Design Parameters for Online Psychometrics

Shiny Verghese

A thesis submitted in partial fulfilment of
the requirements of Teesside University
for the degree of Doctor of Philosophy

October 2018

School of Social Sciences, Humanities and Law
Teesside University
Middlesbrough, United Kingdom

DECLARATION

While registered as a candidate for the degree of Doctor of Philosophy the author has not been registered for any other award with any other university or institution. No part of the material in this thesis has been submitted for any degree or other qualification at any other institution by the author or, to the best of her knowledge and belief, by any other person. The thesis describes the author's original work.

ABSTRACT

In contemporary society, where technology is rapidly spreading, the traditional method of (offline) testing through pen(cil) and paper is being converted to online psychometric test administration. There is a plethora of research available on the advantages of online administration of questionnaires. However, much of this work addresses comparisons between online and offline administration, factual questionnaires rather than psychometric questionnaires or online psychometric questionnaires without addressing presentation- or interaction design. In human-computer interaction online psychometric questionnaires are, for example, used to measure customers' perceived quality of (online or offline) services and to measure users' interaction experience with a Web site in terms of flow experience. Compared to the popularity of web-based surveys, there is little research available to aid the design of online psychometric questionnaires and to ensure sound measurement. Because psychometric questionnaires do not measure factual information, it is more likely that the responses given are influenced by external factors, such as the presentation design of the particular questionnaire that is being administered. Research reports that reading speed is affected by font size which (in turn) could apply for online psychometrics in terms of completion time of questionnaires. It is essential to further develop the scientific understanding of how presentation-design factors affect people's responses in psychometric measurement and design guidance. The aims of this research are to develop a technical system to support the required research, gather data for online psychometrics with manipulation of design parameters, provide empirical evidence of the effect of design parameters on online psychometric measurement and finally provide design guidance for online

psychometrics. The results could be extended to various other settings such as educational assessments.

DEDICATION

I dedicate this thesis to my philosopher and guide Paul van Schaik, my loving daughter Riya Grace, my beloved parents Thomas Verghese and Elizabeth Usha Verghese and my dear brother Reuben Verghese.

ACKNOWLEDGEMENTS

The Almighty's mercy and grace has led me to this accomplishment. These past seven years have been a very eventful period in my life. Reaching this point was once a dream and getting this far with my studies was not easy. I have definitely not done it all by myself. First and foremost, I am deeply indebted to my director of studies, Paul van Schaik, for his constant and invaluable guidance and support throughout the whole process. Without your support and guidance Paul, I would have been like the sailor without his mariner's compass. I take this opportunity to express my sincere gratitude to my second supervisor, Wilson Christopher, for his thoughtful support during needy times and useful feedback on this document.

A special thank you to Ewan Ingleby for the couple of times he arranged to have me as a distance presenter through Skype for the Postgraduate Forum and for all the progression meetings arranged through Skype. I am also grateful to the following Teesside university staff: Martin Leyland, Lorraine Smeltzer, Julie Wright and Barbara Cox for their timely assistance to all my requests. Special thank you to Gabor Aranyi, for the guidance with the formatting of my thesis.

I wish to thank Kameleddine Benameur, Iqbal Al Shammari and Mariam Kinawi who were the members of the translation committee from GUST. My gratitude to Abdulmohsen Yousef Jamal, Ayman Bakr, Taher Ali and Mahmoud Elish, faculty members of GUST who kindly allowed me to visit their classes and request student participation for my experiments.

I extend my gratitude to Abdulraheem Taha for his invaluable support in the development of the web-based tool. You have been the compiler and debugger in all my needy times. Mariam Kinawi my best friend, Ahmed Ali Salman my ex-student, and to all my friends, here is my heartfelt gratitude for your words of support and encouragement that will never be forgotten.

I would have never made this dream come true without the support of my parents, and most of all my now thirteen year old girl Riya. Your warm hugs were my source of energy and during all these crucial years, Riya trained herself to become as independent as possible because in her words, “my mama has to complete her PhD”. Your support and sacrifice to save me extra time so I can complete my PhD journey is so much cherished and thanked. Last but not the least, a big thank you to my cat Kittu for being there with me on all the sleepless nights as my companion with purrs of encouragement. All of you made this, the most pleasurable part of my life and for me, this is not the end but just the beginning.

Thank you

ABBREVIATIONS

ANOVA	Analysis of variance
DIS	Disorientation
DPI	Dots per inch
ER	Entity relationship
F	F-test
HCI	Human-computer interaction
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
M	Mean
mm	Millimetre
MVC	Model-View-Controller
n	Number of cases
N	Total number of cases
NASA-TLX	NASA-Task Load Index
OnPQDT	Online psychometric questionnaire design tool
p	Significant probability
PE	Perceived enjoyment
PEU	Perceived ease of use
PPI	Pixels per inch
PU	Perceived usefulness
PSSUQ	Post system study usability questionnaire
r	Pearson’s product moment correlation coefficient
RDBMS	Relational Database Management System
SD	Standard deviation

SPSS	SPSS Statistical Package for the Social Sciences
SQL	Structured Query Language
SUS	System usability scale
t	t-test
TAM	Technology acceptance model
UX	User-experience
WPRE	Web-based psychometric research environment

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Contents

1.1 Overview

The primary aim of this chapter is to present the background and rationale to the research undertaken and presented in this thesis. Given the proliferation of online psychometric questionnaires on the Internet and other platforms, the design of online psychometrics becomes increasingly important to ensure good measurement properties. The chapter starts with a brief discussion of psychometrics and human-computer interaction. Next, the design of questionnaires specifically in online psychometrics is introduced and a need for research is identified. Three broad research questions are formulated, the research approach is presented and the aims of the current research are designated. The chapter closes with an outline of the structure of this thesis.

1.2 Psychometrics and its general mainstream application

Psychometrics is a well-developed field. The goal of psychometrics is to establish the quality of psychological measures (J. C. Nunnally, 1978). Psychometrics is a branch of psychology that focuses on the operationalization of variables for the purposes of measurement (Vogt, 2005). It is defined as the science of measurement. Traditionally, there have been two main types of psychometric instruments: ability tests and personality tests. The psychometric test is an assessment tool that consists of any standard procedure for measuring sensitivity, memory, intelligence, aptitude or personality (Colman, 2009) Early applications of psychometrics were in educational assessment, and clinical or occupational testing. For example, psychometric tests were and are still used to select appropriate candidates in an organisation.

There is also a growing market for personality questionnaires, for example questionnaires that show evidence of extrovert nature in candidates. No matter where psychometric tests are used within recruitment, team development, personality assessment, they can add substantial and valid information to decision-making processes. However, it is important that detailed psychometric work is conducted to validate scales in any discipline before mainstream use (Copping, Campbell, Muncer, & Richardson, 2017, p. 2). Currently, a significant impact on the application of psychometrics in both educational and occupational fields are due to the great advances made in statistical modelling (Rust & Golombok, 1989). The lead is coming from psychologists analysing large-scale survey data including questionnaires in different domains such as epidemiology and evolutionary psychology among many others. One such example can be found in the research reported by Copping, Campbell, and Muncer (2014) conducted among a large British sample (n = 809). The researchers raised concerns regarding the use of psychometric indicators of lifestyle and personality as proxies for life history strategy when they have not been validated. This is was because of the detailed psychometric work they undertook to assess the factor structure and validity of the High K Strategy Scale (HKSS: Giosan, 2006).

1.3 Psychometrics in human-computer interaction

Human-computer interaction (HCI) developed as an interdisciplinary field and a community in the early 1980s.

A distinction needs to be made between human-computer interaction as an activity (people's interaction with interactive computer systems) and the field of human-computer interaction, abbreviated as HCI ("a discipline concerned with the design, evaluation and implementation of interactive computer systems for human use and with the study surrounding them" (Hewett et al., 1992, p. 5). The theoretical roots of HCI encompass a number of other disciplines outside of computer sciences, including psychology, computing, ergonomics and social sciences. Some of the classic HCI publications of the 1980s such as the psychology of human-computer interaction (Card, Moran, & Newell, 1983), Norman's analysis of human error (Norman, 1983), Carroll's 'training wheels' approach to interface design (Carroll & Carrithers, 1984), and Shneiderman's work on direct manipulation (Shneiderman, 1983) are still relevant today. As the Internet and the web gained wide acceptance, a number of research fields under the umbrella of HCI increased. One such example is the research by McDonald and Stevenson (1996) that reported disorientation in hypertext and the effect of text structures on navigation performance. The implications of this research extended to computer-assisted learning systems. According to Ben Shneiderman: "The old computing is about what computers can do, the new computing is about what people can do" (Shneiderman, 2002, p. 1). HCI seeks to develop theoretical knowledge regarding the design and use of interactive computer systems, and offer practical guidance to practitioners in interaction design, usability and UX. One such example is the design and use of questionnaires via the Internet (online psychometrics). Research in this area provides knowledge about the effect of design parameters (e.g. response target size) on outcomes (e.g. completion time) and guidelines for designing online psychometrics.

There is a major difference between the application of psychometrics in personality testing and usability evaluation. In the former, relatively stable traits such as Extraversion, Agreeableness, Dependability, Stability and Culture are measured (Goldberg, John, Kaiser, Lanning, & Peabody, 1990). However, participants in a usability study give their responses in the context of using an interactive system. ISO-9241-11 defines usability in terms of effectiveness, efficiency and satisfaction in a particular context of use (Bevan, Carter, & Harker, 2015). This may, in itself, influence participants' ratings. They may feel their role is to be more critical of the user interface than they might otherwise be.

In the related technology-acceptance literature, usability is described by the concepts of perceived usefulness and perceived ease of use. Perceived usefulness is defined as “the degree to which a person believes that using a particular system would enhance his or her job performance” (Davis, 1989, p. 320). To measure both perceived ease of use and perceived usefulness, validated scales have been developed in the technology acceptance literature (Venkatesh et al, 2003). According to Norman and Nielsen (2016), "user experience" (UX) encompasses all aspects of the end-user's interaction with the company, its services, and its products. Usability can be considered as one of these aspects. This is also reflected in professional practice: in 2012, the Usability Professionals Usability Association (UPA) changed its name to User Experience Professionals Association (UXPA).

An example of the application of usability testing is the use of think-aloud method and psychometrics to explore users’ experience with a news web-site (Aranyi, Schaik, & Barker, 2012). Thinking aloud may be the single most valuable usability engineering method (Nielsen, 1993 as cited in Holzinger, 2005). Research by McDonald, Edwards, and Zhao, (2012) identified the gap between theory and practice of think-aloud methods usage through an exploratory study. The implications of this research provided insight into current think-aloud practice in usability studies and the underlying reasons determining the approaches adopted. The think-aloud technique is useful in formative evaluation. The aim is to test designs and use the test results to guide the further design and development of interactive systems. This important formative testing work complements summative evaluation. The aim is to test a fully functional interactive system at the end of the development phase, for example in terms of task performance. Psychometric measurement then takes place after test users have completed their tasks to complement the task performance results. For example, a high degree of psychometrically measured disorientation may explain why test users take a long time to complete their tasks.

HCI researchers have applied psychometrics, to measure the quality of human-computer interaction. For example, usability researchers have used psychometrics to develop and evaluate questionnaires to assess usability (Sweeney & Dillon, 1987). Psychometrics helps measure usability constructs with reliability and validity (Lewis, 2002). Van Schaik and Ling (2005) tested five psychometric scales for online measurement of the quality of human-computer interaction in websites and also established the psychometric properties of the scales.

Another important consideration of psychometric instruments in HCI is their sensitivity to experimental manipulations (Lewis, 2002), for example design parameters of web pages (van Schaik & Ling, 2001a; 2001b, 2003a, 2003b). Therefore, the application of psychometrics within human-computer interaction within computer science. In this research, psychometric questionnaires that measure (1) usability such as PSSUQ (see section 3.5), SUS (see section 3.6) and (2) the quality of human-computer interaction (see section 3.4) such as Disorientation, Perceived ease of use, Perceived usefulness are used.

1.4 Design parameters in online psychometrics

With the application of psychometrics in developing and evaluating standardised usability and user experience questionnaires (Lewis, 2018), the design of online psychometrics becomes increasingly important to ensure good measurement properties. In contrast to research on survey design guidelines (e.g., Andrews, Nonnecke, & Preece, 2003; Couper, 2008; Larossi, 2006; Toepoel, 2017), few studies exist on the design of online psychometrics (van Schaik & Ling, 2003, 2005a, 2005b, 2007; van Schaik et al, 2015). Because psychometric questionnaires do not measure factual information, it is more likely that the responses given are influenced by external factors, such as the presentation design of the particular questionnaire that is being administered. An example of objective (factual) information is the demographics when answered truthfully are not likely to be affected by an external factor such as a design parameter for example font size because there is only one obvious correct answer.

However, a response to question for subjective (nonfactual) information such as a user's perception of their interaction with a system does not have a correct answer and is therefore subject to external influences (such as presentation design); therefore, variability in answers as a function of design parameters is likely. For example, research has shown a substantial impact of presentation format on response quality (van Schaik & Ling, 2003; van Schaik & Ling, 2007; van Schaik et al, 2015)

1.5 Research questions

The use of computers has rapidly changed society. As a result, human-computer interaction has increased in many aspects of our daily lives including assessment, information search and diagnostics. Various research studies demonstrate how usability science, along with other research in HCI, can benefit from the application of psychometrics in different situations (e.g., Lewis, 2015; Tuch, Schaik, & Hornbæk, 2016)

Because psychometrics models human psychological characteristics, it is important for instruments administered online to be sound and standardised in terms of measurement. Within current work in online psychometrics, little research exists that addresses psychometric measurement in human-computer interaction through web pages or mobile applications. Similarly, there is lack of research systematically investigating the psychometric quality of a range of major measures of interaction as a function of design parameters when presented online. This systematic investigation is important because the effects of design parameters have to be identified based on which design guidelines could then be developed.

Moreover, investigations of empirical research show a significant impact of presentation format on response quality (van Schaik & Ling, 2003; van Schaik & Ling, 2007). However, large-scale empirical research in terms of wide-ranging large number of participants, extensive collection of psychometric questionnaires and a comprehensive set of design parameters, is required to develop a complete understanding of online psychometrics.

Accordingly, the research presented in this thesis systematically studies design parameters of online psychometrics through a combination of innovation and experimental empirical research to make an original contribution. Specifically, the research presented in this thesis develops (a) an instrumentation for the development and deployment of online psychometrics that allows the systematic manipulation and testing of design parameters, and further develops (b) the scientific understanding of how presentation-design factors affect people's responses in psychometric measurement and (c) design guidance. Accordingly, the research presented in this thesis addresses the following research questions.

1. What technical system is required to support research on design parameters in online psychometrics?
2. What are the effects of design parameters in online psychometric measurement?
3. How can the knowledge acquired by answering Research Questions 1 and 2 be applied to guide system design?

Research Question 1 can be addressed by examining current existing software tools available for questionnaire design and administration and the suitability of these tools for research in online psychometrics. Once the specifications have been identified and mapped against the requirements for research in online psychometrics, either such an existing tool can be adopted or if there exists a need, a technical system may have to be developed to support online psychometric research. Research Question 2 can be addressed by formulating and creating human-computer interaction experiments using the technical system, as a basis for further developing the knowledge about the effects of design parameters on people's responses to online psychometric questionnaires. Furthermore, the results from the experiments can be analysed to test the effects of design parameters on people's responses. The psychometric properties in particular, reliability, validity and factor structure of the online questionnaires can also be assessed in these experiments. Research Question 3 relates to the application of the results of the research to derive design guidance. A research approach related to the three research questions developed for this project is presented in Figure 1.1.

Multifactorial research designs – combinations of design parameters of line psychometrics manipulated:

1 font size;
2 text/background colour/polarity;
3 response target size;
4 type of response format.

↓

Online psychometric experiments run

↓ ↓

Online psychometric data | Interaction outcome data

↓ ↓

Psychometric analysis by design parameter combination: | Multifactorial analysis of variance

1 factor structure;

2 reliability;

3 validity.

Figure 1.1. Research approach for empirical studies.

1.6 Contribution to knowledge

In summary, the expected impact of the research project presented in this thesis and the contribution to knowledge is threefold. First, the project contributes a novel tool for researchers interested in online psychometrics research, specifically design parameters of online psychometrics. Second, by establishing and quantifying the effects of design parameters on interaction outcomes in online psychometrics a comprehensive understanding of psychometric measurement in human-computer interaction in terms of completion time, perceived enjoyment and workload is formulated. Finally, design guidance is derived to aid the design of online psychometric questionnaires.

1.7 Structure of this thesis

The three research questions presented in the previous section are addressed in six chapters in this thesis. Chapter 2 outlines the concept of psychometrics and human-computer interaction. A literature review of psychometrics in human-computer interaction research is presented. Chapter 3 identifies a set of design parameters in online psychometrics, followed by the development of the hypotheses for the chosen design parameters that are manipulated in the research presented in this thesis. Research Question 1 is addressed in Chapter 4 in terms of the development of an online tool for web-based psychometric research environment. Chapters 5, 6 and 7 address Research Question 2.

Three separate studies: Study 1, Study 2-A and Study 2-B with the data collected from Kuwait are presented. The implications of Kuwait at the location where the experimental studies were conducted are as follows.

Although there is not much difference in the procedure of conducting the experimental research in Kuwait compared to other regions of the world, clearly, there is lack of published research evidence regarding the existence and validity of usability questionnaires in Arabic (AlGhannam, Albustan, Al-Hassan, & Albustan, 2017). Therefore, it is important to establish the psychometric properties of translated or new questionnaires in Arabic. Consequently, this research not only addresses the effect of design parameters on human-computer interaction outcomes, but also the translation and validation of selected psychometric instruments.

Study 1 is presented in Chapter 5, and manipulates and tests three design parameters (font size, text-background colour and response target size) on mobile devices. Study 2-A presented in Chapter 6 and Study 2-B presented in Chapter 7, manipulate and test three design parameters (font size, text-background polarity and response format) on mobile devices (Study 2-A) and desktop computers (Study 2-B). Research Question 3 is addressed in Chapter 8 where design guidance is derived based on the results from Chapters 5, 6 and 7. Finally, a discussion of the studies, their limitations and suggestions for future work are presented in Chapter 8.

Chapter 2

Psychometrics

Contents

2.1 Overview

The goal of this chapter is to provide (1) a background and (2) a rationale for the research undertaken and presented in this thesis. The chapter starts with a brief introduction to psychometrics, its uses and a closer look at its application in human-computer interaction. The structure of psychometric tests and the content of psychometric items are discussed. Next, the method of online psychometric measurement is reviewed and research that compares online psychometrics with psychometrics using a pen(cil) and paper are summarised. Finally, design parameters for online psychometric measurement are discussed with a description of design options and results of existing research.

2.2 Introduction

Modern psychometrics dates back to Sir Francis Galton (1822-1911), who demonstrated that objective tests on human participants could provide meaningful measures. Psychometrics is the study of measuring complex psychological concepts, or constructs, such as a person's motivation, anger, personality, intelligence, attachment, fear (Nunnally, 1978) or as described by van Schaik et al. (2015, p. 52) "to measure people's abilities, attitudes or perceptions in various domains, including human-computer interaction (e.g. disorientation experienced by the users of a website, Ahuja and Webster, 2001)." Psychometrics can be understood as a discipline that models human psychological characteristics mathematically. In almost every aspect of our daily life (e.g., education or work), we are continuously assessed in different forms such as interviews, examinations, practical, multiple-choice tests.

Despite the wide variety of assessments, what remains constant is the aim of good-quality measurement. An aim of psychometrics is to maximise the quality of assessment. Irrespective of where psychometrics is used, it can provide substantial valid information to inform decision-making processes.

The use of computers in our daily lives has rapidly changed the way present-day societies exist and function. As a result, human-computer interaction has increased in many aspects of our daily lives, including assessment, information search and diagnostics. Various research studies demonstrate how usability science, along with other human-computer interaction research, can benefit from the application of psychometrics in different situations. The most common application for psychometrics and human-computer interaction is questionnaire administration and data collection. Within the current work in online psychometrics, specifically, there is lack of research investigating the psychometric quality of a range of significant measures of human-computer interaction and taking into account the way psychometric items are presented online.

An important rationale behind research on design parameters in online psychometrics is the need to investigate and report a comprehensive set of design parameters so that a framework for design guidelines can be established. With these guidelines a high-quality user experience in online psychometrics can then presumably be promoted.

This research aims at providing comprehensive knowledge on the effect of presentation formats of online questionnaires on participants' responses.

In this chapter, the basics of psychometrics are reviewed, followed by a discussion of the transition of psychometric questionnaires using pen(cil) and paper to an online questionnaire mode. Finally, a literature review of existing research on online psychometrics in the field of human-computer interaction is presented.

2.3 Psychometrics in Human-Computer interaction

Human-computer interaction (HCI) is the study of interaction between people (users) and computers. Since the use of psychometrics from the early 1900s, it was during the late 1970s, that psychometric questionnaires started to be used to measure the quality and process of human-computer interaction (Kirakowski, Claridge & Whitehand, 1998). In this chapter the emphasis is on the task of providing reliable, valid and useful scales for the applied discipline of HCI. There is a vast number of scales that measure different HCI aspects such as usability, satisfaction and experience. It is important that the constructs these scales measure are clearly defined. HCI researchers have therefore applied psychometrics, the science of measurement, to measure usability satisfaction and experience. Usability questionnaires such as SUMI (Software Usability Measurement Inventory) by Kirakowski and Corbett (1993), QUIS (Questionnaire for User Interaction Satisfaction) by Chin, Diehl and Norman (1988) and MPUQ (Mobile Phone Usability Questionnaire) by Ryu and Smith-Jackson (2006) were developed by following psychometric approaches. Sauro and Lewis (2009) employed factor analysis, a statistical method widely used in psychometrics, to identify the fundamental factors or aspects of usability for the System Usability Scale (SUS).

