic evaluation vs. laboratory testing. In *Proceedings of the Proceedings of the 15th annual international conference on Computer documentation*, Salt Lake City, Utah, United States1997 ACM.

[78] Kirstein, M., Burney, K., Paxton, M. and Bergstrom, E. 2001. Moving towards broadband ubiquity in U.S. business markets.

[79] Kjeldskov, J. and Graham, C. 2003. A review of mobile HCI research mothods. In *Mobile HCI*, L. CHITTARO Ed. Springer-Verlag, 317-335.

[80] Kline, R. 2003. Resisting Consumer Technology in Rural America: The Telephone and Electrification. In *How Users Matter: The Co-Construction of Users and Technology*, N. OUDSHOORN AND T. PINCH Eds. MIT Press, Cambridge, 51-66.

[81] Kraut, R. and Kiesler, S. 2003. The social impact of Internet use. *Psychology Science Agenda*, 8-10.

[82] Kubey, R. and Csikszentmihalyi, M. 1990. *Television and the quality of life: how viewing shapes everyday experience*. Lawrence Erlbaum Associates Inc., Hillsdale, New Jersey,USA.

[83] Lacohée, H. and Anderson, B. 2001. Interacting with the telephone. *International Journal of Human-Computer Studies 54*, 665-699.

[84] Lagasse, P. and Moermann, I. 2005. Broadband communication. In *True visions: The emergence of ambient intelligence*, E. AARTS AND J.L. ENCARNACAO Eds. Springer, Berlin, Heidelberg, 185-207.

[85] Lapides, P., Sharlin, E. and Greenberg, S. 2009. HomeWindow: an augmented reality domestic monitor. In *Proceedings of the Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, La Jolla, California, USA2009 ACM.

[86] Lee, H., Wu, C. and Aghajan, H. 2011. Vision-based user-centric light control for smart environments. *Pervasive and Mobile Computing 7*, 223-240.

[87] Lewis, C. and Wharton, C. 1997. Cognitive Walkthroughs. In *Handbook of Human-Computer Interaction (Second Edition)*, 717-732.

[88] Li, A.X. and Bonner, J. 2010. Enhancing social relationships through human-like intelligences. In *The 9th International Workshop on Social intelligence Design (SID2010)*, Egham, UK.

[89] Li, A.X. and Bonner, J.V.H. 2011. Improving control panel consistency of wizard-of-oz design and evaluation studies. In *the 17th International Conference on Automation & Computing*, W.-H. CHEN, X. CHEN AND Z. XU Eds., Huddersfield, UK, 163-168.

[90] Li, H. and Greenspan, M. 2011. Model-based segmentation and recogntion of dynamic gestures in continuous video streams. *Pattern Recognition 44*, 1614-1628.

[91] Lindgaard, G. 1994. *Usability testing and system evaluation: A guide for designing useful computer systems*. CHAPMAN & HALL, London, Glasgow, New York, Tokyo, Melbourne, Madras.

[92] Ling, R. 2004. *The mobile connection: The cell phone's impact on society*. Elsevier, San Francisco.

[93] Ling, R. and Yttri, B. 2002. Hyper-coordination via mobile phones in Norway. In *Perpetual contact: Mobile communication, private talk, public performance*, J. KATZ AND M. JAAKHUS Eds. Cambridge University Press, Cambridge.

[94] Liu, C., Rieser, V. and Lemon, O. 2009. A Wizard-of-Oz interface to study Information Presentation strategies for Spoken Dialogue Systems In *The 1st International Workshop on Spoken Dialogue Systems*.

[95] Lo, S.-K. and Lie, T. 2008. Selection of communication technologies--A perspective based on information richness theory and trust. *Technovation 28*, 146-153.

[96] Mann, S. 1996. "Smart clothing": Wearable multimedia computing and "personal imaging" to restore the technological balance between people and their environments. In *The forth ACM international conference on multimedia* ACM Press, New York, 163-174.

[97] Maulsby, D., Greenberg, S. and Mander, R. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, Amsterdam, The Netherlands1993 ACM.

[98] Mavrikis, M. and Gutierrez-Santos, S. 2010. Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. *Computers & Education 54*, 641-651.

[99] Milgram, P. and Kishino, F. 1994. A Taxonomy of Mixed Reality Visual Displays. *Special Issue on Networked Reality E77-D*, 1321-1329.

[100] Moira, B., Cameron, M. and Thomas, L. 2010. Social network activity and social well-being. In *Proceedings of the Proceedings of the 28th international conference on Human factors in computing systems*, Atlanta, Georgia, USA2010 ACM.

[101] Molin, L. 2004. Wizard-of-Oz prototyping for co-operative interaction design of graphical user interfaces. In *Proceedings of the Proceedings of the third Nordic conference on Human-computer interaction*, Tampere, Finland2004 ACM.

[102] Morley, D. and Silverstone, R. 1990. Domestic communication - technologies and meanings. *Media, culture and society 12*, 31-55.