In another work, Lewis evaluated the psychometric properties of four existing IBM questionnaires that were developed for measuring user satisfaction with computer system usability (Lewis, 1995). Further, psychometric properties of PSSUQ questionnaire using data from five years were also established by Lewis (2002). Research by Ahuja and Webster (2001) led to the identification of two new scales to measure disorientation and perceived ease of use to explain experiences with Web-based systems. In another study by Davis and Wiedenbeck (2001), a new scale was developed to measure flow, a psychological state of a person to feel cognitively efficient or motivated and happy (Moneta & Csikszentmihalyi, 1996) while using websites. In establishing the quality of users' interactions with web sites, the research by van Schaik and Ling (2003), was the first to investigate the psychometric properties of three existing quality of interaction scales: disorientation, ease of use and flow. In addition, the researchers also examined the influence of response format on the quality of interaction with web pages. Van Schaik and Ling (2005) further tested five psychometric scales for online measurement of the quality of human-computer interaction in websites. Van Schaik and Ling (2007) also studied the effect of design parameters of rating scales (Likert scale using radio buttons and visual analogue scale) on four questionnaires Disorientation, Perceived ease of use, Perceived usefulness and Flow, tested for online measurement of the quality of human-computer interaction. Overall, the study concluded that the instruments demonstrated good psychometric properties for both response formats for measuring the perceived quality of interaction with web sites. In addition, van Schaik et al. (2015) conducted a study that was directed at a different important consideration in the design of online psychometrics: questionnaire layout.

This study established that psychometric questionnaires administered online need to be administered using single-item layout to reduce completion time and facilitating attention in questionnaire completion. The results of these research studies (van Schaik & Ling, 2003, 2005, 2007; van Schaik et al., 2015) form the starting point for the current study: establishing design parameters of online psychometrics for sound measurement of quality human-computer interaction. In online psychometrics, it is important that the questionnaires meet the requirements on the following characteristics (Lewis, 2002): (1) factor structure, (2) reliability (consistency of measurement) and (3) validity (measurement of intended attribute)..

2.3.1 Factor Structure

Factor analysis is a statistical procedure that is used to estimate factors or reduce a large number of variables to fewer ones (Pedhazur & Pedhazur-Schmelkin, 1991). It examines the correlations or covariance among items to discover groups of related items. In psychometrics, factor analysis is often used to identify the underlying constructs that might exist in the data. These constructs are called as factors. Two types of factor analysis are distinguished: exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis (EFA) is used when the researcher is ambiguous about the theoretical conceptualisation of the construct. This method helps to explore the underlying factors of the construct. Thus, this method provides an opportunity to improve the theory at an early stage of scale development. Confirmatory factor analysis (CFA), on the other hand, is used when the researcher has a more specific theory about the conceptualisation of the construct of interest; CFA is usually conducted on scales that have first been developed and analysed with EFA. The researcher will have a clear idea on the number of factors that is expected to emerge. Based on this theory, the researcher builds a model and gathers data to examine whether the data fit the hypothesised model. Within exploratory analysis, there are a number of ways to extract factors. One of the common methods is the principal component analysis (PCA). This method is widely used for determining a first set of loadings. All the variance in each variable is analysed in PCA. Strictly speaking, this method is not factor analysis, because it is based on a different measurement model. However, the pattern of PCA results is often the same as that of other extraction techniques. The most common extraction method in factor analysis is the principal axis factoring (PAF). PAF is a measurement model of the latent variable. Only the shared variance is analysed in the PAF.

Factor rotation is an important consideration in factor analysis. By maximizing high item loadings and minimising low item loadings, rotation helps to produce a more interpretable factor analysis solution. There are several rotation techniques. The most commonly used rotations are varimax (orthogonal, producing uncorrelated factors) and direct oblimin (oblique, producing correlated factors). Rotated factors are typically used to assess the underlying structure of the questionnaire items in terms of factors.

2.3.2 Reliability

In a general sense, “Reliability refers to the degree to which the test scores are free from errors of measurement” (American Psychological Association, 1985, p. 19).

Two main types of reliability are distinguished: internal-consistency reliability and test-retest reliability. Internal-consistency reliability of psychometric instruments is the degree to which the items that make up a factor are related. A questionnaire's reliability is a quantitative assessment of its consistency. This is assessed by employing Cronbach's coefficient alpha (Cortina, 1993; Nunnally, 1978). Cronbach's alpha α ranges from zero (no reliability) to one (perfect reliability). The test-retest reliability is the ability of a measure to produce consistent results when the same entities are tested at two different points in time (Field, 2013).

2.3.3 Validity

The validity of the psychometric instruments is described as the extent to which an instrument evaluates what it intends to measure. Validity of different types exist and the type of validity maps to the purpose of the scale. Criterion-related validity measures the relationship between the measure of interest and a different concurrent or predictive measure. Discriminant validity determines the level of differentiation between measures of distinct constructs. In both types of validity, the Pearson correlation coefficient r is used for assessment. Correlations in the range of 0.30 – 0.40 are deemed sufficient to demonstrate the validity of psychometric instruments (Nunnally & Bernstein, 1994).

2.4 Structure of psychometric tests

As discussed previously, a psychometric instrument is a tool for measuring human psychological characteristics mathematically. The structure of psychometric tests is composed of items, subscales and scales.

2.4.1 Items

An item represents the fundamental unit of measurement in a psychometric test. Items are structured into attribute description and a response part. The attribute description is the item stem that is composed of full sentences, phrases or single words. The response part is the measure that describes the degree of the attribute description and varies according to different response formats of scale points (choices) or anchored phrases (yes-no, agree-disagree). A detailed explanation of the different response formats is presented in Section 2.5. For every item, a set of properties (item parameters) is estimated. When an item level analysis is performed, feasibility and difficulty of each item are determined. Item difficulty is a measure of, for example, the ability of the people who responded correctly to an item. As discussed in Streiner and Norman (2008) item difficulty is determined by an inspection of the mean and endorsement frequency for each item. When items measure what they are intended to measure, the item-total correlation value is between 0.2 and 0.7 (Streiner & Norman, 2008). Item redundancy is observed when items are similar. Item redundancy is noted when correlations exceed 0.7. Similarly, when correlation falls below 0.2 the item is observed to measure an entirely different construct. A classic example of a psychometric development for HCI can be seen in Davis (1989). Various researchers adopt factor analysis for item selection in a multidimensional construct while other researchers use the non-factor analytic internal-consistency method such as reliability analysis for item selection when developing a scale. Internal-consistency- or reliability analysis is estimated by an index such as coefficient alpha (Cronbach, 1951).

However, the latter method on its own without factor analysis is flawed because there are no universally agreed standards regarding what level of reliability is considered acceptable. Although a minimum of 0.80 and 0.90 is recommended in the research by Nunnally (1978), some contemporary researchers characterise reliabilities in the 0.60s and 0.70s as good or adequate (Clark & Watson, 1995). Internal consistency is a necessary but not sufficient condition for determining whether the scale items assess a single underlying factor or construct (Briggs & Cheek, 1986; Cortina, 1993; Green, Lissitz, & Mulaik, 1977). With only reliability analysis and without factor analysis, the number of constructs and the factor structure underlying the data remains unknown. A scale is composed of several individual items and the quality of the scale depends on the quality of its items (Rust & Golombok, 1989).

2.4.2 Scales

As mentioned earlier, scales are composed of items. A summated item score can be used to measure each construct. Since no true score for a test exists, it is important that instruments constructed from multi-item scales have high reliability and validity. One of the main types of reliability is internal consistency. It can be defined as the degree to which the items that make up a factor are related and is usually assessed by employing Cronbach’s coefficient alpha. If alpha is sufficiently high (> 0.70), then the items are often added up or averaged to produce a scale, thereby reducing the larger set of item scores to a single scale value for the underlying construct.

Careful selection of the initial scale items helps to assure the scales will possess content validity, defined as "the degree to which the score or scale being used represents the concept about which generalizations are to be made" (Bohrnstedt, 1970, p. 91, as cited in Davis 1989). Apart from content validity, scales must also be tested for convergent and discriminant validity. This can be done using multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959, as cited in Davis 1989), factor analysis and other methods. To demonstrate convergent validity, items that measure the same trait should correlate highly with one another (Campbell & Fiske, 1959, as cited in Davis 1989). The test for discriminant validity is that an item should correlate more highly with other items intended to measure the same trait than with different items used to measure a different trait (Campbell & Fiske, 1959, as cited in Davis 1989). Factor-analytic methods can also be used to study unidimensionality and discriminant validity of scales. A good example once again for scale construction and validation of a scale for HCI is provided by (Davis 1989). In the same study, new scales for two specific variables, perceived usefulness and perceived ease of use, were developed and validated. These have been hypothesised and demonstrated to be fundamental determinants of users' acceptance in the fields of information systems and HCI (e.g. Davis, 1993).

2.4.3 Subscales

Scales in psychometrics can sometimes be divided into subscales. In other words, subscales are hypothesised to be manifestations or dimensions of a more general construct (Clark & Watson, 1995).

It is normally important to determine for an individual scale whether it is unidimensional (a single scale score is then calculated) or a multidimensional (a subscale score is then calculated for each subscale) is required. Factor analysis can be used to explore the subscale structure. Such subscale structures can be particularly informative for further diagnosis. Factor analysis of all the test items is useful to decide whether separate subscales are required. Subscales are recommended to adequately assess each major scale construct, which otherwise may result in an incorrect dimensionality of measurement. Subscales must exhibit content, discriminant and conceptual validity. A good example is the flow scale devised by Davis and Wiedenbeck (2001) with the two dimensions of involvement and control within the construct of flow experience. Another example is the PSSUQ questionnaire is also composed of three subscales System Usefulness (SYSUSE), Information Quality (INFOQUAL) and Interface-Quality (INTERQUAL); the studies by Lewis (1995, 2002) have confirmed good psychometric properties for the overall scale and its subscales.

2.5 Content of psychometric items

As discussed in the structure of psychometric tests, a psychometric item represents the fundamental unit of measurement. It is composed of an attribute describing the item stem and a response part. The response part is a measure that describes the degree of the attribute description.

2.5.1 Stem

When developing a scale, it is important to write items in simple, straightforward and appropriate language.

It is worth consulting the available literature on item writing (e.g. Angleitner & Wiggins, 1985; Comrey 1988; Kline 1986; as cited in Clark & Watson, 1995). Lindenberger and Nesselroade (1999), presented a comprehensive framework for item selection and inclusion in a scale. In their paper, the authors detail how item representativeness competes against internal consistency and further state that "Selecting variables in psychological research has been a long-standing concern, even though the volume of attention has been relatively low" (p. 193). Relevant content is of high significance when writing items. Each item's content must reflect the intended psychological variable. Many critical psychological constructs are broad in scope, having several facets or modes of manifestation. For example, for flow experience, Davis and Wiedenbeck (2001) conceptualised the items regarding involvement and control. Thus, the items reflected both the concepts to represent the flow scale.

2.5.2 Response part

The response part is a measure that describes the degree of the attribute description. Important considerations are the way in which items are presented, and responses are obtained.

2.5.2.2 Semantic differential scale.

The semantic differential scale is a descriptive response format represented by describing words with opposite meaning at both ends and between these there are no intermediate points. Charles Osgood developed it in the 1950s (Osgood, 1952; ; Osgood et al., 1957). Currently, it is an established measurement tool used in many fields (e.g., psychology).

This idea that exists in the World Wide Web (WWW) for computerised data collection is not new and historical evidence by McReynolds and Ludwig (1987) reveal that at the beginning of the nineteenth century a device similar to semantic differential was existent.

2.5.2.1 Graded/discrete response format.

The most common type of scales among researchers is the Likert scale. For a Likert-type scale, the item's text is available with a response option associated to a numeric value. The Likert-type rating scale is used especially in psychological and health research using psychometric instruments. Likert scales are usually composed of five or more response categories. In HCI, discrete response formats are typically used rather than analogue ones (Gillan & Cooke, 1995). In order to fill the missing gap of a scientific justification for this choice, van Schaik and Ling (2003) conducted empirical research to establish the psychometric properties of questionnaires when discrete and analogue rating scales were used. The results showed the same pattern of results with both response formats.

2.5.2.3 Visual analogue scale/continuous scale.

Visual analogue Scale (VAS) are scales that have no intermediate scale points. They can be visualised as long straight lines with guidance to the directionality of the rating for the respondent. They are commonly used to indicate the intensity of pain. Visual analogue measures have existed over the years, but have become prominent, along with the Likert-type scale, with the increased use of computer administration. Van Schaik and Ling (2003; 2007) tested the effect design parameters of rating scales (Likert scale using radio button and visual analogue scale) on online questionnaires. The studies reported the same psychometric properties for each of the scales and there was no response bias associated with the visual analogue response format, which did not produce more extreme scores than the Likert format.

2.5.2.4 Comparison between discrete and continuous rating scales

Response formats each have their own advantages and disadvantages. Advantages of Likert and visual analogue scale formats reported by van Schaik and Ling (2003) is presented in Table 2.1.

Table 2.1. Advantages and disadvantages of Likert and visual analogue scale response formats (Source: van Schaik & Ling, 2003).

	Likert	Visual analogue scale
Advantages	Relatively easy to learn because all possible responses are presented	Effect of individual interpretation of Likert graduations avoided
	Relevant changes in scores more easily interpreted by researchers	Better match between subjective state and response through very large response range
Disadvantages	Poorer match between subjective state and response because of restricted range of responses	Difficulty in (learning to) use because of lack of indication of intermediate points (only end-points are displayed)
	Variability due to individual interpretation of Likert graduations	Extra work required to convert analogue responses into numeric scores after data collection

In the light of these considerations, it is not possible to conclude that one type of format is preferred over the other. However, when used intelligently, both formats yield reliable and valid scales. The research presented in this thesis will provide knowledge whereby the results of the experiments conducted in this study will help map the effect of response formats in online psychometrics.

2.6 Online psychometric measurement

The traditional method of psychometric measurement using paper and pen(cil) is rapidly diminishing while the method of administering via the Internet (online psychometrics) has gained acceptance for various reasons, such as increased efficiency and reduced expense of data collection (Birnbaum, 2004). This transformation in the mode of administration is supported by various factors such as saving time for the researcher (Schmidt, 1997) and eliminating costs involved in data collection (Buchanan & Smith, 1999; Hertel, Naumann, Konradt & Batinic, 2002). In the field of clinical psychology, it has been shown that psychometric properties observed in a paper and pencil mode of administration are not necessarily retained when transferred to Internet-administered measures (Buchanan, 2002). The practice of simply adopting a paper-and-pencil instrument for the Internet mode by assuming that the measurements are equivalent to the original mode is not recommended (Buchanan, 2002; Coles, Cook, & Blake, 2007; Hewson & Charlton, 2005; Noyes & Garland, 2008). Extensive research on the comparative study of both methods exist, providing evidence that access to larger, diverse samples enhance the validity of data collected online (Coles et al., 2007; Riva & Davide, 2003; Schmidt, 1997).

Several studies did not find substantial differences between web-based and paper-based modes of administration (Denscombe, 2006; Ritter, Lorig, Laurent, & Matthews, 2004; Van De Looij-Jansen & De Wilde, 2008). Most comparative results showed largely equivalent psychometric properties for the two administration formats (Cronbach's alpha between 0.79 and 0.95), along with high and significant correlations between the Internet and the paper-and-pencil versions (Carlbring, Brunt, & Bohman, 2007). Due to the flexibility of easy administration without any constraint on time and place, research participants prefer online administration (Naglieri, Drasgow, & Schmit, 2004; Pettit, 1999). In the research by Brock, Barry, Lawrence, Dey and Rolffs (2012) there were three aims: (a) to determine if the self-report questionnaires for paper-pencil administration remain reliable when administered over the Internet, (b) to examine quantitative equivalence of written and Internet methodologies and (c) to examine qualitative equivalence among measures across written and Internet methodologies. There were no significant differences in the test-retest reliabilities and internal-consistency reliability for all administered questionnaires was in the acceptable range (Cronbach's alpha ranging from .76 to .99). Similarly, quantitative equivalence was demonstrated across written and Internet administration. Thus, these results emphasise that online administration may be a reliable and valid alternative. The results of qualitative equivalence were found to be generally adequate thus increasing the confidence of researchers that they were assessing the constructs they intended to measure during online administration. Nevertheless, the research reported that instruments converted from the traditional method to be administered online must be examined for equivalence of all psychometric goals before administration.

2.7 Design parameters for online psychometric measurement

With the growing competition in the world market, websites have gained great importance for organisations (Roy et al., 2014) and there is a plethora of research on design parameters of websites for enhanced usability (Tarafdar & Zhang, 2005; Cebi, 2013; van Schaik & Ling, 2008; 2011). In the context of online questionnaires, administration via computers is becoming increasingly widespread; however, little is known about the impact of the design parameters on the psychometric properties of these measures (Norman, Friedman, Norman, & Stevenson, 2001; van Schaik and Ling, 2003, 2007; van Schaik et al., 2015). A failure to establish adequate psychometric properties in a typical Internet sample would suggest that the scale is not suitable for use in online studies. As an analogy, well-designed websites enhance users' interaction and hence design is crucial (van Schaik & Ling, 2001; Ling & Schaik, 2002, 2006). Similarly, a well-designed presentation of an online questionnaire should result in sound psychometric measurement. In the experiments conducted by van Schaik and Ling (2007), parameters of questionnaire designs were investigated for their effect on psychometric questionnaires. Important parameters included response format, questionnaire layout and interaction mechanism.

2.7.1 Response format

With the lack of research that studies instruments measuring the quality of human-computer interaction online, the experiment conducted by van Schaik and Ling (2007) contributed results about the equivalence of two response formats (Likert scale using radio buttons and visual analogue scale).

Even though a strong preference for Likert scale was observed, overall psychometric results regarding factor structure, reliability and validity for both response formats converged (van Schaik & Ling, 2007). Based on the results of the study by van Schaik and Ling (2007), the design parameters of response format will be investigated in this study to compare the results between small screen devices and desktop computers.

2.7.2 Questionnaire layout

The study by Norman, Friedman, Norman and Stevenson (2001) investigated four different ways of partitioning surveys for online presentation. In online psychometrics, research into human-computer interaction design of online psychometrics is still scarce (but see van Schaik and Ling, 2003; 2007; van Schaik et al., 2015). Research by van Schaik and Ling (2007) and van Schaik et al. (2015) investigated the presentation of questionnaire layout as whole-form (all items in one page) and as single-item presentation (one item per page). Both the studies reported consistently that items presented singly rather than as a whole-form exhibited sound psychometric properties. The flexibility of online questionnaires makes this a feasible option that would not be otherwise possible in paper-based questionnaires, due to the immense amount of paper requirement. The findings of these research studies that single-item presentation layout is faster with some advantage to psychometric structure will be adopted in this study.

2.7.3 Interaction mechanism

Two types of interaction mechanism are discussed in van Schaik and Ling (2007): direct and indirect.

In direct interaction, the user can immediately select from a set of visible options (e.g., radio buttons). In indirect interaction, the user can choose from a set that will become visible when interacting with the control, (e.g. drop-down list). The findings of the study exhibited little effect on psychometric properties of the questionnaires. However, it was reported that questionnaire items took longer to complete with indirect interaction than with direct interaction. Therefore, direct interaction will be used here.

2.8 Conclusion

From the literature review, it is evident that the results of the use of online psychometrics can be influenced by design factors such as presentation format (Norman et al., 2001; van Schaik & Ling, 2003, 2007; van Schaik et al., 2015). The current research focusses on creating a computer-based environment to support the systematic study of the effects of design parameters on the results of online psychometrics. An extensible system architecture in the form of a database model and a runtime system using the database is proposed. The system architecture for such an extensive study is detailed in Chapter 4. The development of hypotheses for the design parameters in online psychometrics is presented in the next chapter (Chapter 3), along with the literature review of the questionnaires that will be used in this research study. Since the research is conducted among Arabic speakers, Arabic translation of the questionnaires is required. However, at the time of this study, no translated version for the questionnaires was available and therefore the process of translating the questionnaires to be used in the study is also detailed in Chapter 3.

Chapter 3

Design Parameters in Online Psychometrics

Contents

3.1. Overview

With the application of psychometrics in developing and evaluating standardised usability and user experience questionnaires (Lewis, 2018), the design of online psychometrics becomes increasingly important to ensure good measurement properties. In contrast to research on survey design guidelines, very few studies exist on the design of online psychometrics (van Schaik & Ling, 2003, 2005a, 2005b, 2007; van Schaik, Wong, & Teo, 2015). This chapter begins with the development of a model of design parameters in online psychometrics, followed by the development of the hypotheses for the chosen design parameters that are manipulated in the research presented in this thesis. Following the hypotheses, a concise literature is presented on the review of the questionnaires that are used to measure the human-computer interaction for online psychometrics. The translation process of the questionnaires used in the experiments within this thesis is detailed. The chapter concludes with a summary.

3.2 Model of Design Parameters in Online Psychometrics

Among the various design parameters, font size, response target size, text/background colour and response format have been chosen and manipulated for Study 1 and Study 2 in this research. The effects of design parameters on interaction outcomes are tested and psychometric properties are evaluated. The results form the basis for design guidelines.

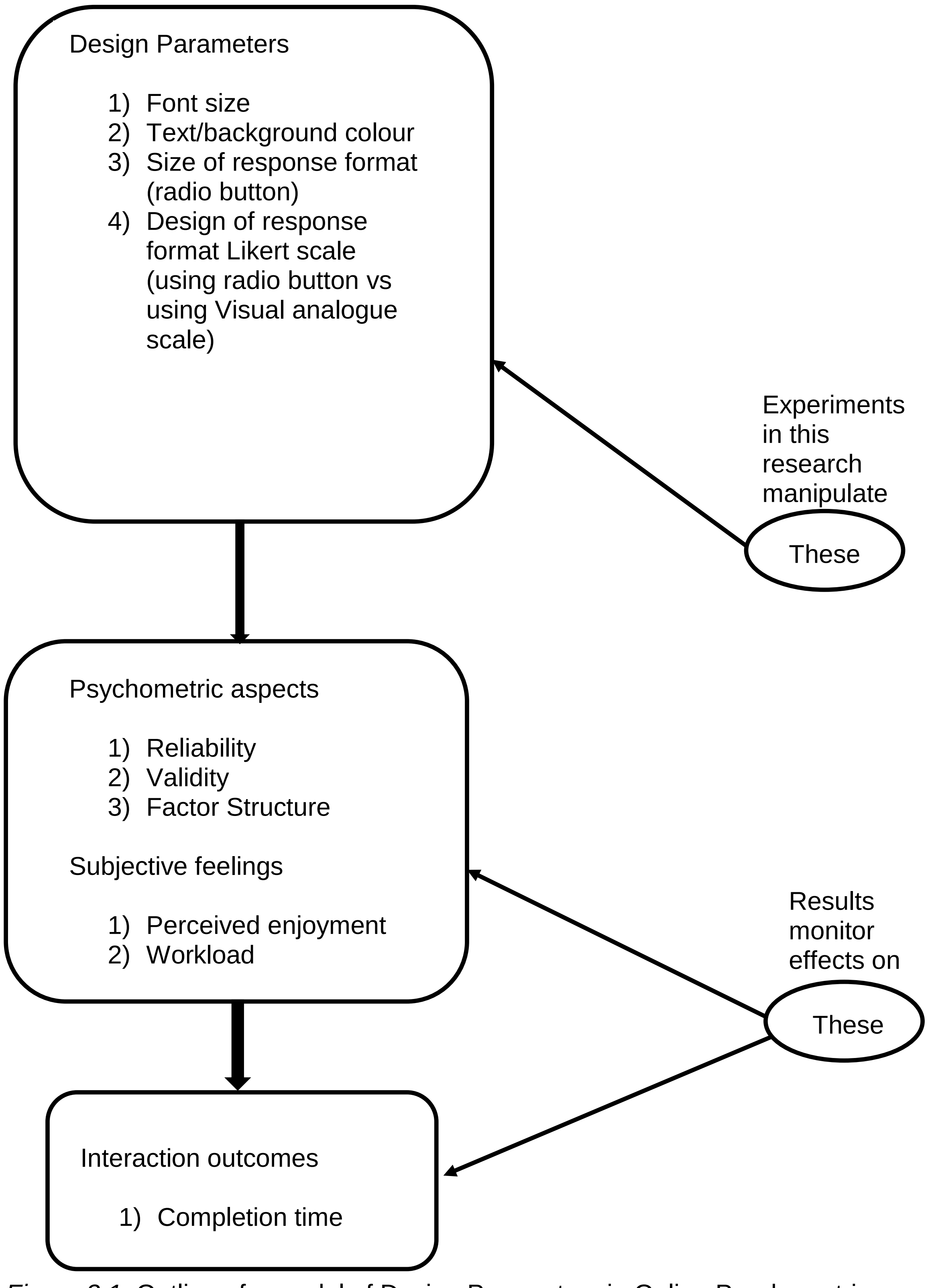


Figure 3.1. Outline of a model of Design Parameters in Online Psychometrics

3.3 Development of Hypotheses

In a study by Segall, Doolen and Porter (2005), efficiency is defined as the resources consumed to achieve a goal. Time is a resource of great interest to human-computer interaction (HCI): an efficient task will consume less of a user's time. Therefore, efficiency measurements often include time to complete a task, time to learn how to perform a task and time spent on recovering from errors. In this research, efficiency was quantified as the time required to complete a questionnaire or a set of questionnaires.

There is a lack of research on the design of online psychometric questionnaires (van Schaik & Ling, 2007). Among the various design parameters, font size, text/background colour, response target size and response format have been chosen and manipulated in the two studies, Study 1 and Study 2. This is because research suggests that these design parameters most likely affect reading comprehension (e.g. Ramadan, Mohamed, & El-Hariry, 2010) and task completion time (e.g. Parhi, Karlson, & Bederson, 2006; Van Schaik & Ling, 2007).

Study 1 is designed to test the effect of the design parameters font size, text/background colour and response target size on small-screen devices. Study 2 comprises of Study 2-A (on small-screen devices) and Study 2-B (on desktop computers). The design parameters font size, text/background colour – polarity and response format are manipulated in Study 2.

3.3.1 Font Size

A plethora of research exists to determine the best font size to maximise the readability of text on desktop computers while little research exists for small screen devices. Furthermore, little research exists that provides guidelines for online psychometrics on small-screen devices and desktop computers. Research by Darroch, Goodman, Brewster, and Gray (2005) for handheld devices recommend that designers creating applications for reading text on a small-screen with a resolution of 640 $\times$ 480 should offer the choice of small (font size 8 point), medium (font size 10 point), or large (font size 12 point) to accommodate the broadest range of users. The device used for research by Darroch et al. was the iPAQ a handheld computer designed to be used in mobile situations. In the context of desktop computers research by Banerjee and Bhattacharyya (2011) recommend font size 14 point as best for reading on computer screens. In another study conducted on 17-inch computer screens, Rello, Pielot and Marcos (2016) recommend using at least font size 18 point to optimise readability and comprehension of web text content. Research studies on readability of text other than the English language are very few. It is to be noted that in this research questionnaires are administered in Arabic. There are many differences between the English and Arabic texts. A main difference is that Arabic text written and read from right to left while English text is written and read from left to right. English text alphabets have an uppercase and lowercase while in Arabic, characters take different forms at different times (e.g. at the beginning of the word, middle of the word). In addition, Arabic text is cursive.

The characters are connected to each other with no spaces forming like a block (Al-Mutawa, 1999, as cited in Ramadan, 2011). Considering Arabic, among the few research studies that exist on Arabic font size and type, Ramadan et al. (2010) recommended Simplified Arabic font type, font size 13 to have a good performance for reading comprehension on desktop computers. Another study by Ramadan (2011) recommended Simplified Arabic font type, font size 14 point as the best combination for e-book reading. Furthermore, Abubaker and Lu (2012) found that font type Simplified Arabic and font sizes 14 point and 18 point were readable to pupils aged 9-12.

Because of the difference in screen sizes between desktop computers and mobile devices, a site, when viewed on a desktop computer, will look and behave differently from that same site when viewed on a smartphone. Indeed text with font size 14 point and 18 point readable on desktop computers, appear small and are hard to read when administered on small-screen devices. To avoid this problem, font sizes 36 point, 44 point and 64 point were chosen. In the current research, questionnaires are administered on both desktop computers and small-screen devices. Therefore, it seems logical to choose large font sizes for small-screen devices to match the font sizes on desktop computers. To test font size for online psychometrics on desktop computers and small-screen devices, the task of responding to a psychometric questionnaire was presented with two font sizes 36 point and 44 point in Study 1, and 44 point and 64 point in Study 2. Online questionnaires can also be presented on desktop computers. Then again, the font size selected for mobile devices will appear more prominent on desktop screens due to the difference in screen size (see Figures 3.2, 3.3, 3.5 and 3.6).

Figure 3.2. Font size 64 point on a desktop computer (21 inches diagonally).

Figure 3.3. Font size 64 point on mobile screen (5 inches diagonally).

Figure 3.4. Font size 20 point on desktop computer (21 inches diagonally)

Figure 3.5. Font size 44 point on a desktop computer (21 inches diagonally).

Figure 3.6. Font size 44 point on mobile screen (5 inches diagonally).

I felt disoriented

Figure 3.7. Font size 12 point on desktop computer (21 inches diagonally)

Therefore, font sizes 12 point and 20 point were chosen for the desktop Study 2-B and font sizes 44 point 64 point were chosen for the mobile Study 2-A.

Because larger font size increases readability, completion time may be faster with large font size. Therefore,

Hypothesis 1: completion time decreases with increasing font size.

3.3.2 Response Target size

For survey input design, research by Stapleton (2013) recommends vertical radio buttons as the input type for all questions on mobile devices because this input type leads to less biased data and is displayed consistently on mobile devices. In the context of online psychometrics on desktop computers, a study conducted by van Schaik and Ling (2007) resulted in the same psychometric properties of scales when Likert (radio buttons) and the visual analogue scale was used both in horizontal layout. Van Schaik and Ling (2007) also report that presentation of single items with direct interaction mechanism produced faster completion time of questionnaires. As before, the main difference between desktop computers and mobile devices is the screen size. Response options in the horizontal layout, especially when radio buttons are used, are close to each other, which may cause respondents to unknowingly select the wrong response option (Jones, Marsden, Mohd-Nasir, Boone, & Buchanan, 1999; Parush & Yuviler-Gavish, 2004).

A consideration with the touch response mode for mobile devices is that the response option target size (width of the radio button) must be designed according to the size of the human fingertip. Table 3.1 details the target sizes in different units. A study by Mizobuchi, Ren and Yasumura (2002), on small-screen device a PDA with a pen, showed that with a minimum target size of 5 mm (Table 3.1), the pen (stylus) could point to targets quicker than with smaller target sizes. The minimum target size of 5 mm was also confirmed in a study by Brewster (2002), in which the participants used a Palm III handheld computer with a stylus. In the study by Brewster, the target size could be reduced with the help of sound to 2.5 mm (Table 3.1), but the mental workload then increased with the smaller target size. Thus the minimum target size that was recommended was 5 mm.

Table 3.1. Target sizes in different units

Millimetre (mm)[a]	Pixels (px)[a]	Point[a]
2.5	9.45	7
3.8	14.36	11
5.0	18.90	14
5.8	21.92	16
6.5	24.57	18
7.0	26.46	20
7.7	29.10	22
9.2	34.77	26
9.6	36.28	27
10.0	37.79	28
11.5	43.46	32
12.7	48.00	36
15.5	58.58	44
22.6	85.42	64

Note: [a] Calculated on screen size: 21.5 inch (diagonal) display

Fitts's model for motor movement (1954), defined the time required to quickly move to a target area as function of the ratio between the distance to the target and the width of the target. A study by MacKenzie and Zhang (2001) based on Fitts' law tested text entry rates on two sizes of soft keyboards (target size: 6.5 mm, 10 mm Table 3.1). It was concluded that it was harder to press keys the farther they are but more comfortable to press keys the bigger they are. Parhi et al. (2006) conducted a study that examined interaction between target size and task types such as pointing tasks (activating buttons, radio buttons or checkboxes), a serial sequence of taps (text entry) for one-handed use of touchscreen-based handhelds. The target sizes used in their study were 3.8, 5.8, 7.7, 9.6 and 11.5 mm (Table 3.1). It was observed that time-on-task decreased significantly as target sizes grew, but the error rate did not. Thus, they recommended that target sizes should be at least 9.2 mm (Table 3.1) for single-target tasks such as a tap and at least 9.6 mm (Table 3.1) for serial tasks such as text entry. Similarly, on touchscreen handheld devices, Perry and Hourcade (2008) examined whether targets on the edge of the screen enable participants to be more accurate in selection, than targets not on the edge. For this, the target sizes reported in Parhi et al. (2006) were chosen. Although the authors did not explicitly recommend the target size, their research reported that the best result was obtained with the largest target size 11.5 mm. Microsoft's Windows Phone Silverlight development guidelines recommend that a touch target size greater than or equal to 9 mm square is acceptable. Microsoft allows their developers to use 7 mm (Table 3.1) square as a minimum target size when a smaller hit target is warranted (as cited in Umami, Arezes, & Sampaio, 2016).

The minimum target size recommended by Microsoft is the same as the size recommended by Google. Google’s Android UI Guidelines suggest 7 mm square (as cited in Umami et al., 2016). The iOS Human Interface Guidelines allow a 6.5 mm (Table 3.1) square for their developers (as cited in Umami et al., 2016).

In this research, to test the questionnaire completion time, depending on the response target size when administered on small-screen devices, participants were asked to perform the task of responding to a set of psychometric questionnaires under different target sizes. The target sizes reported in Parhi et al. (2006) and Perry and Hourcade (2008) were used in experiments conducted on small-screen devices that had a screen dimension of 3.5 inches diagonally. In this research, the screen size of the devices is a minimum of 4.7 inches or larger diagonally. Response target sizes greater than the target sizes reported in the study by Parhi et al. (2006) were chosen for this research (12.7 mm, 15.5 mm, 22.6 mm). These target sizes, closely matched the chosen font sizes (36 point, 44 point, 64 point). Target size of width 7 mm appears to be small on screen size 4.7 inches and therefore will give rise to difficulty for selection, as it is evident from all of the studies mentioned, that larger target sizes improved the speed rate of the task and reduced errors. Applying Fitts's law (1954) in Study 1 predicts that as the response target size (radio button) increases, the selection time is faster. Therefore,

Hypothesis 2: completion time decreases with larger response target size.

3.3.3 Text/Background polarity

A plethora of research studies exist examining text/background colour combinations on computer screens while the literature on colour combinations for small-screen devices is scarce. Often, best colour combinations suitable for the desktop view are also applied for small-screen devices. However, some established colour combinations may not always be best suited. Furthermore, there is a lack of research on the Arabic language compared to the English language. In the context of online psychometrics on small-screen devices, systematic research of colour combinations for the text/background design parameter begins with relevant work obtained from previous research studies. Contrast is essential in any written text. Without contrast, reading is not possible. Whether in print or on-screen displays, low contrast can be irritating and fatiguing to readers. Contrast is the value (intensity) difference between two areas; the value is the amount of lightness or darkness in colour. Text can be depicted by luminance contrast (i.e., differences in luminance between characters and background) or by colour contrast (i.e., differences in chromaticity) (Legge, Parish, Luebker, & Wurm, 1990). The text can be well-read when colours with high contrast are used (Neilsen, 2000). There are various studies on text/background colour combination for English text in various contexts on computer screens (e.g., Bonnardel, Piolat, & Le Bigot, 2011; Buchner, Mayr, & Brandt, 2009; Gradišar, Turk, & Humar, 2010; Greco, Stucchi, Zavagno, & Marino, 2008; Grobelny & Michalski, 2015; Hall & Hanna, 2004; Humar, Gradišar, Turk, & Erjavec, 2014; Lin, Wu, & Cheng, 2013; Ling & van Schaik, 2002; Pearson & Van Schaik, 2003; Rello & Baeza-Yates, 2017; Timpany, 2009).

The most often recommended web design guideline for appropriate colour combinations between text and the background colour is the traditional black on white. Black on white colour combination has the highest contrast (ratio: 21.00:1). Colour contrast ratio is given by the luminance of the brightest colour (e.g. white) to that of the darkest colour (e.g. black). With regards to primary colours (red, blue, green), Kaya and Epps (2004) stated that colours are often described in temperature terms such as “cool” (e.g. blue, green) and “warm” (e.g. red). The authors further describe that colour-induced associations have the potential to activate alternative types of regulatory focus. Regulatory focus theory suggests that people can achieve their goals in two different ways, either with a promotion or a prevention regulatory focus (Higgins, 1987). Based on these studies, research by Mehta and Zhu (2009) concluded that blue colour activated a promotion focus and thus enhanced performance on creative tasks. Similarly, research by Ling and van Schaik (2002) and Timpany (2009) recommend using blue text on white background (ratio: 8.59:1). The research studies by Buchner and Baumgartner (2007) and Buchner et al. (2009) concluded that reading text from computer screens is better when text is printed in dark letters on a light background (positive polarity – black on white) than when it is printed in light letters on a dark background (negative polarity – white on black). Even though the positive polarity advantage seems to be a robust phenomenon (Piepenbrock, Mayr, & Buchner, 2014; Piepenbrock, Mayr, Mund, & Buchner, 2013) the empirical basis for this conclusion is limited to mostly for proofreading performance. The study by Piepenbrock et al. (2014) also reported that reading was faster in the positive polarity condition than in the negative that more words were read with increasing character size.

According to (Piepenbrock et al., 2014) the implications of positive polarity are important for the design of text on displays as those of computers, automotive control and entertainment systems, as well as smartphones that are increasingly used for the text-based media and communication. Thus, polarity is important for the use of both small-screen devices and desktop computers in online psychometrics. In this research, it is to be noted that questionnaires were administered in Arabic. For the Arabic text, there exists little research regarding text/background colour combinations. Among these research studies, Ramadan (2011) conducted a study on reading e-book materials. Four text-background colour combinations (white-on-black, black-on-white, blue-on-white, and white-on-blue) were studied in their work. The study concluded for reading speed with better comprehension black-on-white was the best combination. The study also reported that in terms of preference rating, along with black-on-white, participants also preferred blue-on-white. As far as the author is aware, there is no research available on the effect of text/background colour for online psychometrics especially on mobile devices and in Arabic. Most of the literature recommends colour combinations for faster reading, and therefore it is essential to test the effect of text/background colour on task completion time in online psychometrics. For this research, the background colour white and two font colours (black and blue) have been chosen for Study 1. For Study 2, black text on white background and white text on black background combinations have been chosen. The author has based these choices on the research studies by Ling and van Schaik (2002), Ramadan (2011) and (Piepenbrock et al., 2014).

These studies identified various combinations such as blue-on-white (Ling & van Schaik, 2002; Ramadan, 2011) and black-on-white (Piepenbrock et al., 2014; Ramadan, 2011) as beneficial for accuracy, reading speed, the speed of visual search as well as user preference. Given that high contrast between text and background improves readability,

Hypothesis 3a: completion time decreases with increasing colour contrast.

Hypothesis 3b: positive text/background polarity decreases completion time.

3.3.4 Response format

The development of graphic rating scales (Hayes & Patterson, 1921) brought about the visual analogue scale for continuous rating. Likert scales were proposed in 1932 (Likert, 1932). Research is available on the effectiveness of rating scales in different fields of study and survey administration. In human-computer interaction (HCI), discrete response formats are more commonly used than continuous ones (Gillan & Cooke, 1995). There exists a lack of empirical evidence for the choice of discrete response format over continuous ones in HCI. The study by van Schaik and Ling (2007) compared the two response formats used in three multi-item instruments (such as disorientation and perceived ease of use developed by Ahuja and Webster (2001) and the flow questionnaire developed by Davis and Wiedenbeck (2001). The authors also conducted psychometric analysis and reported finding similar results for both response formats. Van Schaik and Ling further reported that the direct interaction mechanism (Likert scale) produced faster completion of questionnaires and was therefore recommended. Then again, the effect of the design parameters of rating scales for online psychometrics was not tested on small-screen devices.

Little research comparing Likert scale and visual analogue scale on small-screen devices for surveys exist (Buskirk, Saunders, & Michaud, 2015; Funke, 2016; Funke, Reips, & Thomas, 2011; Stapleton, 2013). Therefore, this research addresses the research gap that exists on the effect of response format design in online psychometrics especially on small-screen devices.

Conventionally, the visual analogue scale employs the slider bar that makes use of the ‘drag-and-drop’ principle. With regards to ease of use in this research, (responding to 43 items on a small-screen device), the ‘drag-and-drop’ principle seemed problematic. However, according to Toepoel (2017), a better way to use visual analogue scale is to employ the ‘point-and-click’ principle. Therefore, visual analogue scale with a slight variation was adopted with point-and-click. Tick marks indicating five response options were provided (Figure 3.8). Radio buttons are round button images that can be clicked to provide an answer (Figure 3.9).

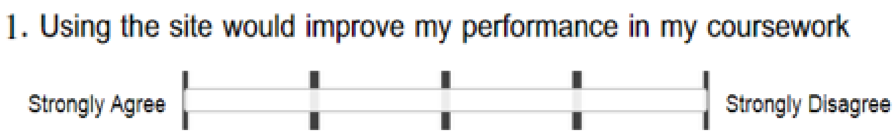


Figure 3.8. Likert scale using visual analogue scale.

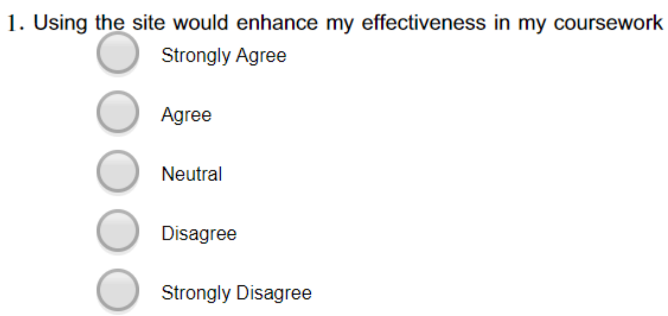


Figure 3.9. Likert scale using radio button.

However, they differ in their orientation in this research as presented in Figures 3.8 and 3.9. The difference in orientation is because the radio button options (target sizes: 15.5 mm, 22.6 mm) when horizontally placed, overlap each other and when placed horizontally with spaces in between do not fit on one line (due to the target size width). Thus, the radio buttons were placed vertically (Figure 3.9). The visual analogue scale is horizontally oriented and needs a limited amount of space. Therefore, it especially suits the limited space available onscreen for small-screen devices. Research on the influence of response format on completion time particularly for surveys on desktop computers and mobile devices exist (e.g., Couper, Traugott, & Lamias, 2001; Funke et al., 2011). In online psychometrics, research by van Schaik and Ling (2007) provided empirical evidence for faster completion time of questionnaires with direct interaction (radio button format). It is evident that empirical evidence to address the research gap that exists on the effect of response format design in online psychometrics both on small-screen devices and desktop computers for task completion time is required. Based on existing research, in the field of surveys, it is evident that the preference of visual analogue scale mostly implemented with the drag-drop principle comes with longer completion times. Although, in this research the visual analogue scale is implemented with the point-and-click principle, the author proposes

Hypothesis 4: completion time of Likert scales is shorter with radio button format than visual analogue scale format.

3.3.5 Cognitive Load

Measuring response times has enjoyed a long tradition in social psychology and survey research (Couper & Kreuter, 2013; Yan & Tourangeau, 2008) and has been proven as a useful strategy to investigate cognitive effort (Bassili & Fletcher, 1991; Bassili & Scott, 1996; Fazio & Olson, 1990; Yan & Olson, 2013). It is generally assumed that the time of processing corresponds (directly) to the cognitive effort required to answer a question (Höhne, Schlosser, & Krebs, 2017). The inference is that the longer a respondent needs to respond, the higher the cognitive effort must be. Therefore, from Hypotheses 1, 3b and 4 follows,

Hypothesis 5: workload decreases with increasing font size.

Hypothesis 6: positive text/background polarity decreases workload.

Hypothesis 7: workload is lower with radio button format than with visual analogue scale format.

3.3.6 Perceived Enjoyment

A study by Davis, Bagozzi, and Warshaw (1992) defines perceived enjoyment as the degree to which the activity of using technology is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated. Research findings of Venkatesh, Morris, and Davis (2003) support the essential role of perceived enjoyment in the Technology Acceptance Model (TAM) and the correlation between perceived ease of use and perceived usefulness. As much as the author is aware, perceived enjoyment has not been used as an external variable relating to respondents’ experience in online psychometrics.

In this research, the effects of the variables: font size, font colour and response format on respondents' experience is analysed. As discussed in Section 3.3.5, it is assumed that the faster completion, the lower cognitive effort. When little cognitive effort is required, it is apprehended that perceived enjoyment will be higher. Therefore, the author proposes from Hypotheses 1, 2,3a, 3b and 4

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 9: perceived enjoyment increases with increasing response target size;

Hypothesis 10a: perceived enjoyment increases with increasing colour contrast;

Hypothesis 10b: positive text/background polarity enhances perceived enjoyment;

Hypothesis 11: perceived enjoyment is higher with Likert scale using radio button format than with visual analogue scale.

3.4. Scales that measure the quality of human-computer interaction

3.4.1. Disorientation, Perceived Ease of Use, Perceived Usefulness

Disorientation as described in (Woods, 1984), occurs when "the user does not have a clear conception of relationships within the system" (Woods, 1984, p. 229). In other words, according to McDonald and Stevenson (1998), disorientation, is the tendency to lose the sense of location on a Web site. This is one of the most common problems faced by users navigating through hypertext, and it can lead to frustration, loss of interest and a decline in efficiency. Therefore, it is plausible to measure the users' perception of the quality of information architecture through perceived disorientation (Ahuja & Webster, 2001).