[103] Newman, W.M. and Lamming, M. 1995. *Interactive system design*. Addison-Wesley Longman Publishing Co.,Inc.,.

[104] Nielsen, J. 1994. Usability inspection methods. In *Proceedings of the Conference companion on Human factors in computing systems*, Boston, Massachusetts, United States1994 ACM.

[105] Nielsen, J. and Molich, R. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people*, Seattle, Washington, United States1990 ACM.

[106] Noble, G. 1987. Individual differences, psychological neighbourhoods and use of the domestic telephone. *Media Informaiton Austrialia 44*, 37-41.

[107] Novielli, N., de Rosis, F. and Mazzotta, I. 2010. User attitude towards an embodied conversational agent: Effects of the interaction mode. *Journal of Pragmatics 42*, 2385-2397.

[108] O'Halloran, K., Tan, S., Smith, B. and Podlasov, A. 2010. Challenges in designing digital interfaces for the study of multimodal phenomena. *Information Design Journal 18*, 2-21.

[109] Oviatt, S., Darves, C. and Coulston, R. 2004. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas ACM, 300-328.

[110] Paiva, A., Costa, M., Chaves, R., Piedade, M., Mourao, D., Sobral, D., Hoo k, K., Andersson, G. and Bullock, A. 2003. SenToy: an affective sympathetic interface. *International Journal of Human-Computer Studies 59*, 227-235.

[111] Palen, L. and Hughes, A. 2007. When home base is not a place: parents' use of mobile telephones Springer-Verlag, 339-348.

[112] Papacharissi, Z. and Zaks, A. 2006. Is broadband the future? An analysis of broadband technology potential and diffusion. *Telecommunications Policy 30*, 64-75.

[113] Plomp, J. and Tealdi, P. 2004. Ambient intelligent technologies for wellbeing at home. In *Proceedings of the Proceedings of the 2nd European Union symposium on Ambient intelligence*, Eindhoven, Netherlands2004 ACM.

[114] Purdy, J.M. and Nye, P. 2000. The impact of comunication media on negotiation outcomes. *The International Journal of Conflict Management 11*, 162-187.

[115] Ramey, J., Boren, T., Cuddihy, E., Dumas, J., Guan, Z., Haak, M.J.v.d. and Jong, M.D.T.D. 2006. Does think aloud work?: how do we know? In *Proceedings of the CHI '06 extended abstracts on Human factors in computing systems*, Montreal, Quebec, Canada2006 ACM.

[116] Ramos, C., Augusto, J.C. and Shapiro, D. 2008. Ambient intelligence - the next step for artificial intelligence. *Intelligent systems 23*, 15-18.

[117] Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M. and Eilola, T. 2003. Mind-reading ability: Beliefs and performance. *Journal of Research in Personality 37*, 420-445.

[118] Regenbrecht, H.T. and Wagner, M.T. 2002. Interaction in a collaborative augmented reality environment. In *Proceedings of the CHI '02 extended abstracts on Human factors in computing systems*, Minneapolis, Minnesota, USA2002 ACM.

[119] Rogers, E.M. 2003. Diffusion of innovations Free Publisher, New York.

[120] Rosis, F.d., Novielli, N., Carofiglio, V., Cavalluzzi, A. and Carolis, B.D. 2006. User modeling and adaptation in health promotion dialogs with an animated character. *Journal of Biomedical Informatics 39*, 514-531.

[121] Ruyter, B.d., Saini, P., Markopoulos, P. and Van Breemen, A.J.N. 2005. Assessing the effects of building social intelligence in a robotic interface for the home. *Interacting with Computers 17*, 522-541.

[122] Salber, D. and Coutaz, J. 1993. Applying the Wizard of Oz technique to the study of multimodal systems. In *Human-Computer Interaction* Springer Berlin/Heidelberg, Berlin/Heidelberg, 219-230.

[123] Schmandt, C. and Ackerman, M. 2004. Personal and Ubiquitous Computing: Issue on privacy and security Springer-Verlag, 389-390.

[124] Schutze, M., Sachse, P. and Romer, A. 2003. Support value of sketching in the design process. *Research in Engineering Design 14*, 89-97.

[125] Sengers, P., Boehner, K., David, S. and Kaye, J.J. 2005. Reflective design. In *Proceedings of the Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, Aarhus, Denmark2005 ACM.

[126] Shami, N.S., Jeffrey, T.H., Christian, P., Michael, M. and Regan, M. 2008. Measuring affect in hci: going beyond the individual. In *Proceedings of the CHI '08 extended abstracts on Human factors in computing systems*, Florence, Italy2008 ACM.

[127] Sharp, H., Rogers, Y. and Preece, J. 2006. *Interaction design: beyond human-computer interaction*. John Wiley & Sons,Ltd.

[128] Silverstone, R., Hirsch, E. and Morley, D. 1992. Information and communication technologies and the moral economy of the household. In *Consuming technologies: Media and information in domestic spaces*, R. SILVERSTONE AND E. HIRSCH Eds. Routeledge Press, London UK.