In the technology-acceptance literature, two main concepts are perceived usefulness and perceived ease of use.

Perceived usefulness is defined as “the degree to which a person believes that using a particular system would enhance his or her job performance” (Davis, 1989, p. 320). Perceived ease of use is defined as “the degree to which a person believes that using a particular system would be free of effort” (Davis, 1989). Perceived ease of use and perceived usefulness were investigated by Davis (1989) with the conclusion that they formed two key concepts in HCI within the framework of the Technology Acceptance Model.

The instruments perceived ease of use (3 items), perceived usefulness (4 items) and disorientation (7items) used 5-point scales, with the terms "Strongly agree", "Strongly disagree” as the endpoints. The study by van Schaik and Ling (2005a) studied perceived disorientation, perceived ease of use and perceived usefulness questionnaires to measure the quality of human-computer interaction. The psychometric properties of these questionnaires were confirmed in their research. Furthermore, research studies by van Schaik, Luan, and Teo (2015), van Schaik & Ling (2005, 2007) also confirmed the psychometric properties of these questionnaires while manipulating design parameters of rating scales and questionnaire layout in online psychometrics. It is therefore within the scope of the current research study to establish the psychometric properties of the questionnaires while manipulating various other design parameters using an extensible system architecture based on the studies of van Schaik and Ling (2003, 2005a, 2005b, 2007) and van Schaik et al. (2015).

The factor analysis in the studies by Ahuja and Webster (2001), van Schaik and Ling (2005a, 2005b, 2007; van Schaik et al., 2015) consistently resulted in distinct factors for disorientation, perceived ease of use and perceived usefulness. The study by van Schaik and Ling (2003, 2007) also reported that the factor structure of the questionnaires were similar when the two response formats Likert scale and visual analogue scale were used. The reliability results of the three questionnaires DIS, PEU and PU, to measure disorientation, perceived ease of use and perceived usefulness, are reported in Table 3.2.

Table 3.2. Estimates of reliability for disorientation, perceived ease of use and perceived usefulness from previous research studies.

Study	DIS		PEU		PU	
Ahuja and Webster (2001)	0.89		0.87			
van Schaik and Ling (2005a)	0.89		0.87		0.97	
van Schaik and Ling (2005b)	0.92		0.88		0.95	
	SI	WF	NA		NA	
van Schaik et al. (2015)	0.89	0.91				
	LS	VAS	LS	VAS	LS	VAS
van Schaik and Ling (2003)	0.87	0.89	0.88	0.89	NA	NA
van Schaik and Ling (2007)	0.92	0.94	0.87	0.88	NA	NA

Note: SI: single item. WF: whole form. LS: Likert scale. VAS: Visual analogue scale. NA: Not applicable (not used in the study).

The validity (correlation) results of the three questionnaires DIS, PEU and PU are reported in Table 3.3.

Table 3.3. Estimates of validity (correlation) for disorientation, perceived ease of use and perceived usefulness from previous research studies.

Study	DIS-PEU	PEU-PU	PU-DIS
Ahuja and Webster (2001)	*r* = -0.30	NA	NA
van Schaik and Ling (2003)	*r* = -0.343	NA	NA
van Schaik and Ling (2005a)	NA	*r* = 0.47	NA
van Schaik and Ling (2005b)	NA	*r* = 0.45	*r* = -0.27
van Schaik and Ling (2007)	*r* = -0.246	NA	NA
van Schaik et al. (2015)			

NA: Not applicable.

3.5. PSSUQ

To test the assessment of usability following participation in task-based usability tests, several standardised usability questionnaires were developed from the 1980s (Lewis, 2018). The PSSUQ (Post-System Study Usability Questionnaire) was developed by James Lewis (1995,2002) in a major company – International Business Machines Corporation. The PSSUQ is a 19-item questionnaire that assesses user satisfaction with system usability (Lewis 1995, 2002). The items are 7-point scales, with the terms "Strongly agree" for 1, "Strongly disagree" for 7 as the endpoints on the scale. During the construction of the questionnaire, the items loaded on three factors and a group of human factors engineers named the factors as System Usefulness (SYSUSE), Information Quality (INFOQUAL), and Interface Quality (INTERQUAL) (Lewis 1995). The psychometric evaluation of the PSSUQ was once again conducted using data from five years of usability studies by Lewis (2002). Validity was confirmed consistent with results established previously (Lewis, 1995; 2002).

Research by Frughling and Lee (2005) validated the PSSUQ instrument using a larger sample size in a different domain, following the recommendation in the research studies by Lewis (1995, 2002), that it would be pragmatic to collect data in different circumstances and extend the generalizability of the findings.

The Factor analysis results from Lewis' research (1995) indicated there are three-factor sub-scales: System Usefulness, Information Quality, and Interface-Quality. It was reported that these three factors accounted for 87% of the variability in the data. Similarly, the factor analysis results from Lewis' research (2002) also indicated the same three-factor sub-scales. It was reported that the three-factor solution explained 72.5% of the variance in the data. The research by Frughling and Lee (2005) failed to replicate the three-factor sub-scale structure. Instead, Frughling and Lee (2005) reported that the factor analysis resulted in two-factor sub-scales. The sub-scales were named as System Quality and System Usefulness. The research reported that the two-factor solution explained 71.81% of the total variance. Reliability results and factor analysis results from previous studies are reported in Table 3.4.

Table 3.4. Estimates of reliability for PSSUQ from previous research studies.

Study	Overall	SysUse	InfoQual	IntQual
Lewis (1995)	0.96	0.96	0.92	0.83
Lewis (2002)	0.97	0.96	0.91	0.91

Validity (correlation) from previous studies is reported in Table 3.5.

Table 3.5. Estimates of validity (correlation) for PSSUQ from previous research studies.

Study	SysUse-InfoQual	SysUse-IntQual	InfoQual-IntQual
Lewis (1995)	0.71	0.68	0.64
Lewis (2002)	0.72	0.67	0.56

The study by Frughling and Lee (2005) reported that all Cronbach's coefficient alphas exceeded the generally accepted minimum value of .70, demonstrating satisfactory evidence of internal consistency. The reliability coefficient alpha of the two factors in the study by Frughling and Lee were higher than .90. The System Quality reliability coefficient alpha (.933) was higher than those of both the PSSUQ Information Quality (.91) and the PSSUQ Interface Quality (.91). The System Usefulness reliability coefficient alpha (.958) was very close to that of the PSSUQ System Usefulness (.96). The research also reported that the correlation analysis supported the validity of the scales.

3.6. SUS

An increase in the application of human factors psychology to the design and evaluation of office and personal computer systems was seen as early as the 1980s. SUS was one of the first usability questionnaires to be developed (as mentioned by its developer John Brooke, 2013, p. 29). It was developed in Digital Equipment Corporation and was later published in a book on usability engineering (Brooke, 1996).

A number of published translations of the SUS in various languages such as Slovene, Polish, Italian, Persian, and Portuguese are available since 2014, and the latest published study of the Arabic translation was in 2017 by AlGhannam, Albustan, Al-Hassan, and Albustan.

The 10-item SUS was designed to be a unidimensional construct (Brooke 1996). It was reported in (Lewis & Sauro, 2017) that researchers started publishing results showing that the SUS had acceptable reliability and validity. The published results, showed evidence of a two-factor structure (e.g. Bangor, Kortum, & Miller, 2008; Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009). The two-factor structure had Items 4 and 10 aligning on one factor, while all other items loaded on the second factor. Lewis and Sauro named the two factors Usability (all items except 4 and 10) and Learnability (items 4 and 10). However, since 2009, research studies consistently reported the two-factor structure but failed to replicate the two factors (Usability and Learnability) that seemed apparent in 2009 (Kortum & Sorber, 2015; Lewis, Brown, & Mayes, 2015; Lewis, Utesch, & Maher, 2013, 2015; Sauro & Lewis, 2011). The research studies reported SUS to be a bi-dimensional construct with the odd-numbered items loading on one factor and the even-numbered items on another factor. The odd-numbered items had a positive tone, and thus the factor was named Positive while the factor with the even-numbered items that had a negative tone was named Negative. The reliability and validity results of the English version of SUS and other languages are reported in Table 3.6.

Table 3.6. Estimates of reliability for SUS from previous research studies.

Study (English version)	Reliability (coefficient alpha)	Concurrent validity (correlation)
Bangor et al. (2008)	0.91	0.81
Lewis and Sauro (2009)	0.92	NA
Finstad (2010)	0.97	0.96
Sauro and Lewis (2011)	0.92	NA
Lewis et al. (2013)	0.89	0.90, 0.81
Lewis et al. (2015)	0.90	0.50, 0.63
Kortum and Sorber (2015)	0.88	NA
Berkman and Karahoca (2016)	0.83	0.74
Lewis and Sauro (2017)	0.91	NA
Lewis (2018)	0.93	0.83
Borsci et al. (2009) (Italian version)	0.81	0.45, 0.66
Dianat et al. (2014) (Persian)	0.79	NA
Blažica and Lewis (2015) (Solvene)	0.81	0.52
Borkowska and Jach (2016) (Polish)	0.81	0.82
AlGhannam et al. (2017) (Arabic)	0.82	NA

3.7. Interaction-experience questionnaires

3.7.1. Perceived Enjoyment properties as explained for the PSSUQ.

The perceived-enjoyment scale, PE, that is used in the current research was developed by (Davis, Bagozzi, & Warshaw, 1992). Perceived enjoyment has been used and measured in organisational behaviour research (Venkatesh & Speier, 1999), HCI research (Venkatesh & Speier, 2000) and Information Systems research

(Brown & Venkatesh, 2005; Heijden, 2004; Venkatesh, 2000; Venkatesh, Thong, & Xu, 2012). Perceived enjoyment can be associated to the pleasure from the use of a particular technology at a particular time. Thus it may change dramatically over time and across systems. The uni-dimensionality of the scale was repeatedly confirmed and validity of the perceived enjoyment scale was also supported with factor loadings > 0.70 (Brown & Venkatesh, 2005; Fred Davis et al., 1992; Heijden, 2004; Visawanath Venkatesh & Speier, 1999, 2000; Viswanath Venkatesh et al., 2012). Reliability values from previous research is reported in Table 3.7.

Table 3.7. Estimates of reliability for PE from previous research studies.

Study	Reliability (Cronbach's alpha)
Davis (1992) Study 1, Study 2	0.81, 0.92
Venkatesh and Speier (1999)	> 0.80
Venkatesh and Speier (2000)	> 0.90
Venkatesh (2000) Studies 1, 2a, 3, and 2b at each of the three points of measurement	0.90, 0.92, 0.93
van der Heijden (2004)	0.86
Brown and Venkatesh (2005)	0.85
Venkatesh et al. (2012)	0.86

3.7.2. NASA-TLX

Among the many subjective procedures that exist to measure the mental workload, a well-known measure is the NASA-TLX (Task Load Index) (Hart & Staveland, 1988). The National Aeronautics and Space Administration-Task Load Index (NASA TLX), a multi-dimensional rating procedure, provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demand, Physical Demand, Temporal Demand, own Performance, Effort and Frustration (Hart & Staveland, 1988). Mental workload is a measure of usability when the task involves continuous demand on a user's attention for monitoring and control (Brewster, 2002). The NASA TLX has been used in a variety of fields (Hart, 2006). It has been used in studies involving the evaluation of visual and auditory displays, vocal and manual input devices, and virtual/augmented vision (Cao, Chintamani, Pandya, & Ellis, 2009).

With regards to reliability and validity, Hart and Staveland (1988) reported that NASA-TLX is more pragmatic and less sensitive to individual differences. Other researchers have also established the reliability and validity of the NASA-TLX questionnaire (e.g., Battiste & Bortolussi, 1988; Hill et al., 1992).

A limitation noted by Hart (2006) was the interpretation of the scores and in the research Hart further noted that an analysis of the vast amount of published data could remove this limitation. Following this, Grier (2015), presented a study that defines the range and cumulative frequencies of over 1000 global NASA Task Load Index published scores from over 200 publications, that will enable practitioners to state the percentage of scores that have been reported as higher or lower than the observed score.

As much as the author is aware there is no literature explicitly measuring workload in online psychometrics and knowledge of workload induced by design parameters in online psychometrics is essential for user experience.

Therefore, in this research, the observed score is mapped against the deciles and quartiles of the global NASA-TLX analysis adapted from the research by Grier (2015). Table 3.8 presents the percentiles and the scores.

Table 3.8. The deciles and quartiles of global NASA-TLX analysis (Source: Grier, 2015).

Percentile	Score
Min	6.21
10%	26.08
20%	33.00
25%	36.77
30%	39.45
40%	45.00
50%	49.93
60%	53.97
70%	58.00
75%	60.00
80%	62.00
90%	68.00
Max	88.50

3.8. Usability questionnaires in Arabic

3.8.1. Need for translation

Standardised usability questionnaires are an important tool in usability research (Kirakowski & Murphy, 2009). When questionnaires are available only in English, they are useful only to people who are fluent in English.

According to various studies, cultural differences between English speakers and non-native speakers may affect their validity (Finstad, 2006). Therefore, the motivation for translating and validating these questionnaires is the potential to extend the use of these questionnaires to populations beyond English-speakers.

3.8.2 Arabic-translation process

The committee approach described by Simonsen and Mortensen (1990) was adopted for the translation process. The committee consisted of three academics all proficient bilingual speakers fluent in Arabic and English. The author initiated the translation process and handed over the questionnaires to two committee members for independent translation from English to Arabic. The author then received the translations and then facilitated a meeting between the two translators to ensure conceptual equivalence and clarity in the translation in case of any difference. After this, the author consulted the third member of the committee who then executed the back translation, and minor changes were made to reflect the most appropriate translation for each item. All the five psychometric questionnaires: DIS (7 items), PEU, PU (3 and 4 items), the PSSUQ (19 items) and the SUS (10 items) were translated. The interaction-experience questionnaires perceived enjoyment (3 items) and the workload (NASA-TLX) were also translated. A listing (Table 3.9) of all the original items, translation and the appropriate back-translation is provided for reference. It should be noted that the back-translation reported in this thesis was provided by a fourth member outside of the translation committee, who is a native Arabic speaker with acceptable English language skills.

Back-translation is an important second step in the translation process and is needed to identify any discrepancies between the meaning of the translation and the original questionnaire. Back-translation is conducted by a translator who has not seen the original questionnaire (Bradley, 2013). Thus in the current research, a fourth member outside of the translation committee was consulted to confirm and compare the translation by someone who was not exposed to the original questionnaire in English. Therefore, the limitation observed is that, the words may not map onto the same original words in English although the meaning is conveyed. The back-translation of Disorientation, perceived ease of use and perceived usefulness, is best compared to PSSUQ and SUS.

Table 3.9. Forward and backward translation of the questionnaires used in this research.

	Original Questionnaire Item in English	Forward Translation to Arabic		Backward translation into English
NASA-TLX				
1	How mentally demanding was the task?	كيف كان المجهود العقلي او الإدراكي المطلوب لهذه المهمة؟	1	How was the mental effort needed for this task?
2	How physically demanding was the task?	كيف كان المجهود البدني المطلوب لهذه المهمة؟	2	How was the physical effort needed for this task?
3	How hurried or rushed was the pace of the task?	كيف كانت سرعة وتيرة المهمة؟	3	How was the speed pace of the task?
4	How successful were you in accomplishing what you were asked to do?	كيف كنت ناجحا في تحقيق ما طلب منك؟	4	How you were successful in accomplishing what has been required from you?
5	How hard did you have to work to accomplish your level of performance?	كم من المجهود بذلت لتحقيق هذا المستوى في ألأداء؟	5	How much effort you did to reach this level of performance?
6	How insecure, discouraged, irritated, stressed, and annoyed were you?	كيف أحسست بعدم الأمان، التثبيط أو عدم التشجيع، الانزعاج، القلق أو الضيق؟	6	How did you feel insecure, Inhibition, annoyance, Anxiety or distress?

PSSUQ					
1	Overall, I am satisfied with how easy it is to use this system.	بشكل عام، أنا راض عن سهولة استخدام النظام	1	Generally, I am satisfied with the easiness of using the system.
2	It was simple to use this system.	لقد كان من السهل استخدام النظام	2	It was easy to use the system.
3	I could effectively complete the tasks and scenarios using this system.	لقد استطعت الانتهاء من المهام وماهو مطلوب على نحو فعال باستخدام هذا النظام	3	I was able to finish the tasks and what was required effectively using this system.
4	I was able to complete the tasks and scenarios quickly using this system.	لقد تمكنت من الانتهاء من المهام وماهو مطلوب بشكل سريع باستخدام هذا النظام	4	I was able to finish the tasks and what was required quickly using this system.
5	I was able to efficiently complete the tasks and scenarios using this system.	لقد تمكنت من الانتهاء من المهام وماهو مطلوب على نحو فعال باستخدام هذا النظام	5	I was able to finish the tasks and what was required efficiently using this system.
6	I felt comfortable using this system.	لقد شعرت بالراحة عند استخدام هذا النظام	6	I felt comfortable when I used this system.
7	It was easy to learn to use this system.	لقد كان من السهل تعلم كيفية استخدام هذا النظام	7	It was easy to learn how to use this system.
8	I believe I could become productive quickly using this system.	أعتقد أنه بإمكاني أن أكون منتجة بشكل أسرع باستخدام هذا النظام	8	I think I can be quickly productive using this system.
9	The system gave error messages that clearly told me how to fix problems.	النظام قدم رسائل تبليغ عن الخطأ والتي أبلغتني كيف أتمكن من تصليحه بشكل واضح	9	The system gave me error messages which informed me how I can clearly fix them.
10	Whenever I made a mistake using the system, I could recover easily and quickly.	كلما قمت بخطأ أثناء استخدام النظام، استطعت أن أستعيده بشكل سريع وسهل	10	When I do a mistake while using the system, I was able to get back quickly.

11	The information (such as on-line help, on-screen messages, and other documentation) provided with this system was clear.	المعلومات (مثل المساعدة عبر الانترنت، الرسائل التي تظهر على الشاشة، وغيرها من الملفات الموثقة) التي تم تزويدها لهذا النظام كانت واضحة	11	The information (like help on-line, messages that appears on the screen, and other kinds of verified files) which has been provided with this system was clear.
12	It was easy to find the information I needed.	لقد كان من السهل العثور على المعلومات التي أحتاجها	12	It was easy to find the information that I needed.
13	The information provided for the system was easy to understand.	لقد كان من السهل فهم المعلومات التي يتم تزويدها عن طريق النظام	13	It was easy to understand the information which has been provided by the system.
14	The information was effective in helping me complete the tasks and scenarios.	المعلومات كانت فعالة في مساعدتي على إنهاء المهام وماهو مطلوب	14	The information was effective in helping me to finish the tasks and what was required.
15	The organization of information on the system screens was clear.	ترتيب المعلومات على شاشة النظام كان واضحا	15	The arrangement of the information was clear.
16	The interface of this system was pleasant.	واجهة النظام كانت مرضية	16	The interface of the system was satisfying.
17	I liked using the interface of this system.	لقد أحببت استخدام واجهة النظام	17	I liked using the interface of the system.
18	This system has all the functions and capabilities I expect it to have.	يحتوي هذا النظام على كل الوظائف والقدرات التي أتوقع أن أحتاجها	18	This system contains all the functions and capabilities that I was expecting to have.
19	Overall, I am satisfied with this system.	بشكل عام، أنا راض عن هذا النظام	19	Generally, I am satisfied with this system.

SUS				
1	I think that I would like to use this system frequently.	أعتقد أنني أرغب في استخدام هذا النظام بشكل متكرر	1	I think I would like to use this system frequently.
2	I found the system unnecessarily complex.	أجد هذا النظام معقدا بلا أي داع	2	I find this system complicated.
3	I thought the system was easy to use.	اعتقدت أن النظام سهل الاستخدام	3	I think this system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.	أعتقد أنني قد أحتاج إلى دعم شخص تقني لكي أستطيع استخدام هذا النظام	4	I think I might need the support for a technician so that I can use this system.
5	I found the various functions in this system were well integrated.	وجدت أن العديد من الوظائف المختلفة مدمجة بشكل جيد	5	I found many of the different functions are well integrated.
6	I thought there was too much inconsistency in this system.	توقعت أن هناك تناقضا وتنافرا في هذا النظام بشكل كبير	6	I thought there was inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.	أتخيل أن العديد من الناس سوف يتعلمون استخدام هذا النظام بشكل سريع	7	I presume that many people will quickly learn how to use this system.
8	I found the system very cumbersome to use.	وجدت أن استخدام هذا النظام مرهق للغاية	8	I found using this system is awkward/very exhausting.
9	I felt very confident using the system.	لقد شعرت بالثقة عند استخدام هذا النظام	9	I felt confidence when I used this system.
10	I needed to learn a lot of things before I could get going with this system	احتجت إلى تعلم العديد من الأشياء قبل الاستمرار باستخدام هذا النظام	10	I needed to learn many things before using this system.

PEU				
1	Learning how to use the site was easy.	تعلم استخدام هذا الموقع كان سهلاً	1	Learning how to use this website was easy.
2	Becoming skilful in using the site was easy.	أن تصبح ماهرا في استخدام هذا الموقع كان سهلاً	2	Being skilful in using this website was easy.
3	The site was easy to navigate.	كان من السهل التنقل في هذا الموقع	3	It was easy navigating this website.
PU				
1	Using this site improved my school performance.	استعمال هذا الموقع يحسن أدائي الدراسي	1	Using this website enhances my studying performance.
2	Using this site improves my productivity.	استعمال هذا الموقع يحسن إنتاجياتي	2	Using this website enhances my productivity.
3	Using this site improves my effectiveness in doing my school work.	استعمال هذا الموقع يحسن فعاليتي في أدائي الدراسي	3	Using this website enhances my school work effectively.
4	I find this site useful in doing my school work.	أجد هذا الموقع مفيداً في إنجاز أعمالي الدراسية	4	I find this website very useful in accomplishing my school work.
PE				
1	I find using the system to be enjoyable.	لقد استمتعت باستخدام النظام	1	I enjoyed using the system.
2	The actual process of using the system is pleasant.	طريقة استخدام النظام كانت ممتعة ولطيفة	2	The method of using the system was pleasant .and nice
3	I have fun using the system.	وجدت أن استخدام النظام كان ممتعا	3	I found that using the system was fun.

DIS					
1	I felt lost.	أحسست بالضياع	1	I felt lost.
2	I felt like going on without a destination.	أحسست كمن يدور بدون وجهة	2	I felt like moving with no destination.
3	It was difficult to go back to the page I had previously visited.	كان من الصعب العودة إلى الصفحة التي زرتها سابقاً	3	It was difficult to get back to the page I have visited earlier.
4	Navigation between pages was a problem.	التنقل بين الصفحات كان مشكلاً	4	Navigating between the pages was problematic.
5	I didn't know how to find the desired page.	لم اكن أعرف كيفية الوصول إلى الصفحة المبتغاة	5	I didn't know how to reach the desired page.
6	I was confused.	كنت مرتبكاً	6	I was confused.
7	After browsing for a while, I had no idea where to go.	بعد التصفح لفترة لم أكن أعرف أين أتجه	7	After browsing for a while, I didn't know where to go.

3.9. Conclusion

The primary objective of this chapter was to develop research hypotheses, review the literature and report the translation of all the questionnaires that are used in this research. The literature review shows a lack of Arabic versions of these questionnaires at the time of the start of this research. Therefore, this process is an initiative for cross-cultural studies in online psychometrics. In particular, the author focuses on the study of design parameters in online psychometrics, and for this purpose, an online psychometric questionnaire design tool needs to be developed. This tool must be able to accommodate both English and Arabic languages for Questionnaire administration. It must all be able to manipulate design parameters administer experiments and collect data. The system architecture and the development process of such a web-based tool are presented in the next chapter.

Chapter 4

System Architecture

Contents

Chapter 4: System architecture

4.1 Overview

With a plethora of tools for creating online surveys that are available at present, a question is why there is a need for another one. In contrast to other currently available survey software tools, Online Psychometric Questionnaire Design Tool (OnPQDT) is developed solely for research in online psychometrics, particularly testing design parameters. Hence only researchers have access to contribute, collect and manipulate the repository of data available within its databases. In this chapter, first, an outline of the features of OnPQDT is presented along with the available features of existing survey tools in the market. Next, the design of an extensible database system architecture is modelled. Further, the structure and the functional specifications are detailed. Finally, the user interface (dashboard) of the system the web interface, is detailed along with its functional specification.

4.2 Overview of the features of OnPQDT

OnPQDT is a web-based purposefully developed research tool for creating and running online psychometric experiments dealing with the manipulation of design parameters. The tool includes an online editor supporting the questionnaire creation, administration and data collection. As online survey tools are plentiful in the consumer market, the answer to the question of why is there a need for another tool lies in the following observed disadvantages such as,

(a) separation between content, style and design of questionnaires is not implemented in the commercial survey tools available in the market;

(b) currently existing tools (such as SurveyMonkey, Qualtrics, and QuestionPro, among many others) provide various templates and styling features (design parameters). However, these design parameters cannot be manipulated individually and have to be applied as a template to the questionnaire when created;

(c) the existing tools present a choice overload to users with an ever-increasing number of options to choose from. Research by Reutskaja and Hogarth (2009), report that there is reduced choice satisfaction when the complexity of the offered options increases. The choice overload is due to the increased cognitive effort needed to make a choice.

Due to the above mentioned limitations, the research facility especially to test the effect of design parameters in online psychometrics using the currently existing tools is not feasible. Specifically, for a large-scale research such as that reported in this thesis, there is a lack of functionality to support the manipulation of design parameters in the existing tools. As far as the author is aware, in the existing survey tools include there is (1) an inability to manipulate the size of response formats (such as the size of the radio button, length of the visual analog scale), (2) a lack of facility to store and download values to design parameters (e.g., font size, text/background colour), (3) a user interface with many features that are not required for online psychometrics such as skip-logic and multiple question types that are more suited for surveys, (4) a lack of an automatic move option to the next page when a single-item or multiple items are presented, (5) a limitation of features on free accounts such as only one active survey at any given time, limited number of responses, limited number of questions per survey and limited number of outgoing emails and (6) an involvement of cost to avail features such as unlimited responses and surveys compared to the cost of hosting the proposed web-based research tool (OnPQDT) that can inclusively facilitate the unlimited responses and surveys.

The focus of the development of this online tool was on the design of a reliable and extendable system architecture to support research experiments for online psychometrics. The visualisation of questionnaire items remains separated from the questionnaire content, and this separation allows researchers to efficiently create experiments to test the effect of design parameters (see Section 4.2.1) and presentation styles (see Section 4.2.2) in online psychometrics. At this point, it becomes necessary to explain the concepts of design parameter and presentation style.

4.2.1 Design parameter

In the context of online psychometrics, it is possible to define design parameters as visual elements that play an important role in displaying the questionnaire content such as questionnaire items and response type. The study by Van Schaik and Ling (2007) reported two experiments investigating three parameters (response format, questionnaire layout and interaction mechanism) of online questionnaire design. The study examined the effect of these parameters on the psychometric properties of the questionnaires. It was further reported in the study that for Likert scale a direct interaction mechanism with radio button should be used as it seemed to have some advantage in terms of completion time. The radio button can be designed with varying widths and manipulation of the width of the radio button is important especially for touch interfaces (see Section 3.3.2). In the same way, other design parameters of questionnaire items include font size, text/background colour, size of the response format (radio button), line spacing between answer choices and line spacing between questions and answers. Empirical research is required to investigate and report if these design parameters have an effect on the psychometric properties of the questionnaires. Figure 4.1 illustrates the different design parameters incorporated in OnPQDT. For the research reported in this thesis, design parameters such as the size of the radio button, font size, text/background colour and response format have been manipulated (see Sections 3.3.1 to 3.3.4).

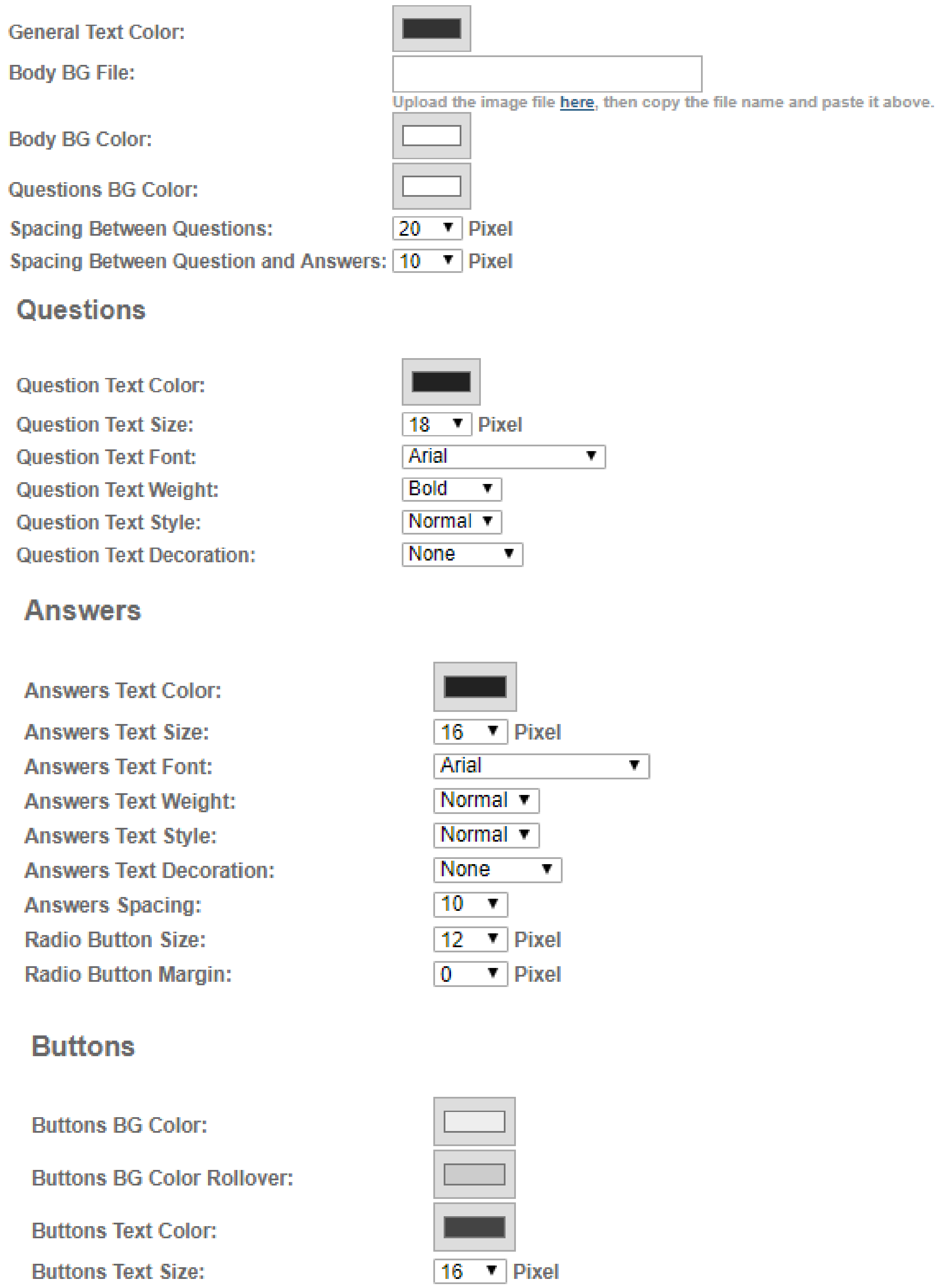


Figure 4.1. Design parameters incorporated in OnPQDT.

4.2.2 Presentation style

Presentation style in the context of online psychometrics can be defined as sequencing of items and the navigation. One such presentation style is questionnaire layout described in the research by Norman, Friedman, Norman, and Stevenson (2001). In the context of online psychometrics, research by van Schaik and Ling (2007) and van Schaik et al., (2015) found that single-item questionnaire layout was faster and had some advantage in terms of psychometric structure. In the same way, empirical research is required to investigate and report the effect of questionnaire layouts such as whole-form, single-item and semantic partitions or screen-sized pages and column layout. Associated to questionnaire layout, there are other useful online presentation styles such as an automatic jump to next page, lock the feature to go back and modify answers, lock the zoom-in and zoom-out facility for touch screen devices. The different presentation style options incorporated in OnPQDT are seen in Figure 4.2.

4.3 Overview of OnPQDT use

OnPQDT permits the creation of questionnaires and collection of user responses. Using this tool, authorised administrators/researchers can create and administer questionnaires without the involvement of specialists in web programming or web design. Once a questionnaire is created, the questionnaire can be published as a URL which can further be emailed to a group of respondents. The questionnaire can be accessed through the URL on any device that has access to the Internet through a web-enabled browser. The questionnaire responses are collected in a relational database with a timestamp as every participant completes the questionnaire. The tool also enables the export of the collected data in comma-separated values (CSV) format which can then be statistically analysed through other external data analysis software (e.g. SPSS Versions 23, 24).

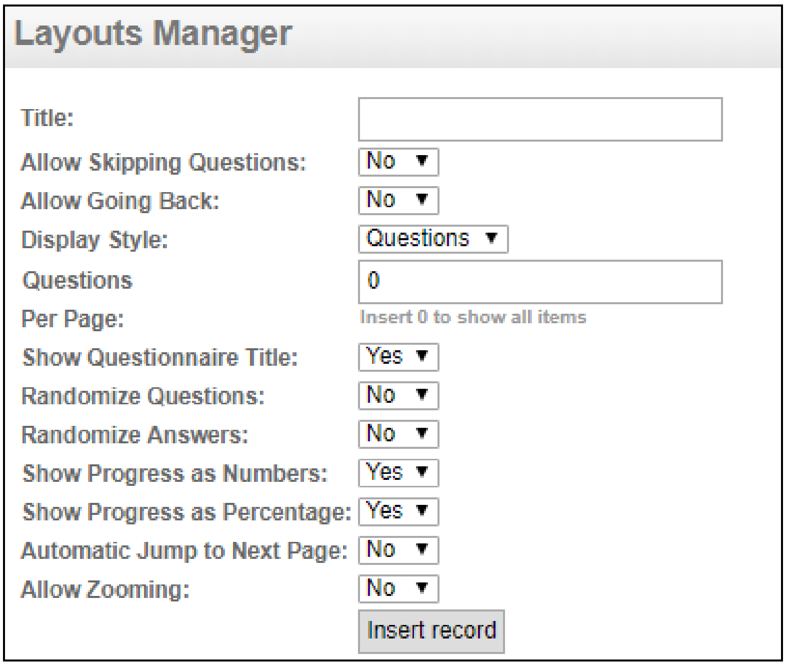


Figure 4.2. Presentation Styles incorporated in the OnPQDT.

With an array of different devices, OnPQDT supports administered questionnaires to fit different screen dimensions by allowing the questionnaire designer the ability to provide values to design parameters and presentation style using Responsive Web Design (RWD) (Marcotte, 2010). Due to the nature of the research project that is reported in this thesis, the design and implementation of the Online Psychometric Questionnaire Design Tool (OnPQDT) are primarily presented from a researcher's point of view. The following sections detail the software architecture, software design and implementation.

4.4 Architecture of OnPQDT

The software architecture is concerned with the selection of architectural elements, their interactions, and the constraints on those elements and their interactions necessary to provide a framework, to satisfy the requirements and serve as a basis for the design (Perry & Wolf, 1992). OnPQDT is a web-based tool built based on the Model-View-Controller (MVC) architecture and is a software development pattern that emphasises the separation of data representation code from methods that interact with data or process the data (Reenskaug, 1979, 2003). Reenskaug (1979) first formulated the MVC architecture for the SmallTalk programming language. MVC is a three-layered architecture and this enables the application to be more maintainable, easy to modify, update and enhance layers separately (Figure 4.3). The Controller is an interface between the View and the Model. It handles the user input and transfers the information to the Model. The Model consists of all the data of the application. It consists of classes that connect to the database. The Model updates its state and writes the data into the database. The data is fetched by the Controller and sent to the View. The View is the presentation layer and represents the user interface of the application. It receives the updated data from the Model through the Controller and responds accordingly. The user interacts with the View such as clicking on a link or submitting a form. This user interaction is again fetched by the Controller and is passed on to the Model. The solid lines represent the direct associations and the dashed lines are the inferred associations (Figure 4.3). As OnPQDT is a web-based application, the program resides on the server in three folders model, view and controller. The Table 4.1 details the technical specifications used for the development of OnPQDT.

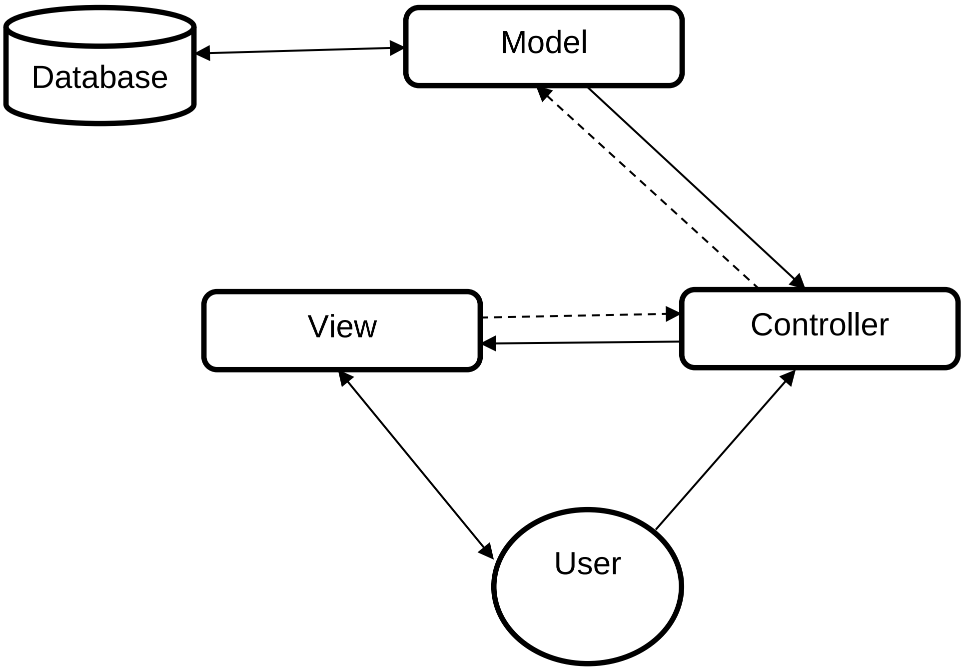


Figure 4.3. The MVC architecture

Table 4.1. Technical Specifications of OnPQDT.

Site info	*www.onlinepsychometrics.org/admin*
Server-side programming language	PHP 5.4.45 Model-View-Controller resides on the server
Client-side programming	HTML, CSS, JavaScript
Database	MySQL
Web server	Apache
Web hosting provider	GoDaddy

4.5 Design of OnPQDT

The design is concerned with the modularization and detailed interfaces of the design elements, their algorithms and procedures, and the data types needed to support the architecture and to satisfy the requirements (Perry & Wolf, 1992).

In order to guide the design of the development of the OnPQDT, design principles for tools to support creative thinking were chosen (Resnick et al., 2005; Shneiderman & Plaisant, 2010). Here the application of some of these principles is explained. A full treatment is beyond the scope of this thesis. The basic design decision to keep a simple user interface for the OnPQDT, is based on the first principle “Know thy user” mentioned in the study by Hansen (1971). The principle means that the interaction and interface of the system in design should cater to the needs and capabilities of the target user of the system. The users of OnPQDT will be researchers and hence simplicity will be the basis rather than a sophisticated user interface. Simplicity of OnPQDT refers to the plain user-interface format in terms of font colour, font type number of options to choose from with regards to question type. The background design of the web-page is also plain with no use of graphics. Other design principles such as understanding the task (identifying the sequence and structure of subtasks) (G. J. Kim, 2015), providing drop down lists for minimal memory load on the user (users should not be required to remember information) (Smith & Mosier, 1986) and offering informative feedback such as a preview or visual presentation of design parameters (see Section 4.5.2.3b) in the system (Shneiderman & Plaisant, 2010) were applied.

Figure 4.4 details the components of the Online Psychometrics Questionnaire Design Tool (OnPQDT). The administrator interacts with the web-based control panel (also called the dashboard) of the OnPQDT to create a questionnaire, administer it and retrieve the data collected in the online database of the OnPQDT. The respondents interact through an Internet-enabled browser with the online questionnaire on any given device.

Thus, the two main components are the database and the control panel of the OnPQDT as they are the core of the back and front-ends of this system, respectively. The respondent has direct access to the questionnaire through the URL provided by the administrator, and the web-enabled browser acts as the interface between the respondent and the administered questionnaire.

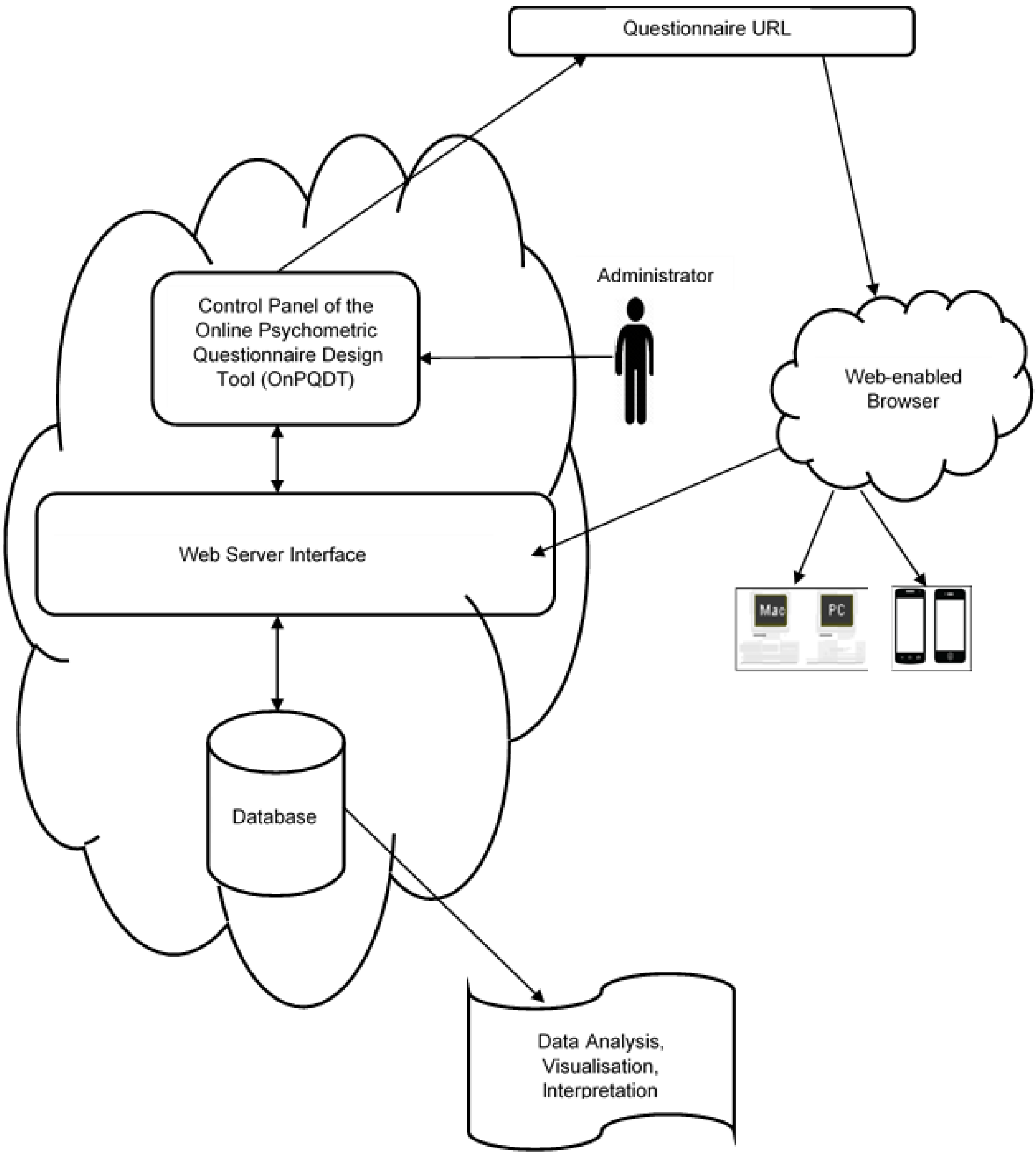


Figure 4.4. Components of the Online Psychometrics Questionnaire Design Tool.