[129] Slavkovic, A. and Cross, K. 1999. Novice heuristic evaluations of a complex interface. In *Proceedings of the CHI '99 extended abstracts on Human factors in computing systems*, Pittsburgh, Pennsylvania1999 ACM.

[130] Snyder, C. 2003. *Paper prototyping: The fast and easy way to design and refine user interfaces*. Elsevier Science, San Francisco.

[131] Stewart, J. 2003. The social consumption of information and communication technologies (ICTs): insights from research on the appropriation and consumption of new ICTs in the domestic environment *Cognition, Technology & Work 5*, 4-14.

[132] Tamura, H., Yamamoto, H. and Katayama, A. 2001. Mixed Reality: Future Dreams Seen at the Border between Real and Virtual Worlds IEEE Computer Society Press, 64-70.

[133] Taylor, A.S., Harper, R., Swan, L., Izadi, S., Sellen, A. and Perry, M. 2007. Homes that make us smart. *Pers Ubiquit Comput*.

[134] Tscheligi, M., de Ruyter, B., Markopoulus, P., Wichert, R., Mirlacher, T., Meschterjakov, A., Reitberger, W., De Carolis, B., Mazzotta, I. and Novielli, N. 2009. NICA: Natural Interaction with a Caring Agent. In *Ambient Intelligence* Springer Berlin / Heidelberg, 159-163.

[135] Tsukahara, W. and Ward, N. 2001. Responding to subtle, fleeting changes in the user's internal state. In *The SIG-CHI on Human Factors in Computing Systems* ACM, Seattle, WA., 77-84.

[136] Vall, N.D. 1988. *Domestic technology: a chronology of developments*. G.K. Hall, Boston, Mass.

[137] Venkatesh, A. 2001. The home of the future: An ethnographic study of new information technologies in the home. In *Advances in consumer research volume XXVIII*, M. GILLY AND J. MEYERS-LEVY Eds. Association for Consumer Research, Valdosta,Georgia, 88-96.

[138] Vetere, F., Gibbs, M.R., Kjeldskov, J., Howard, S., Mueller, F.F., Pedell, S., Mecoles, K. and Bunyan, M. 2005. Mediating intimacy: Designing technologies to support Strong-Tie relationships. *CHI 2*, 471-480.

[139] Vitalari, N.P., Venkatesh, A. and Gronhaug, K. 1985. Computing in the home: shifts in the time allocation patterns of households. *Communications of the ACM 28*, 512-522.

[140] Voida, S., Podlaseck, M., Kjeldsen, R. and Pinhanez, C. 2005. A study on the manipulation of 2D objects in a projector/camera-based augmented reality environment. In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems*, Portland, Oregon, USA2005 ACM.

[141] Wania, C.E., Atwood, M.E. and McCain, K.W. 2006. How do design and evaluation interrelate in HCI research? In *Proceedings of the Proceedings of the 6th conference on Designing Interactive systems*, University Park, PA, USA2006 ACM.

[142] Weiser, M. 1991. The computer for the 21st century. *Scientific American 265*, 94-104.

[143] Weiser, M. 1993. Some computer science issues in ubiquitous computing. *Commun.ACM 36*, 75-85.

[144] Wharton, C., Rieman, J., Lewis, C. and Polson, P. 1994. The cognitive walkthrough method: a practitioner's guide. In *Usability inspection methods* John Wiley \& Sons, Inc., 105-140.

[145] Whittaker, S. and Sidner, C. 1996. Email overload: Exploring personal information management of email. In *CHI 96 Conference on Human Factors in Computing Systems* ACM Press, 276-283.

[146] Whittaker, S., Walker, M. and Moore, J. 2002. Fish or fowl: A wizard of oz evaluation of dialogue strategies in the restaurant domain. In *Proceedings of the Language Resources and Evaluation Conference*2002.

[147] Winograd, T. 1997. From computing machinery to interaction design. In *Beyond calculation: the next fifty years of computing*, P. DENNING AND R. METCALFE Eds. Springer-Verlag, Amsterdam, 149-162.

[148] Wirén, M., Eklund, R., Engberg, F. and Westermark, J. 2007. Experiences of an in-service Wizard-of-Oz data collection for the deployment of a call-routing application. In *Proceedings of the Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, Rochester, New York2007 Association for Computational Linguistics.

[149] Wooffitt, R., Fraser, N.M., Gilbert, N. and McGlashan, S. 1997. *Humans, computers and wizards: Analysing human (simulated) computer interaction*. Routledge, London.

[150] Xu, Y., Ueda, K., Komatsu, T., Okadome, T., Hattori, T., Sumi, Y. and Nishida, T. 2009. WOZ experiments for understanding mutual adaptation. *AI & Society 23*, 201-212.

[151] Yang, B. and Lugger, M. 2010. Emotion recognition from speech signals using new harmony features. *Signal Processing 90*, 1415-1423.

[152] Yildirim, S., Narayanan, S. and Potamianos, A. 2011. Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language 25*, 29-44.

[153] Zhang, T., Hasegawa-Johnson, M. and Levinson, S.E. 2006. Cognitive state classification in a spoken tutorial dialogue system. *Speech Communication 48*, 616-632.

Appendix 1.1

Definitions of Key Terms

Reliability and validity of WoZ – The reliability of WoZ is defined as the extent to which the evaluator could generate consistent operations throughout WoZ studies. The validity of WoZ is defined as the degree of how evaluator could follow the schema and return expected operation.

Schema – This refers to a set of pre-planned rules which are used to guide evaluator to generate proper operations to subject activities.

Rigorousness of schema design – This refers to the extent to how strict and specific the schema rules are planned to guide evaluator's operation.

Control panel – This refers to the operation tool that is provided to evaluator to generate operations.

Evaluator management and control – This refers to evaluator's overall activities that aim to facilitate subject's activities by using the control panel.

Domestic communication – This refers to the interaction between human and smart devices in the home, and its primary focus is on the communication between humans.

Domestic communication technologies – This is defined as the interactive technologies that are designed in the home to assist the communication between humans and smart systems.

Smart devices – This is defined as systems that provide human-like intelligence and interact with humans in natural way, such as natural-language dialogue system.

Smart homes – This is defined as domestic environment which is equipped with context-aware smart devices.

Rigorousness level - This is defined as the degree of how specific the schema rules are designed to guide evaluator's system operation. The higher rigorousness level, the more specific the schema rules are designed, and the less operation flexibility the evaluator has.

WoZ Components – This refers to the general elements that consist of WoZ studies. Of these used in this research, these included schemas, control panels, tools for subject's system interactions, and the laboratory setting of the study.

Proactive prompts – This is defined as a type of operation that the evaluator anticipates subject's anticipations and makes advance prompts.

Improvisational operations – This is defined as those operations that are made by evaluator's improvisational judgements instead of schemas.

Synchronisation between evaluators – This is defined as the communication between evaluators that aims to set up a way of consistent system management and control.

Natural interaction – In this research this is defined as the interaction supported by a system that takes natural human speech and gestures as an effective interaction way.

Appendix 3.1

Application to University Ethical Committee

UNIVERSITY OF HUDDERSFIELD

SCHOOL OF COMPUTING AND ENGINEERING

PROJECT ETHICAL REVIEW FORM

Applicable for all research, masters and undergraduate projects

Project Title:	REALITIES IN THE MAKING: USING WIZARD-OF-OZ METHODOLOGY FOR DOMESTIC COMMUNICATION STUDIES
Student:	XIANGDONG LI
Course/Programme:	REALITIES IN THE MAKING: USING WIZARD-OF-OZ METHODOLOGY FOR DOMESTIC COMMUNICATION STUDIES
Department:	INFORMATICS
Supervisor:	JOHN BONNER
Project Start Date:	OCT/2008

ETHICAL REVIEW CHECKLIST

		Yes	No
1.	Are there problems with any participant's right to remain anonymous?	☐	☒
2.	Could a conflict of interest arise between a collaborating partner or funding source and the potential outcomes of the research, e.g. due to the need for confidentiality?	☐	☒
3.	Will financial inducements be offered?	☐	☒
4.	Will deception of participants be necessary during the research?	☒	☐
5.	Does the research involve experimentation on any of the following?		

	(i) animals?	☐	☒
	(ii) animal tissues?	☐	☒
	(iii) human tissues (including blood, fluid, skin, cell lines)?	☐	☒
6.	Does the research involve participants who may be particularly vulnerable, e.g. children or adults with severe learning disabilities?	☐	☒
7.	Could the research induce psychological stress or anxiety for the participants beyond that encountered in normal life?	☐	☒
8.	Is it likely that the research will put any of the following at risk:		
	(i) living creatures?	☐	☒
	(ii) stakeholders (disregarding health and safety, which is covered by Q9)?	☐	☒
	(iii) the environment?	☐	☒
	(iv) the economy?	☐	☒
9.	Having completed a health and safety risk assessment form and taken all reasonable practicable steps to minimise risk from the hazards identified, are the residual risks acceptable (Please attach a risk assessment form)	☒	☐

STATEMENT OF ETHICAL ISSUES AND ACTIONS

If the answer to any of the questions above is yes, or there are any other ethical issues that arise that are not covered by the checklist, then please give a summary of the ethical issues and the action that will be taken to address these in the box below. If you believe there to be no ethical issues, please enter “NONE”.

The ethical issues are concerned in aspects of:

1) Collecting participants’ publicly personal information, the information to collect includes weekly or daily appointments, hobbies and interests, and personal habits and preferences in terms of news reading, pet keeping, driving and the use of some domestic appliances.

2) Participant deceptions. The project hides some system components and functions from participants before and during the experiments. In wizard-of-oz experiments the system is simulated by an experimenter actually, while participants are presented apparent illusions without being told the truth until the experimental procedures are done.