4.5.1 Database Architecture

Data environments are comprised of data, hardware, software, people and procedures. A database is a structured set of data stored in a computer. The information age revolution has highlighted the role of the database management system (DBMS) as a key enabling technology. DBMSs are currently the technology of choice for modelling, storing, managing, and querying large amounts of information. A relational database is a set of tables from which data can be accessed in many different ways without having to reorganize the database tables. The standard user and application programming interface of a relational database is the Structured Query Language (SQL). Relational Database Management System (RDBMS) is the basis for SQL, and for all modern database systems such as MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access. MySQL has been used in this research and some of the fundamental characteristics of the database that are within the scope of this project are detailed in the following sub-sections.

4.5.1.1 Fundamental characteristics

(a) Control of data redundancy

Data redundancy is similar to storing one file in five different locations. However, with the help of a database, in this research study, information that is administered and data collected from the respondents will all be stored in one place, with easy access for the administrator.

(b) Data sharing

Current databases support a multi-user system that enables data sharing and collaboration among a group of people. In this research study, the database will allow many researchers to have access to the data of interest.

(c) Enforcement of integrity constraints

The database management system provides the ability to manage and store valid data. Among the many types of integrity constraints, data uniqueness is one of them, for example, the ability to generate and store a unique identification key associated with every respondent. Another instance of data uniqueness is the unique identification generated for each Questionnaire. Such uniqueness ensures no data is duplicated.

(d) Security and privacy

In a web-based environment, numerous individuals can have access to the database and it becomes important to protect the data and information from unauthorised access. Thus, authorised access for the concerned researchers is provided based on the profiles created in the database. Access privileges can also be controlled thus securing the data from unauthorised access.

(e) Backup and recovery facilities

Database management systems are equipped with backup and recovery facilities. Even in the case of computer system failures, the database system can recover and restore the database to its original state.

The database component of OnPQDT, in addition to facilitating the systematic collection and storage of data along with enhanced querying and report generation, further enables systematic development of the knowledge regarding the effects of design parameters on people's responses to the online psychometric questionnaires. The following section details the entity-relationship model that guides the implementation of the database.

4.5.1.2 ER Model

The Entity Relationship Diagram (ERD) organises the data into entities and defines the relationship between the entities. An ERD, first introduced by Chen (1976, 2002), is a major database modelling tool, often used by system analysts to produce proper database structures. An ERD has are three main components: entity, attribute and relationship. An entity is defined as an 'object' which can be distinctly defined (Chen, 1976). All database tables are entities. Attributes are the characteristics of these entities, and the relationship is an association among the entities. Figure 4.5 illustrates the Entity relationship diagram for the Online Psychometrics Questionnaire Design Tool (OnPQDT). There are a total of seventeen entities in the database component of OnPQDT. The entities are divided into five modules (1) Administration, (2) Content, (3) Design, (4) Structure and (5) Responses. These five modules are explained in the following sections.

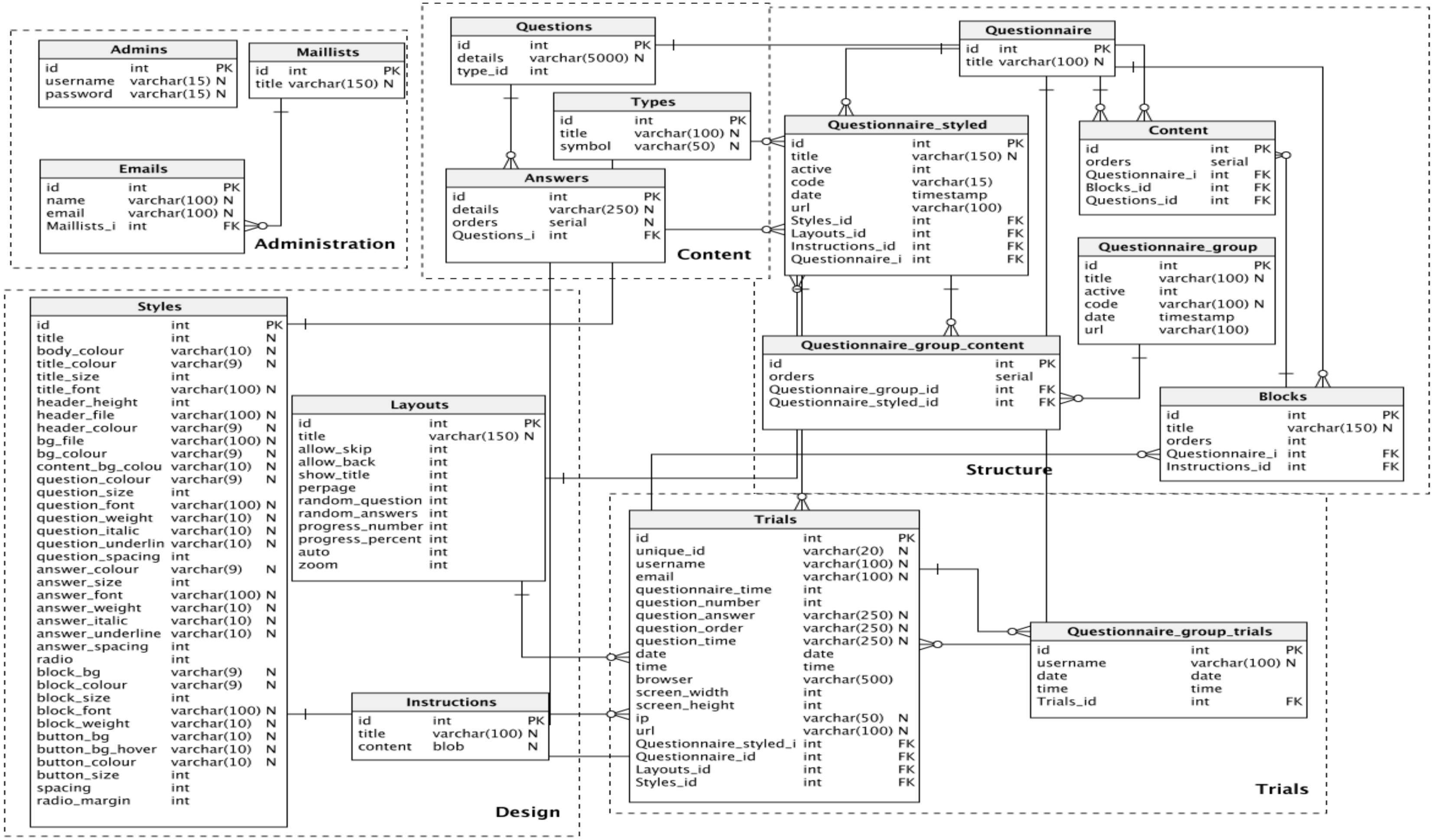


Figure 4.5. Entity relationship diagram of OnPQDT.

4.5.1.2.1 Administration module

Entry into the system is restricted to authorised access only. The author is currently the only administrator of the Online Psychometrics Questionnaire Design Tool (OnPQDT). The administrator can register other researchers as administrators, thus providing unique usernames and passwords for authorised access as required. The role of 'administrator' can be used interchangeably with 'questionnaire designer'. The administration module (Figure 4.6), consists of three entities: Admins, Emails, and Maillists. The credentials of authorised administrators such as username and password are stored in the Admins table. The Maillist table maintains the name of the mailing lists and the Emails table facilitates the storage of email addresses to whom the questionnaire must be emailed.

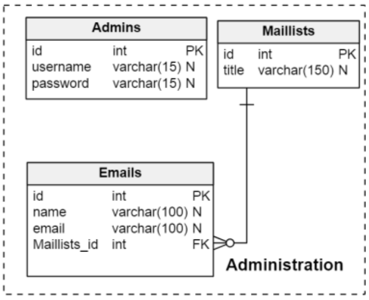


Figure 4.6. Entities in the OnPQDT's administration module.

4.5.1.2.2 Content module

The content module (Figure 4.7) is designed to contain the contents of psychometric tests such as the psychometric items, responses and the response options. As seen in Figure 4.7, three entities have been defined for this purpose: Questions, Answers and Types.

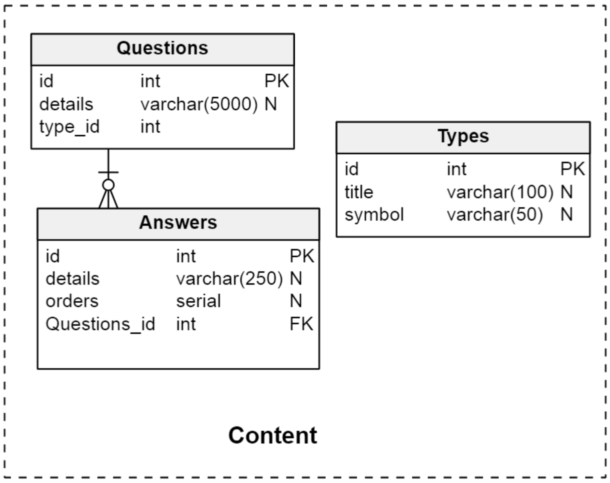


Figure 4.7. Entities in the OnPQDT's content module.

4.5.1.2.3 Design module

Sections 4.2.1 and 4.2.2 discuss the different design parameters and presentation styles within the scope of the study reported in this thesis. In the design module (Figure 4.8) three entities facilitate the storage of the design parameters (Figure 4.1), presentation styles (Figure 4.2), and instructions for each questionnaire. The entities are Styles (for design parameters), Layout (for presentation styles) and Instructions (for instructions). The design parameters and presentation layouts are chosen through the web interface by the questionnaire designer, stored in the corresponding tables and are later applied to the questionnaire at a later stage during questionnaire publishing.

The presentation styles (e.g., questionnaire layout) are stored in the Layout table and the design parameters (e.g., font size, text/background colour) are stored in the Styles table.

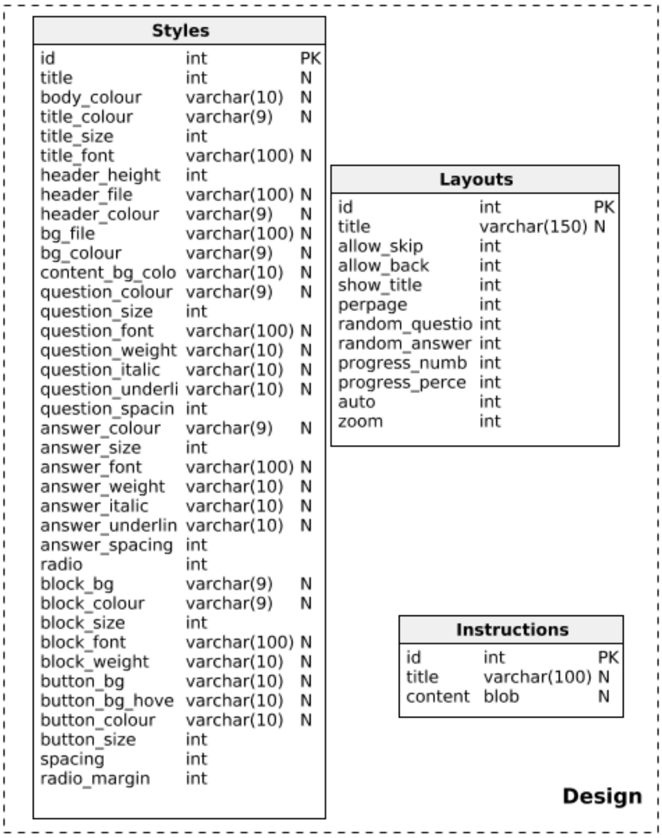


Figure 4.8. Entities in the OnPQDT's design module.

4.5.1.2.4 Structure module

The Structure module (Figure 4.9) details the entities required to generate and store questionnaires. In order to facilitate the generation of a questionnaire, two entities are used: Questionnaires and Blocks. The table Questionnaires, records the name of a psychometric construct. Question items already created can be grouped if required, and the Blocks entity facilitates the storage of such grouped question items if any.

A questionnaire is generated by importing the questionnaire items and is stored in the entity: Content. Other entities that belong to this module are Questionnaires_styled, Questionnaires_groups, and Questionnaires_group_content. The Questionnaires_styled entity is used to record the details of presentation styles (see Section 4.2.2 and Section 4.5.1.2.3) that are applied to the published questionnaire. It also contains the published URL link of the questionnaire. The design of an experiment may require the administration of more than one questionnaire. In such situations, to store the group of different questionnaires, two tables Questionnaires_groups and Questionnaires_group_content are used. The entity Questionnaires_group contains the published URL link for the questionnaire group along with the title of the questionnaire group. The entity Questionnaires_group_content stores the published questionnaires to be administered together.

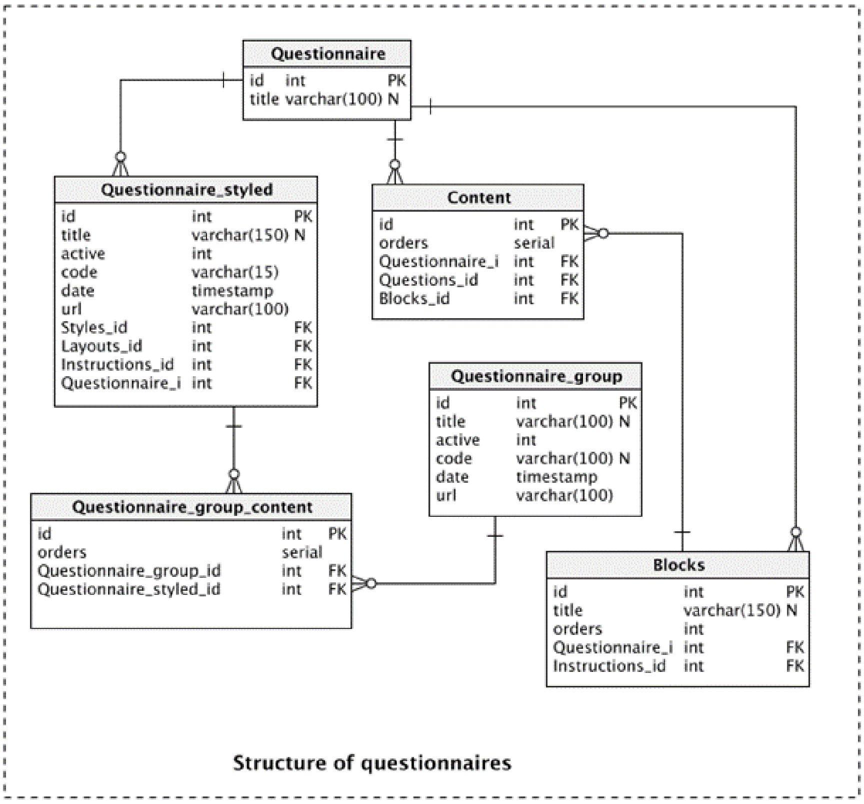


Figure 4.9. Entities of the OnPQDT's structure module.

4.5.1.2.5 Responses module

The responses of each respondent for the questionnaires is recorded. Two entities Trials and Questionnaires_group_trial (Figure 4.10) facilitate the storage of data. The term 'paradata' introduced by Groves & Couper (1998) into the survey research methodology field, refers to the additional data that can be captured during the process of administering a survey (Kreuter, 2013). The table Trials records all the responses, and paradata (e.g., screen size, IP address, date). When questionnaires are grouped, the details of the table Trials are stored in the Questionnaires_group_trial entity.

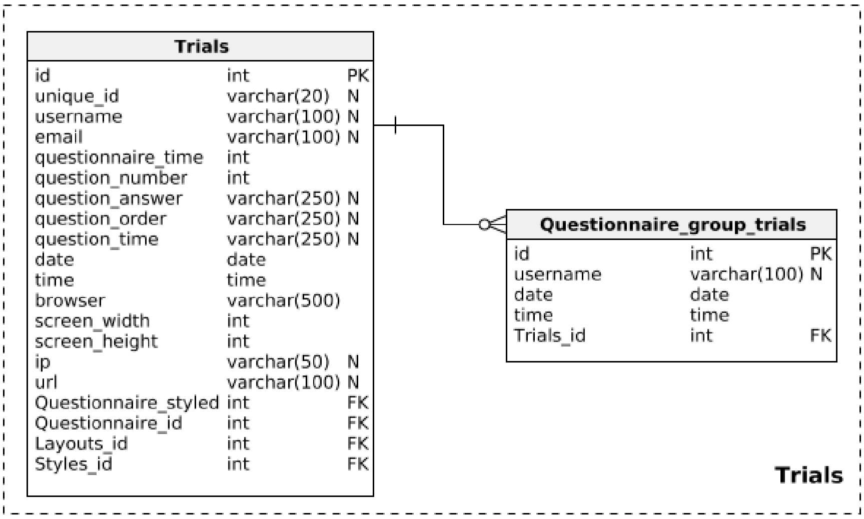


Figure 4.10. Responses Module.

4.5.2 The Control Panel of the OnPQDT

The questionnaire creation, administration, and presentation (publishing) process are achieved through the Control Panel of the Online Psychometrics Questionnaire Design Tool (OnPQDT). The control panel is interchangeably referred to as the dashboard or user-interface. The dashboard provides the facility to the administrator to create other authorised administrators to use the features of the OnPQDT for research purpose. Once the administrator has created and published one or more questionnaires, the URL can be distributed to various groups of respondents. The respondents access the questionnaire(s) through an Internet-enabled web browser on any device such as desktop/laptop or small-screen devices such as tablets/mobile phones. The authorised-administrator can access the control panel (Figure 4.11) by the use of the Internet address http://www.onlinepsychometrics.org/admin. A quick statistics is provided on the homepage of the control panel to present briefly the usage of the OnPQDT.

The Questionnaire count is the total number of basic questionnaires (without Styles and Layout applied) that have been created using the OnPQDT. The Formatted Questionnaires are questionnaires published with different Styles and Layouts. The control panel is divided into five modules: (1) Administration module, (2) Items module, (3) Design parameters module, (4) Questionnaire creation module and (5) Data module.

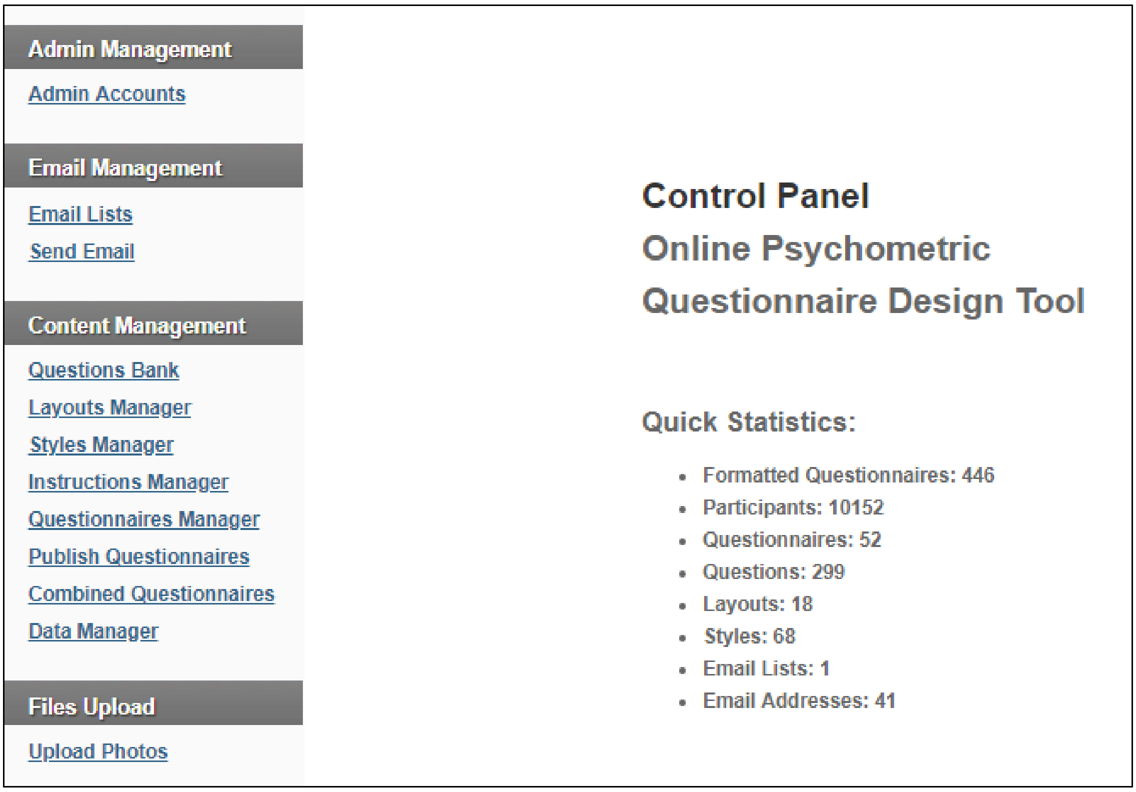


Figure 4.11. Control Panel.

4.5.2.1 Administration module

The administration module (Figures 4.12, 4.13 and 4.14) enables the administrator to (a) create further authorised administrators as required, (b) create mailing lists and (c) send emails from within the interface of the OnPQDT.

The send email option does not provide a facility for attachments. However, rich HTML-formatted emails could be sent with embedded pictures and URLs.

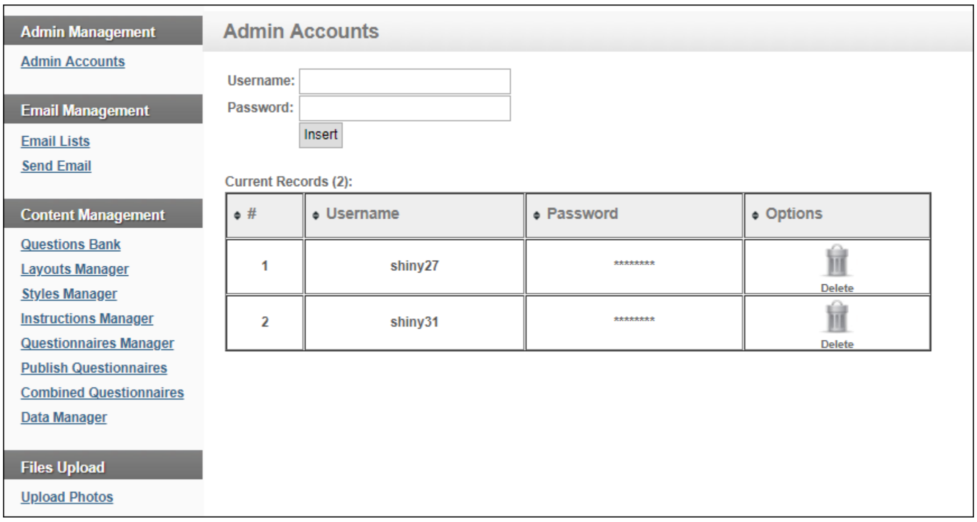


Figure 4.12. Create further Administrator.

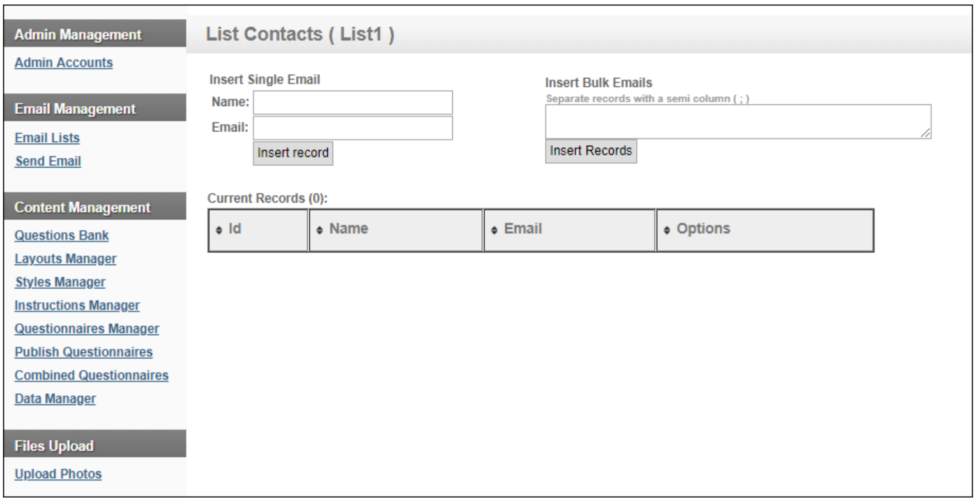


Figure 4.13. Create mailing list.

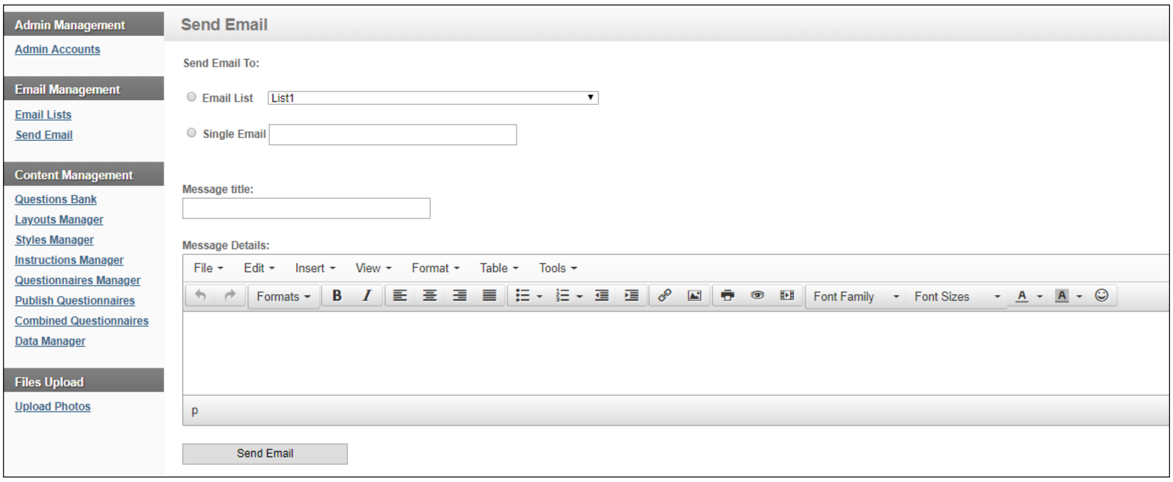


Figure 4.14. Send rich HTML-formatted email.

4.5.2.2 Items module

The items module handles the creation of question items (Figure 4.15). The Question Details contain the description (text) of the question item. The Answers Style is the response format (e.g., radio button, visual analogue scale). The Answers List contains the response options (e.g., strongly agree, strongly disagree). The question items that are created are later displayed in the user interface during the time of questionnaire creation for the questionnaire designer.

Figure 4.15. Creation of Question item (Items module).

4.5.2.3 Design parameters

Design parameters and presentation layouts are manipulated before the administration of questionnaires through (a) the Layout manager (Figure 4.16) and (b) the Styles manager (Figure 4.17, 4.18 and 4.19). The steps to create instructions that can be presented before the questionnaire at the time of administration to respondents is created through the (c) Instructions manager (Figure 4.20).

(a) The Layouts manager

The Layouts manager (Figure 4.16) provides the interface to handle two different questionnaire layouts (Whole-form, single-item) (Norman, 2001; van Schaik and Ling, 2007; Van Schaik et al., 2015). By providing a value '0' in the Questions per page feature a Whole-form layout (all questions on one page) is presented. A single-item layout presentation (one question item per page) is achieved by providing the value '1' indicating one question item per page. To save time in a single-item layout presentation, an automatic jump feature is programmed so the respondents can move to the next page automatically without clicking the 'Next' button. In the process of testing design parameters, considering the different types of devices that will be used by respondents, it is necessary to lock the zoom feature for touch interfaces, otherwise the size of design parameters (e.g., font size) may vary failing to produce the desired effect. Therefore, the facility to lock the 'zoom' feature is provided.

(b) The Styles manager

The Styles manager provides the interface to manipulate the design parameters of the questionnaire. In the Styles manager the design parameters are presented on one page. They are grouped into sub categories such as Title, Header, Questions, Answers and Buttons.

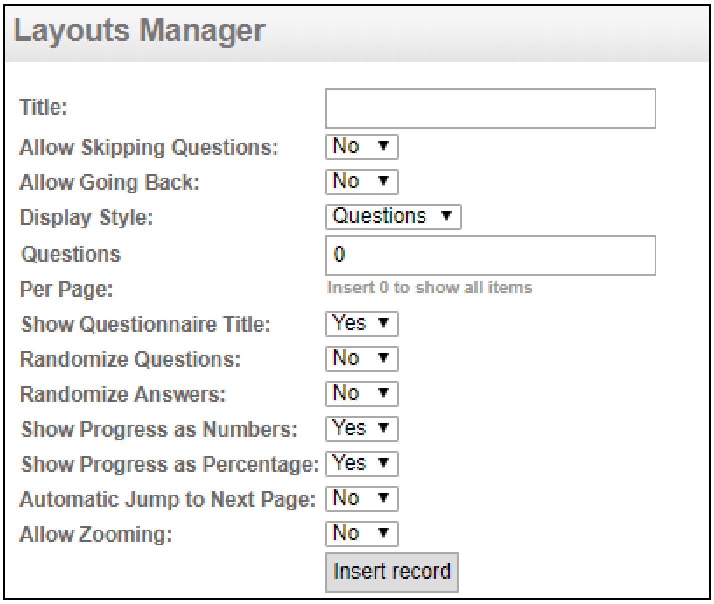


Figure 4.16. The Layouts Manager.

A visual presentation or preview is also provided on the same page of the Styles Manager to enable the Administrator receive a visual feedback based on the values provided for the design parameters (e.g., font size, font colour). Figure 17 presents the design parameters for the Title and the Header. As the name suggests, the Title applies for the title of the questionnaire. The default font type is Arial, colour is Grey and font size is 38 pixel. These values can be changed as required. The width of the Header is determined by providing a value (in pixels) for Header Height. By default, value 100 pixels is already provided. An image can optionally be provided for the Header. In situations where the width of the Header banner image is smaller than the Header height, a background colour can be provided. By default background colour white is provided for the Header background.

Figure 4.17. Styles Manager (Title and Header).

Figure 4.18 presents the design parameters for the Questions and Answers in the Styles manager. The text/background colour, font type and size can be manipulated. The weight provides one option of 'bold'. The style provides one option of 'italic'. The decoration provides one option of 'underline'. Apart from these parameters, the line spacing between the answer choices can be changed if required. The size of the response format (radio button) is manipulated through the feature Radio Button Size. The Radio Button Margin indicates the margin width from the edge of the screen (for small-screen devices).

Figure 4.18 presents a further set of design parameters for Buttons (Submit and Next). The Submit button will be presented at the end of every questionnaire that is administered. The Submit button enables the data to be written into the database. However, the Next button will be visible only if the Automatic Jump option (Layouts Manager) is turned off (see Section 4.4.2.3). Both the buttons are programmed to have the same style. The design parameters such as font size and text/background colour can be manipulated. Apart from this, an option of providing a colour change when the mouse pointer rolls over is also provided.

(c) The Instructions manager

The instructions manager (Figure 4.20) provides the interface to upload different types of instructions for participants before the presentation of a questionnaire. The types of instructions include text, image and video. For example, text-based instructions were uploaded before the workload questionnaire in Study 2, to help respondents understand how to answer the questions.

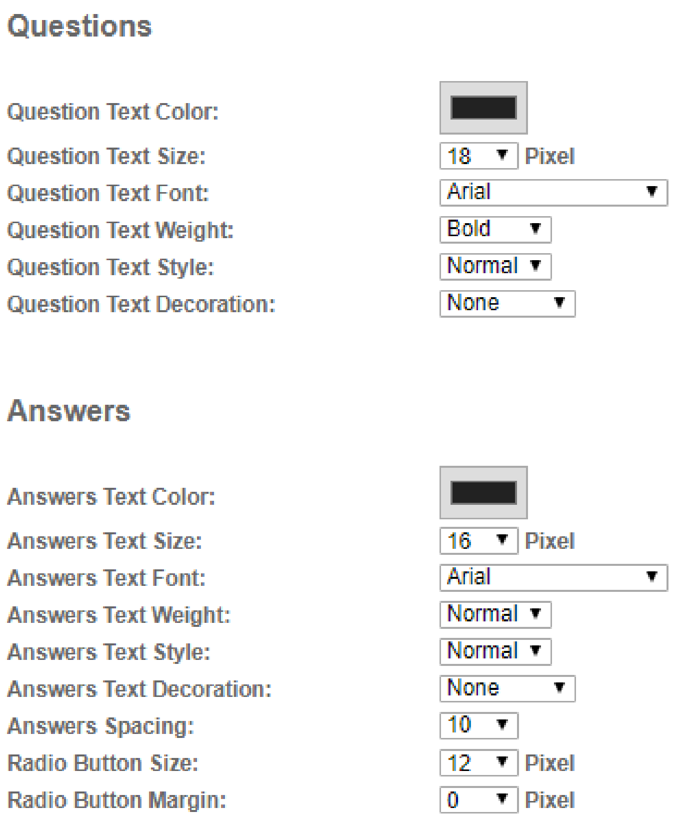


Figure 4.18. Styles Manager (Questions and Answers).

Figure 4.19. Styles Manager (Buttons).

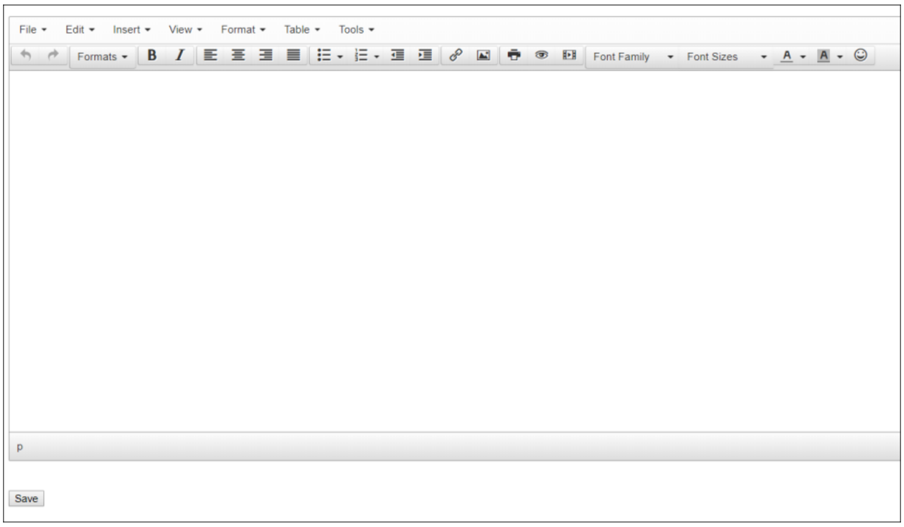


Figure 4.20. Instructions Manager.

4.5.2.4 Questionnaire creation module

The first step to create a questionnaire is to import the questions that have been created (see Section 4.5.2.2). The import feature is available for the administrator in the Questionnaires manager (Figure 4.21).

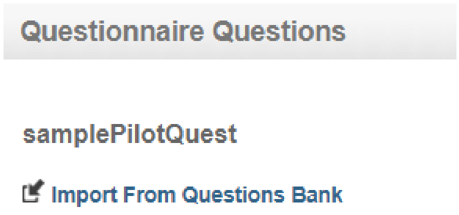


Figure 4.21. Import questions into the questionnaire.

The second step is to publish the questionnaire and apply the design parameters that the administrator/questionnaire designer has chosen (Figure 4.22). The Title identifies the published questionnaire and is different from the name of the questionnaire. The field Questionnaire as seen in (Figure 4.22) is the name of the questionnaire that contains the imported set of questionnaire items. The fields Style and Layout identify the chosen design parameters and presentation layouts (see Section 4.5.2.3). Instructions (if any) (see Section 4.5.2.4) that has been created is chosen to be published as part of the questionnaire.

The Active feature is by default set to the ‘Yes’ option and can be turned off to a ‘No’. Once the option is set to a ‘No’, the questionnaire is deactivated and a respondent who has the URL will not be able to access the questionnaire. For additional security purposes, an Access Code feature is provided. If left blank, a respondent with an Active URL can directly access the questionnaire. However, if a text or numeric code is provided for the ‘Access code’, the respondent with the active URL can access the questionnaire by providing the same access code. Thus, it is recommended that administrators provide an Access Code. In addition, it is the responsibility of the administrator, to provide the set Access code to the respondents so they can gain access to interact with the questionnaire. The URL through which the questionnaire can be accessed is provided as the URL Title.

Care should be taken to avoid spaces and the same instruction is also provided in the interface as a precaution to the administrator/questionnaire designer (Figure 4.22). The URL Title allows the inclusion of all the characters including numbers except spaces.

Once a questionnaire is published, it can be viewed in the user interface along with all the details such as Style, Layout, Access code and URL (see Figure 4.23). The hyperlink of the URL can be clicked to access the questionnaire and alternatively, can be emailed to respondents. When more than one questionnaire needs to be administered, a 'combined questionnaire group' is created. More than one 'published' questionnaire (Figure 4.23) is added to the 'questionnaire group' in the order it needs to be administered (Figure 4.24). The provided URL Title will be published as the active link for the questionnaire group (Figure 4.25). By default, the URL link is set to be Active and an optional Access code can be provided for additional security.

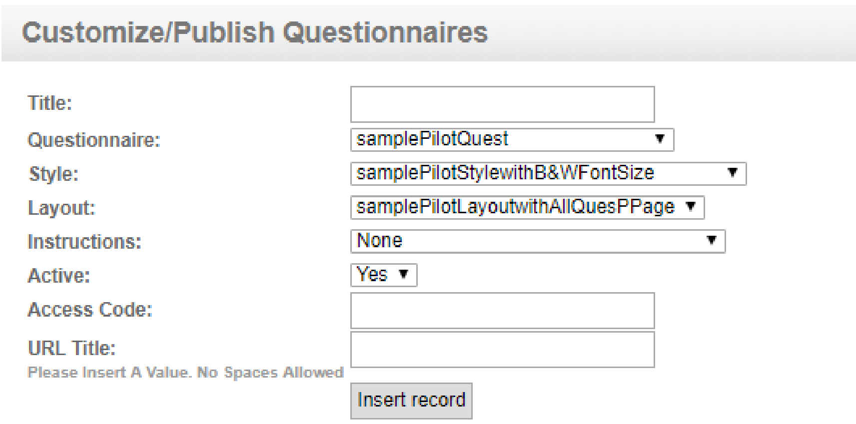


Figure 4.22. Publish Questionnaire.

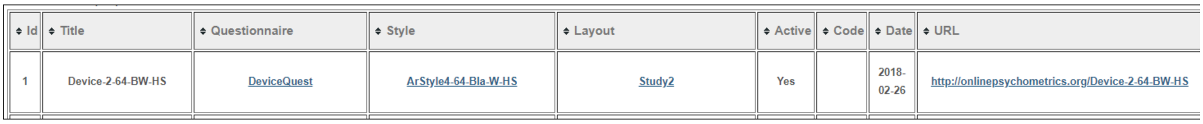


Figure 4.23. Questionnaire URL.

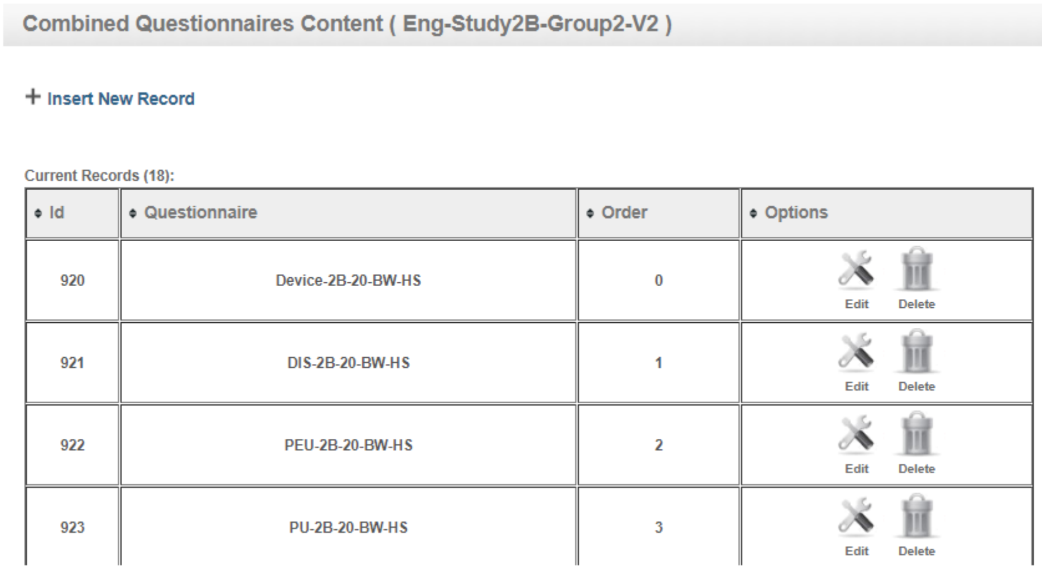


Figure 4.24. Questionnaire group.

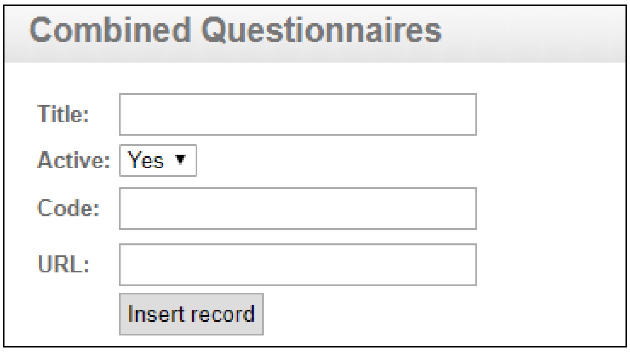


Figure 4.25. URL of the Questionnaire group.

4.5.2.5 Data module

The administrator for the Online Psychometric Questionnaire Design Tool (OnPQDT) has access to all the data stored in the database, through the Data Manager (Figure 4.26). The data are downloaded in the comma-separated value (CSV) format.

The responses of each respondent are stored in the database for each questionnaire, along with the IP address, date, local time, screen resolution (pixels), the total completion time for each questionnaire, and the name of the browser used by each respondent. Besides, the administrator can also request for the details of the manipulated design parameters and presentation layouts (Figure 4.26) of the particular questionnaire (e.g., font size, text/background colour, size of the response format – radio button). The quantitative analysis of the downloaded data can be accomplished by using external software programs such as IBM SPSS by the administrator.

Figure 4.26. Data Manager (Design parameters and Presentation Layout options).

4.6 Technical Evaluation of OnPQDT

Any technical innovation can be evaluated from various perspectives. Here the innovative tool OnPQDT is evaluated from the perspective of usability. This is because the future development of the tool may involve making this suitable for multiple users who may not be specialists in tool development.

Among the several tools and checklists available to evaluate technical innovation in terms of usability, a heuristics evaluation is applied (Nielsen, 1993). Usability evaluations often consist of usability experts evaluating an interactive system so that they may gauge whether any violations or usability issues remain in the current design. However, in this research, the author was the developer and the sole user of the OnPQDT. Hence the tool has been evaluated only by the author. In addition, the author also completed a free usability evaluation provided in the website https://www.uruit.com/ux-quiz/home. The summary of this quiz based evaluation is provided in Table 4.2 and the full results are provided in the appendix as reference. The results show high compliance with the principles *Match between the system and real world* (2), *Consistency and standards* (4) and *Aesthetic and minimalist design* (8). However, compliance was poor with the principles and *User control and freedom* (3) and *Flexibility and efficiency of use* (7). In addition, there was a complete lack of compliance with the principle *Help and documentation*.

Table 4.2. Summary of the results of the heuristics evaluation of OnPQDT with the online quiz.

How usable is OnPQDT	Score
1. Visibility of system status	50%
2. Match between the system and real world	100%
3. User control and freedom	30%
4. Consistency and standards	100%
5. Error prevention	50%
6. Recognition rather than recall	80%
7. Flexibility and efficiency of use	25%
8. Aesthetic and minimalist design	100%
9. Help users recognise, diagnose and recover from errors	60%
10. Help and documentation	0%

1. Visibility of system status: the users of the system must always know where they are, what are they doing and what is the result of the action they took, through appropriate feedback within reasonable time. For example, when a questionnaire is accessed through a URL, it is appropriate that a system message is displayed such as “waiting for the page to load”. In the implementation of OnPQDT, the author did not observe the need of displaying a status while the processing is being done, because the system handled mostly text-based instructions and displays with a short response time. In one instance where data is downloaded, a popup message appears informing the user on the system status through appropriate feedback (e.g., see Figure 4.27).

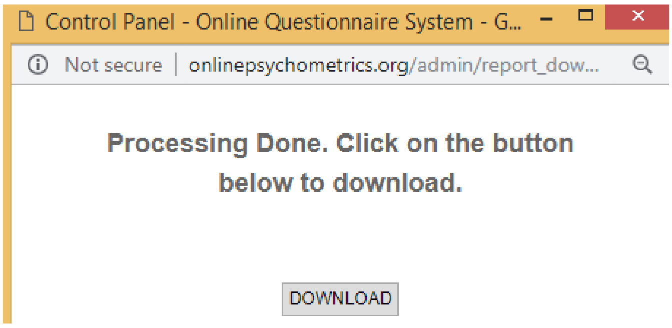


Figure 4.27. Feedback for data download through OnPQDT.