Actions taken to address these issues:

1) Formal invitations are sent to participants before the experiment.

2) The invitation contains clear requirements of personal information, and further procedures we are going to do with the information.

3) A consent form is provided after revealing the experiment system truth, it promises the information collected for the experiment will be confidential and for research use only and will not be used for any other purpose without the participant's permissions. If participants do not agree with the consent form then all relevant data collected from the experiment will be destroyed.

STATEMENT BY THE STUDENT

I believe that the information I have given in this form on ethical issues is correct.

Signature: Xiangdong Li Date: 15/12/2010

AFFIRMATION BY THE SUPERVISOR

I have read this Ethical Review Checklist and I can confirm that, to the best of my understanding, the information presented by the student is correct and appropriate to allow an informed judgement on whether further ethical approval is required.

Signature: John Bonner Date: 15/12/2010

SUPERVISOR RECOMMENDATION ON THE PROJECT'S ETHICAL STATUS

Having satisfied myself of the accuracy of the project ethical statement, I believe that the appropriate action is:

The project proceeds in its present form	
The project proposal needs further assessment by an Ethical Review Panel. The Supervisor will pass the form to the Ethical Review Panel Leader for consideration.	

RETENTION OF THIS FORM

- The Supervisor must retain a copy of this form until the project report/dissertation is produced.
- The student must include a copy of the form as an appendix in the report/dissertation.

OUTCOME OF THE ETHICAL REVIEW PANEL PROCESS, <u>WHERE REQUIRED</u> **Tick One**

1. Approved. The ethical issues have been adequately addressed and the project may commence. ☒

2. Approved subject to minor amendments. The required amendments are stated in the box below. The project may proceed once the form has been amended in line with the requirements and signed by the Supervisor in the box immediately below to confirm this. ☐

I confirm, as Supervisor, that the amendments required have been made:

Signature: _____ Date: _____

3. Resubmit. The areas requiring further action are stated in the box below. The project may not proceed until the form has been resubmitted and approved. ☐

4. Reject. The reasons why it will not be possible to address the ethical issues adequately are stated in the box below. ☐

For any of the outcomes 2, 3 or 4 above, please provide a statement in the box below.

AFFIRMATION BY THE REVIEW PANEL LEADER

I approve the decision reached above by the review panel members:

Signature:	Zhijie Xu	Date:	15/12/2010

Appendix 4.1

Study 1 – Training Words and Questions for System Speech Recognition

Training Words	System Operation Commands	Questions
Seaside	System start	My name is ***
Thunder	New user	What am I wearing?
Huddersfield	System error	Will I be rich when I am older?
System	Repeat	
Andol		

Appendix 4.2

Study 1 – Three Schemas with Different Rigorousness Levels

1. High Rigorousness Level Schema

User activities	Evaluator response
Seaside Thunder Huddersfield System Andol	▪ Recognising the word ▪ Typing the word ▪ No extra words/expressions were allowed ▪ If the word was not recognised, typing system message 'please repeat slowly'
System start New user System error Repeat	▪ Recognising the command ▪ Typing the command in text as it was said ▪ No extra words/expressions were allowed ▪ If the word was not recognised, typing system message 'unknown command, please repeat'
My name is *** What am I wearing? Will I be rich when I am older?	▪ Through the surveillance video, recognising the subject who gave the question ▪ If the subject was recognisable, then typing the name, such like 'Steve'; if the subject was unrecognisable, then typing system message 'sorry, out of system capability' ▪ If the cloth was recognisable, then typing the cloth in the format of 'cloth colour + cloth style', such like

	'black top'; if the cloth was not recognisable, then typing system message 'sorry, out of system capability' ▪ To answer questions like 'will I be rich when I am older', if previous dialogues mentioned that before, then only gave straightforward answer like 'yes' and 'no' according to contextual information. ▪ If such questions were asked without any clues, then typing system message 'sorry, out of system capability'.
Other questions	▪ checking if the question asked was relevant with the subject itself, if yes, then giving straightforward answer with 'yes' or 'no'; otherwise, typing system message 'sorry, out of system capability'.

2. General Rigorousness Level Schema

User activities	**Evaluator responses**
Seaside Thunder Huddersfield System Andol	▪ Recognising the word ▪ Typing the word ▪ No extra words/expressions were allowed ▪ If the word was not recognised, then giving a system message either asking for slow repeating or unrecognised word ▪ If not recognisable, guessing a word that the evaluator thought it might be right ▪ It was allowed to make some minor errors in this part, for example, mistaking a word to another one that had similar

	pronoun, thus triggering more conversations
System commands	▪ If the command was recognisable, then following the command ▪ If the command was not recognisable, three options were given to the evaluator: 1) typing unrecognised command system message, 2) prompting a dialogue to confirm the command in forms of 'are you sure to create new user?', and 3) no response and waiting for the subject's second command
My name is *** What am I wearing? Will I be rich when I am older?	▪ Recognising the subject who asked the question ▪ If the subject was identified, then giving answers that the evaluator thought it was suitable with the context ▪ If the question was not answerable, then giving the subject proper information to cover this issue. The evaluator could use human-human interaction style to facilitate speech dialogues
Other questions	▪ Based on the evaluator's personal understanding, it was encouraged to return a proper answer that would satisfy or amuse the subject

3. Combined Rigorousness Level Schema

User activities	**Evaluator responses**
Seaside Thunder Huddersfield System	▪ Recognising the given word ▪ Typing the recognised word strictly as it was given ▪ No proactive prompt was allowed

Andol	
System commands	▪ Recognising system commands ▪ Conducting relevant instructions to carry out the command ▪ If not recognisable, prompting the system message 'sorry, speech not recognised, do you mean *** (e.g. delete this user?)' ▪ Low-level anticipation was allowed, but this should be based on the conversation contexts
My name is *** What am I wearing? Will I be rich when I am older?	▪ Recognising the subject who asked the question ▪ If the subject was identified, then giving straightforward answer with 'yes' and 'no', plus customised expressions were allowed to be added to the answer. ▪ If the subject was unknown, then following the strict schema rules.
Other questions	▪ Checking if the question asked was relevant with the contexts of the recognised subject. ▪ If it was, then giving additional information with the short answer. If it was not, then giving straightforward answer using 'yes' and 'no'.

Appendix 4.3

Study 1 – Template of Video Footage Transcription

Video Footage Transcription Example:

W – wizard(evaluator), S – subject

1	*[01:20:39]*	*S:*	*system start, new user*
2	*[01:20:41]*	*W:*	*((recognising the speech))*
			(checking if the word was recognised correctly)
			Returning text 'new user detected. system start'
3	*[01:20:52]*	*S:*	*seaside (in normal speed)*
4	*[01:20:53]*	*W:*	*((recognising the speech))*
			Returning text 'seaside'
5	*[01:20:55]*	*S:*	*thunder*
6	*[01:20:56]*	*W:*	*honda*
7	*[01:21:03]*	*S:*	*system error.*
			Repeat
			Thun- -der
8	*[01:21:05]*	*W:*	*((speech recognition))*
			Returning the text 'thunder'

The transcription symbols used here are commonly used in conversation analysis research. The following symbols are used in video footage in this thesis.

1 The number in the head of line is the line number of transcribed scripts in the example.

[01:20:39] The content in square brackets is the time stamp of the transcribed event.

[* pronoun] The content following the * in square brackets indicates a similar pronoun of word unrecognised.

(()) The description in double brackets indicates evaluator's speech progress activities.

()	The content in brackets indicates the evaluator or subject's transitional behaviours, that may not evolve finally to formal activities.
- -	A double dash indicates subject's stressed pronunciation of the word.
?	A question mark indicates subject's speech is a question.
‘’	The content between single quotation marks is evaluator's text output.

Appendix 5.1

Study 2 – Study Scenarios

You have got a smart coffee table.

This table can recognise your natural speech and help you organise your daily appointments.

By following these speech command formats, you are going to trial this smart table and tell us lastly how would you like this smart table to be designed.

Appendix 5.2

Study 2 – Task List

Tasks
Checking the appointments of a day
Adding some appointments on different days
Deleting some appointments
When necessary, using the cube to launch some applications for assistance, e.g. the central heating control, videos, and a book

Appendix 5.3

Study 2 – Speech Command Formats

Show an appointment: **DATE**
e.g. today, tomorrow, the day after tomorrow, yesterday, the day before yesterday, July 20th 2010, next Monday,
Add an appointment: **DATE + ADD APPOINTMENT + APPOINTMENT DETAILS + CONFIRM**
e.g. tomorrow, add appointment, go to swim tonight, confirm/ abort July 10th 2010, add appointment, go to play tennis, confirm/abort
Delete an appointment: **DATE + DELETE APPOINTMENT + NO. + CONFIRM**
e.g. the day after tomorrow, delete appointment, second appointment, confirm/abort

Operations	**Command Formats**	**Examples**
Show appointments	DATE	Today / Next Friday / 20th July 2010
Add appointments	DATE + 'ADD APPOINTMENT' + APPOINTMENT CONTENTS + 'CONFIRM/ABORT'	Tomorrow, ADD APPOINTMENT, go swimming, CONFIRM/ABORT

Delete appointments	DATE + 'DELETE APPOINTMENT' + NUMBER + 'CONFIRM/ABORT'	Yesterday, DELETE APPOINTMENT, 2nd appointment, CONFIRM/ABORT

Appendix 5.4

Study 3 – Scenarios and Tasks

Study scenario

You have got a new smart coffee table. This table is able to recognise your natural-language speech. By using this table, you can organise your daily appointments and watch movies in ways of human-human conversation.

Tasks

Introduce yourself, and set up conversation with the table

Check if you have any appointments scheduled today

Add a new appointment you are going to do today

Add some other appointments on other days

Addressing the advices given by system's proactive prompts

Previewing some videos by using the cube and speech respectively

Checking if the videos in the system are updated

Appendix 5.5

Study 3 – Schemas for evaluator's system operation

User Inputs	Wizard's Responses	Notes
[drawing mode] Hand movements	Recognising the drawing if cannot, dialogue says: unrecognised gestures, please try again	
[drawing mode] speeches	Drawing out the spoken symbol If not, dialogue says: speech not recognised, say it again	
[video mode] hand gestures	Returning corresponding operations If cannot recognise, dialogue says: can I help you?	
[video mode] speeches	Returning corresponding operations If cannot recognise, dialogue says: pardon?	
[emotion mode] making emotions	Returning corresponding emotion icons	
[emotion mode] speeches	If relevant to emotions, returning emotion icons If cannot, dialogue says: I am confused	
[search mode] gestures	Unrecognised Dialogue says: what can I help you?	
[search mode] speeches	If recognisable, searching the term If cannot, the dialogue say: pardon?	
Other Occasions		

The weather, your car, pets, hobbies	The dialogue says: system found something you may interested in, fancy a look?
Work meetings, plans	The dialogue says: do you want to make a reminder of this?
…	

Appendix 6.1

Study 4 – Microsoft BING API Code Example

```
//apiID:37B381D76D0B572C3C5A222A860BBB1C0DE03AC2
                string requeststring =
                  "http://api.bing.net/xml.aspx?AppId=37B381D76D0B572C3C5A222A860BBB1C0DE03AC2&Version=2.2&Query="
                    +querycontent
                    +"&Sources=image&Image.Count=15";

XmlNamespaceManager namespacemgr = new XmlNamespaceManager(xmldoc.NameTable);
                namespacemgr.AddNamespace("mms", "http://schemas.microsoft.com/LiveSearch/2008/04/XML/multimedia");
                XmlNodeList webresults = xmldoc.SelectNodes("//mms:ImageResult", namespacemgr);
                System.DateTime currenttime = new System.DateTime();
                currenttime = System.DateTime.Now;
                xmldoc.Save("d:\\andol\\PIC" + currenttime.ToString("mmHHddMMyy") + ".xml");
        private void webquery( string querytype, string querycontent)
        {
            if (querytype == "WEB") {
                //apiID:37B381D76D0B572C3C5A222A860BBB1C0DE03AC2
                string requeststring =
                  "http://api.bing.net/xml.aspx?AppId=37B381D76D0B572C3C5A222A860BBB1C0DE03AC2&Version=2.2&Query="
                    +querycontent
                    +"&Sources=web&Web.Count=15";
                HttpWebRequest webrequest = (HttpWebRequest)HttpWebRequest.Create(requeststring);
                HttpWebResponse webresponse = (HttpWebResponse)webrequest.GetResponse();
                XmlDocument xmldoc = new XmlDocument();
                xmldoc.Load(webresponse.GetResponseStream());

                XmlNamespaceManager namespacemgr = new XmlNamespaceManager(xmldoc.NameTable);
                namespacemgr.AddNamespace("web", "http://schemas.microsoft.com/LiveSearch/2008/04/XML/web");
                XmlNodeList webresults = xmldoc.SelectNodes("//web:WebResult", namespacemgr);
                System.DateTime currenttime = new System.DateTime();
                currenttime = System.DateTime.Now;
                xmldoc.Save("d:\\andol\\" + currenttime.ToString("mmHHddMMyy") +".xml");

                foreach (XmlNode item in webresults) {
                    Label thelabel = new Label();
                    thelabel.Content = item.SelectSingleNode("./web:Title", namespacemgr).InnerText
                        + " >>"
                        +  item.SelectSingleNode("./web:Description", namespacemgr).InnerText;
                    thelabel.Width = 320;
                    wrapr.Children.Add(thelabel);

                    Label theclientlable = new Label();
                    theclientlable.Content = thelabel.Content;
                    theclientlable.Width = client1window.client1wrapr.Width;
                    theclientlable.FontSize = 14;
                    SolidColorBrush thebrush = new SolidColorBrush();
```

```
            thebrush.Color = Color.FromRgb(255, 255, 255);
            theclientlable.Foreground = thebrush;
            client1window.client1wrapr.Children.Add(theclientlable);
        }

    }
    else if(querytype == "PIC") {
        //apiID:37B381D76D0B572C3C5A222A860BBB1C0DE03AC2
        string requeststring =
        "http://api.bing.net/xml.aspx?AppId=37B381D76D0B572C3C5A222A860BBB1C0DE03AC2&Version=2.2&Query="
            +querycontent
            +"&Sources=image&Image.Count=15";
        HttpWebRequest webrequest = (HttpWebRequest)HttpWebRequest.Create(requeststring);
        HttpWebResponse webresponse = (HttpWebResponse)webrequest.GetResponse();
        XmlDocument xmldoc = new XmlDocument();
        xmldoc.Load(webresponse.GetResponseStream());

        XmlNamespaceManager namespacemgr = new XmlNamespaceManager(xmldoc.NameTable);
        namespacemgr.AddNamespace("mms", "http://schemas.microsoft.com/LiveSearch/2008/04/XML/multimedia");
        XmlNodeList webresults = xmldoc.SelectNodes("//mms:ImageResult", namespacemgr);
        System.DateTime currenttime = new System.DateTime();
        currenttime = System.DateTime.Now;
        xmldoc.Save("d:\\andol\\PIC" + currenttime.ToString("mmHHddMMyy") + ".xml");
        foreach (XmlNode item in webresults)
        {
            BitmapImage bitmap = new BitmapImage();
            bitmap.BeginInit();
            bitmap.UriSource = new Uri(item.SelectSingleNode("./mms:MediaUrl", namespacemgr).InnerText);
            System.Windows.Controls.Image theimage = new Image();
            theimage.Margin.Right.Equals(15.00);
            //theimage.Source = bitmap;
            theimage.Width = theimage.Height = 70;
            wrapr.Children.Add(theimage);
            bitmap.EndInit();

            BitmapImage clientbitmap = new BitmapImage();
            clientbitmap.BeginInit();
            clientbitmap.UriSource = new Uri(item.SelectSingleNode("./mms:MediaUrl", namespacemgr).InnerText);
            System.Windows.Controls.Image clienttheimage = new Image();
            clienttheimage.Margin.Right.Equals(15.00);
            clienttheimage.Source = clientbitmap;
            clienttheimage.Width = clienttheimage.Height = 120;
            client1window.client1wrapr.Children.Add(clienttheimage);
            clientbitmap.EndInit();
        }
    }
}
```

Appendix 6.2

Study 4 – Study Scenarios and Tasks

Study Scenario

Easter is coming. You are about to schedule some events for the coming holiday. The first thing is that you are going to organise a party, which you will invite some friends to. Also, you need to make a reminder of shopping food and drink for the party, and you need to remind yourself to pick a friend Alex from university at 5-30 pm to your party after shopping. After the party, you want to plan a charity hike for West Yorkshire Cancer Research with your friend Tom. To make the hike routine, you need to communicate with Tom. He is home now whom you can use the system to search and share pictures and videos with. Tom can express his opinions towards routines through the emotion recognition system, and he can also comment routines sketching on the table remotely.

Study Tasks

- sending invitation messages to your friends
- making a reminder of shopping
- making a reminder of picking up Alex at 5-30 pm
- seeing if Tom is online for hike routine discussions
- searching information related to pictures or web pages

- sharing the information you found with Tom
- telling your thoughts about system functionalities

Appendix 6.3

Study 4 – Study Scenarios

Study scenario

You are going to organise a party, and you have a smart coffee to assist your job. The table understands natural speeches and body gestures, besides it may also make some smart suggestions. To plan the party you have something to schedule using the smart table. You need to buy some food and drink thus making a shopping list, to pick up one of your friend Tom, to check the domestic multimedia system working fine, and to search some pictures for party settings… now let's start.

Tasks

Asking the smart system for suggestions what to shop for the party (asking the coffee table questions, the system will present some feedbacks through dialogues or pictures.)

Making a reminder for shopping (saying the event 'shopping list', and you can check the reminder through picture viewer.)

Making a reminder to pick up your friend Tom from university at 5:30pm (a more complicated reminder to make, saying the event, and you can draw some personal symbols to remind your.)

Checking all reminders you have made using the picture viewer

Checking the multimedia system whether it works fine and its contents are updated (testing all functionalities – playing, pause, stop, volume up/down, fast forward/backward, and changing video channels)

Sharing your emotions with friends/families (the system can sense your emotion status and generate corresponding icons.)

Trying to do some other tasks in your own ways (the task can be anything)

Making some comments on the system, how you felt and what should be improved

Appendix 6.4

Study 4 – System Operation Schemas

Interaction requests	Operation scripts
Hello, I am …	Hi, …, can I help you
Show me tomorrow's appointments	It is a busy day [calendar] clicking the date of today to show events
Add appointments	Please pick a date
Saturday	[calendar] picking the date, then getting ready to typing events
Go shopping	John will go shopping on Saturday, send an invitation? [calendar] typing the events, then confirming the typing Invitation has been sent
delete appointments	Please pick a date
tomorrow	Please pick an appointment to delete
go to cinema	Confirmed to delete this?
Confirms	Deleting appointments done. Want to see some movie posters you may be interested in?
Yes/no	Yes - [photo] launching the photo browser

	No – Er, how about some funny pictures?
Yes/no	[photo] launched photo browser/
-after photo browsing-	Would you like to watch some videos?
Using gestures to control photo browsing	[photo] Up key/down key to correspond related manipulations [photo] |Up key/down key to correspond subject's manipulation speeches
TV	[TV] clicking relevant videos to correspond related gestures [TV] clicking relevant videos to correspond subject's manipulation speeches: play, pause, stop, forward, backward, next, previous
Dynamic spoken interactions	Depending on situation, providing simple and human-like emotional responses