2. Match between system and the real world: the system should speak the way users do. The use of familiar words, concepts and phrases make the adoption easy. This holds true not just for the use of language but also in the visual elements like typography, colour, and icons. As in real life, the colour red with an exclamation mark is instinctively associated with the alert. Similarly, green is often associated with the success of an action. However, in the implementation of the OnPQDT, the interface has been kept extremely simple; nevertheless, real-world conventions, have been implemented to make information appear in a natural and logical order (e.g., see Figure 4.28).

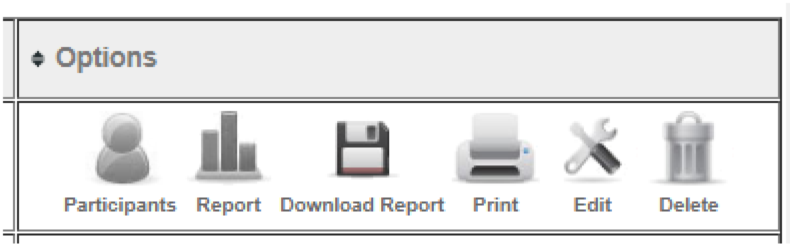


Figure 4.28. Appropriate icons to represent options for published questionnaires.

3. User control and freedom: users must be provided with a clearly marked “emergency exit” to leave an unwanted state without having to go through an extended dialogue.

This gives the user increased confidence to use the system. In OnPQDT, the freedom to cancel any accidental action is provided (e.g. see Figure 4.29). Moreover, “Home”- and “Logout” options consistently appears as the header in every page the user navigates to. The “Home” option returns the user to the homepage and the “Logout” option ends the user’s current session.

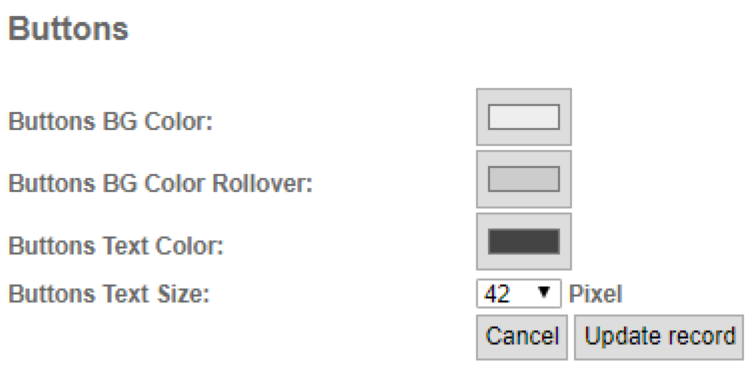


Figure 4.29. Option to cancel any accidental changes.

4. Consistency and Standards: things like the placement of logo on the upper left corner, logout option on the right have almost become standard conventions. This has been consistently followed with the “Logout” option placed on the upper right corner of every page the user navigates to. Likewise, this principle requires consistency to be maintained across different pages of the same website. For example, the submit button on one page should be consistent although it may appear in another page also. This principle has been applied consistently in OnPQDT (e.g., see Figure 4.30).

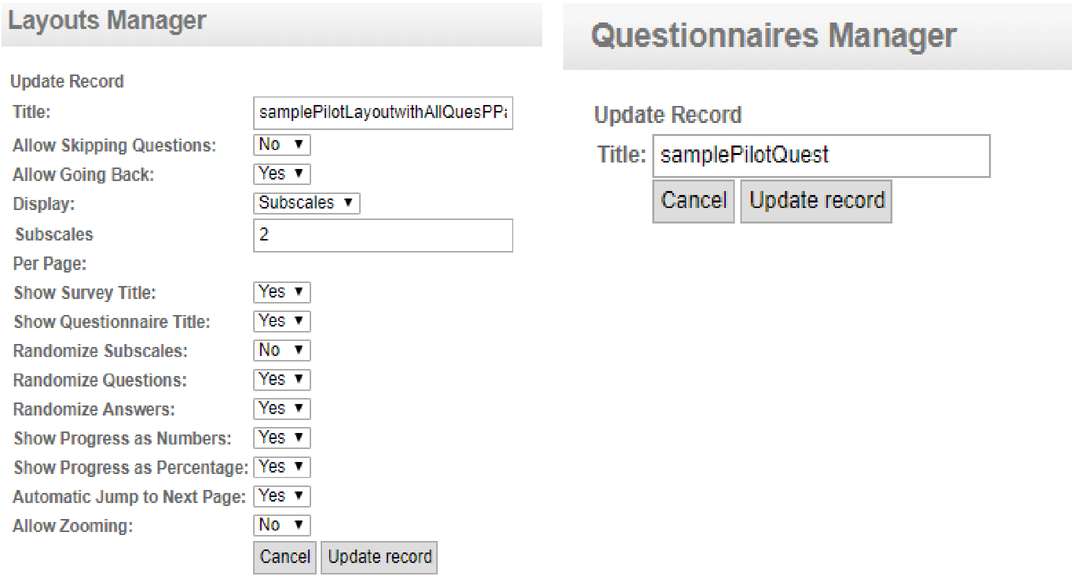


Figure 4.30. The Cancel and Update record buttons on two different pages of the OnPQDT.

5. Error Prevention: a good user-interface should warn the user even before a mistake is about to be committed. In OnPQDT, it is important that the administrator follows the pattern of using text with no spaces, when choosing a URL title. Accordingly, an error prevention message has been provided (e.g., see Figure 4.31).

Figure 4.31. Error prevention message to avoid spaces in the URL title.

6. Recognition rather than recall: recall requires cognitive effort while recognition is retrieving information with the help of multiple options. Usually, recognition makes remembering easy. From the perspective of user experience, all the options that are frequently used or required, must be shown and not hidden.

In addition, the goal of this principle is to minimise user's memory load. In OnPQDT, the administrator has a preview of the design parameter manipulations and therefore, recognition has been applied (e.g., see Figure 4.32).

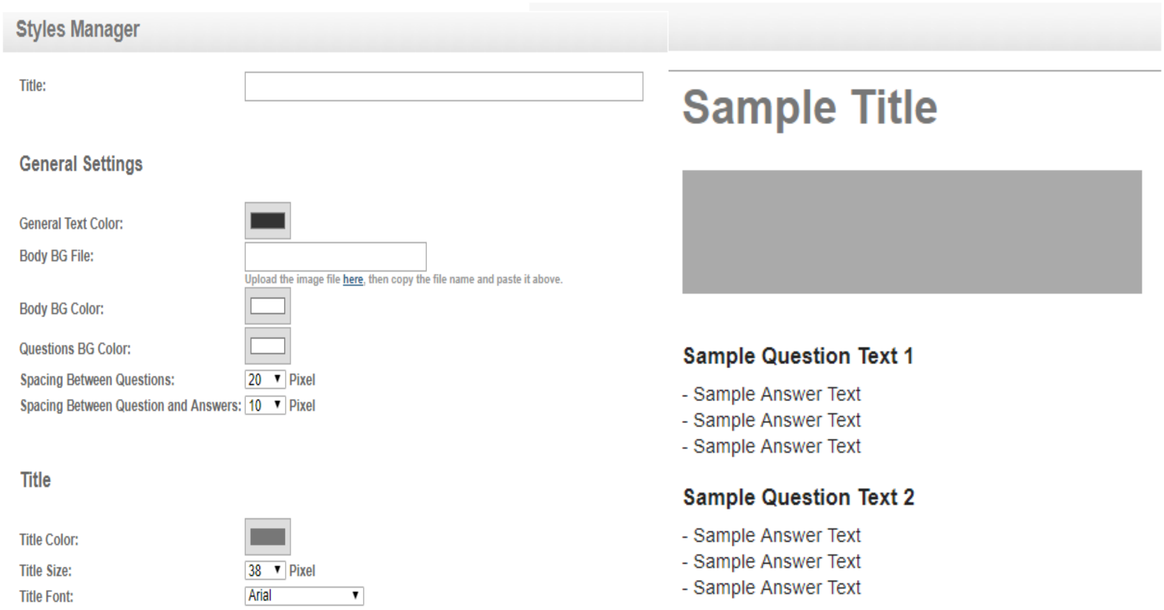


Figure 4.32. Manipulation of design parameters and a preview of the chose values.

7. Flexibility and efficiency of use: the system must be able to accommodate both beginners and experts. The interface of OnPQDT is flexible. However, there are no features that are different for novice and advanced users because currently the system only supports one user. Thus, this principle is not completely applicable for OnPQDT.

8. Aesthetic and minimalist design: minimalist design is a strategy where only the features and content that are required will be shown. More information also creates more cognitive load and more decision time.

The interface of OnPQDT avoids unnecessary designs and colours. The aim was to implement an elegant and simple interface, as it has been solely developed for research purposes. Thus the OnPQDT offers a clutter-free experience to the user.

9. Help users recognise, diagnose, and recover from errors: mistakes can happen. Thus, simple error messages can help the user rectify the action. In the OnPQDT, errors made are explained to the user to a certain extent in a simple and understandable language. For example, it is important to inform the user if the username or password is invalid (see Figure 4.33).

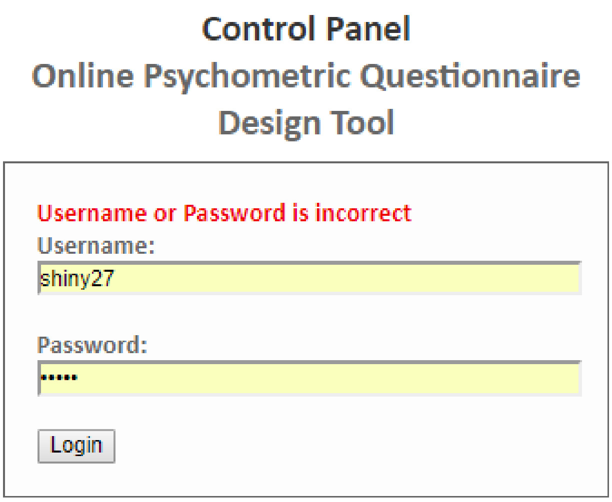


Figure 4.33. Help user recognise and recover from errors.

10. Help and documentation: although the interface may be simple and easy to use, if the user gets confused somewhere, help and documentation must be available to support them. Currently, there is no help and documentation provided as the author was the sole developer and user of the OnPQDT.

A brief evaluation of the heuristics evaluation quiz against the principles is presented in Table 4.2. *The* principle *Visibility of the system status* (1) is half compliant because, the few error messages such as invalid design parameter choices or layout names are consistently shown in the same area of the page. Also, the header of the page is visible to the user at all times and thus the researcher is aware of his current standing. *Error prevention* (5) too is only half compliant because, the system does not provide data entry hints to the user to avoid mistakes. In the few situations where the users are guided to error prevention, the error messages provided are in clear simple language that guide the user to recovery process. Thus, the quiz reveals that the principle *Help users recognise, diagnose and recover from errors* (9) is 60% satisfied. In most places, the user is guided well on the required choices to be made and therefore the principle *Recognition rather than recall* (6) is seen to be 80% compliant. On the other hand, the principle *Flexibility and efficiency of use* (7) is only 25% compliant because, there is no search enabled within the system and neither are there any shortcut keys for the user to perform a task. This being said, the principle *User control and freedom* (3) is limited mainly due to the fact that operations once performed cannot be rolled back or cancelled while in progress. However, clear indications on the button that perform permanent tasks are clearly labelled. It is evident from the quiz that this principle is only 30% compliant. The principles, *Match between the system and real world (2)*, *Consistency and standards* (4) and *Aesthetic and minimalist design* (8) are 100% compliant and have been fully satisfied with regards to the interface design of OnPQDT. However, the principle *Help and documentation* (10) received a score of 0% and will be an area to focus in the future enhancements to the system.

4.7 Limitations

Unlike the commercially available survey generations tools, this online tool has been developed solely for research purposes. Therefore, the user interface has been kept simple. Currently, the administrator can create multiple administrators. In principle, such multiple administrators of a system will have individual memory workspace allocated with access privilege rights to the database. However, currently, although multiple administrators could be registered, no individual memory workspace will be allocated and the administrators share the same settings, questionnaires, and system permissions. In addition, it is to be noted that only the size of the radio button response format can be manipulated currently. The future recommendations will include manipulating the size/length of the visual analogue scale response format and allocating memory workspaces along with access privilege rights to the database for administrators.

4.8 Conclusion

Thus the development of the Online Psychometric Questionnaire Design Tool (OnPQDT) is an important completion to achieve the overall goal of testing design parameters for online psychometrics. Several pilot experiments with data collection were conducted to fix minor issues. In this study, three major experiments have been conducted, and data have been collected for analysis. These experiments are reported in Chapters 5, 6 and 7.

Chapter 5

Study 1

Contents

Chapter 5: study 1

5.1 Overview

This chapter presents a first experimental study of design parameters in online psychometrics (Study 1). The main aim of the experiment was to test the effect of design parameters such as font size, response target size and text/background colour for online psychometrics. The design parameters during the presentation of the questionnaires were manipulated to analyse the effect on the time taken to complete the questionnaires. Five psychometric questionnaires were administered online to test the effect of design parameters. The experiment was conducted on mobile phones among a private university students based in Kuwait. Participants responded to psychometric questionnaires based on their experience with the university's virtual learning environment (the student information system PeopleSoft). They then rated the quality of their experience with online psychometric questionnaires in this session. The chapter begins with the statement of the hypothesis, followed by the description of the design and method of the experiment. An account of the data analysis follows and finally the results and discussion of the study are presented.

5.2 Statement of hypotheses

As described in Section 3.3, the following Hypotheses are presented for Study 1.

Hypothesis 1: completion time decreases with increasing font size;

Hypothesis 2: completion time decreases with larger response target size;

Hypothesis 3a: completion time decreases with increasing colour contrast;

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 9: perceived enjoyment increases with increasing response target size;

Hypothesis 10a: perceived enjoyment increases with increasing colour contrast;

5.3. Method

5.3.1 Design

The experiment used a 2×2×(3) experimental design with three measures. The first, independent-measures, variable was font size with two levels: font size 36 point and font size 44 point (see Section 3.3.1). The second, independent-measures, variable was text/background colour with two levels: blue on white and black on white (see Section 3.3.3). The third, repeated-measures, variable was response target size with three levels: large: 12.7 mm, medium: 15.5 mm and small: 22.6 mm (see Section 3.3.2). Therefore, a total of 12 (= 2×2×3) questionnaire versions were created. The experiment was organised into four groups: according the font size, text/background colour and response target size. The details are as follows:

Group 1: 36 point, black on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 1: 36 point, black on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 2: 36 point, black on white, 22.6 mm/12.7mm/15.5 mm

Order/version 3: 36 point, black on white, 15.5 mm/22.6 mm/12.7

Group 2: 36 point, blue on white, 12.7 mm/15.5mm/22.6mm

Order/version 4: 36 point, blue on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 5: 36 point, blue on white, 22.6 mm/12.7 mm/15.5 mm

Order/version 6: 36 point, blue on white, 15.5 mm/22.6 mm/12.7mm

Group 3: 44 point, black on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 7: 44 point, black on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 8: 44 point, black on white, 22.6 mm/12.7 mm/15.5 mm

Order/version 9: 44 point, black on white, 15.5 mm/22.6 mm/12.7 mm

Group 4: 44 point, blue on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 10: 44 point, blue on white, 12.7 mm/15.5 mm/22.6 mm

Order/version 11: 44 point, blue on white, 22.6 mm/12.7 mm/15.5 mm

Order/version 12: 44 point, blue on white, 15.5 mm/22.6 mm/12.7 mm

This experiment was a retrospective assessment of the student information system. It is to be noted that it was the university's policy that class schedules, attendance system and course grades are maintained on the PeopleSoft website. It seems likely that the last time the students used the system will have been one week previously at most. Participants were not required to use the system within this experiment. The participants in each group responded to five psychometric questionnaires: the disorientation (DIS) developed by Ahuja and Webster (2001), Davis's (1989) perceived ease of use (PEU) and perceived usefulness (PU) scales, and the usability questionnaires PSSUQ, developed by Lewis (1995, 2002), and SUS developed by Brooke (1996). A brief introduction to these questionnaires is provided in Section 3.4. Participants responded to these psychometric questionnaires in a fixed order (PSSUQ, SUS, DIS, PEU, PU) by rating their experience with the university's Student Information System (PeopleSoft) website. Immediately after the participants completed their responses to the five psychometric questionnaires, they completed the perceived enjoyment (PE) scale developed by Davis (1992); specifically, they rated their experience of responding to the psychometric questionnaires they just completed. The experiment was repeated three times for the three different response target sizes for each participant in every group. The dependent variables were the time to complete the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS) and perceived enjoyment.

5.3.2 Participants

A total of one hundred seventy-one university students from Kuwait (51 Male and 120 Female) took part. All the participants were native Arabic speakers and were enrolled for courses 231-Readings in Politics of Kuwait and 101-Reading in Arabic Literature. They were familiar with the Student Information System (PeopleSoft) website, as they frequently accessed it to check assessment grades, register for classes each semester, monitor their attendance (in case of discrepancy from the automated attendance system), check for assessment dates and class schedules.

5.3.3 Materials and equipment

The experiment was administered through the Online Psychometric Questionnaire Design tool (OnPQDT) developed for this research purpose (see Chapter 4).

All participants took part in the study using their own mobile devices. Of the participants, 2% used the Samsung galaxy S6 while 37% used the iPhone 6 model and 61% used the iPhone 6plus (see Table 5.1). Therefore, most of the participants had iOS devices (iPhone 6, iPhone 6plus), while the rest used the Android device (Samsung Galaxy S6) (see Figure 5.1). The dimensions of the devices used by the participants are detailed in Table 5.1.

Figure 5.1. Devices used by participants: iPhone 6, Samsung Galaxy S6, iPhone 6plus.

Table 5.1. Dimensions of mobile devices used in the experiment.

	Numbers of each device type	Screen Size (inches)	Resolution (pixels)	Pixels per inch (ppi)	Dimension (mm)
iPhone 6	64	4.7	750×1334	326	138.10×67.00×6.90
iPhone 6 plus	104	5.5	1920×1080	401	158.10×77.80×7.10
Samsung-Galaxy S6	3	5.1	25600×1440	577	143.40×70.50×6.80

5.3.4 Procedure

The study took place in July 2016. Participants took part individually. There were a total of eight sessions. Each session took place in the classroom where the author was present, along with the course leader and lasted for an average of forty-five minutes each. Every session was led by the author of this thesis. The participants were awarded extra course credits as incentives for participation in the experiment.

Research ethics approval was granted by Teesside University's Research Ethics Committee. The author incorporated the informed consent form electronically within the experiment. The participants were also verbally briefed by the author before the beginning of every session on the reasons for their participation in this research and what the research required of them. Although incentives were offered by the course leader, the students were informed that participation was voluntary and they had the right to withdraw at any point of time within the experiment. In addition, the principle of maintaining confidentiality with the collected data was assured to the participants. For data analysis purposes a unique participant ID was generated for every participant internally by the OnPQDT.

Consecutive sessions of experiments were organized. Every participant in each session received a URL link from the author. The Access code (see Section 4.5.2.4) was also provided to the participants; this was necessary to ensure that the URL is being accessed only by the participants within the session. The URL link when accessed, first provided the participants with the online consent form. Thereafter, brief online instructions (see Section 4.5.2.3) in Arabic were provided on how to complete the five psychometric questionnaires. At the end of the responses to the five psychometric questionnaires, another brief instruction page was administered online on how to complete the perceived enjoyment scale based on their experience of responding to the psychometric questionnaires. The experiment automatically repeated three times according to the three response target sizes (see Section 5.3.1). The font size and the text/background colour remained uniform according to the groups assigned for the sessions. At the end of the experiment, a ‘Thank you’ page was presented to every participant in appreciation for their time and involvement. The author, along with module leader, were present till all the students completed the experiment. Figure 5.2 shows a picture of an experimental session in progress captured by the author.

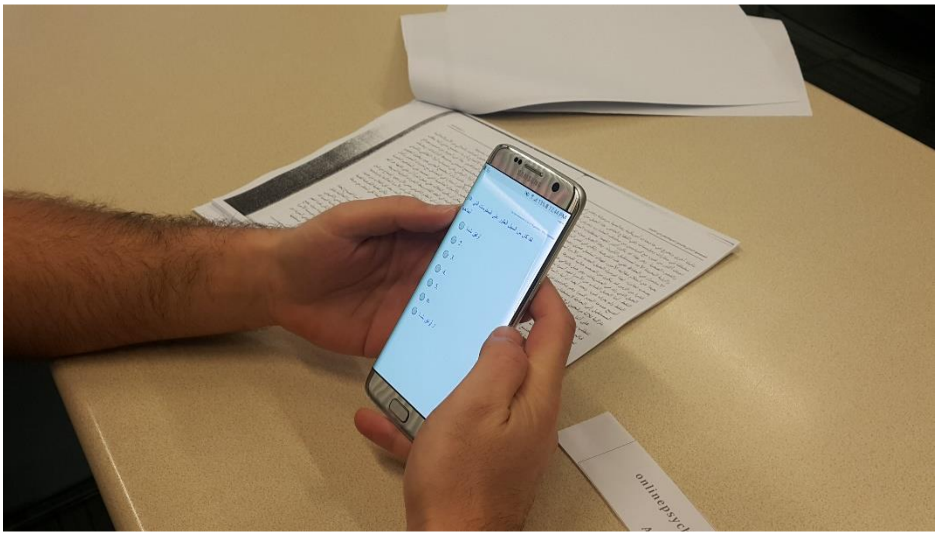


Figure 5.2. Images from the experiment (Study 1).

5.4. Analysis

The data collected for Study 1 was analysed using SPSS and is detailed in the following sections. All the output of the analysis will be provided by the author upon request.

5.4.1 Reliability

Figures 5.3, 5.4, 5.5 and 5.6 illustrate the reliability coefficients for all five psychometric scales (DIS, PEU, PU, PSSUQ and SUS). The clustered bar graph depicts the large, medium and small response target sizes for all questionnaires according to font size and font colour. All the scales were reliable, consistent with the English versions, except for the SUS. The lowest value for the Cronbach's alpha of the SUS was (α = 0.43) (Figure 5.6). In Figure 5.4, we see the Cronbach's alpha of the SUS α > 0.7 for the large, medium and small response target size. In Figure 5.6, the Cronbach's alpha of the SUS α > 0.7 only for the small response target size. For all other cases (Figures 5.3, 5.5 and 5.6), the alpha value of the SUS varied between 0.63 and 0.69.

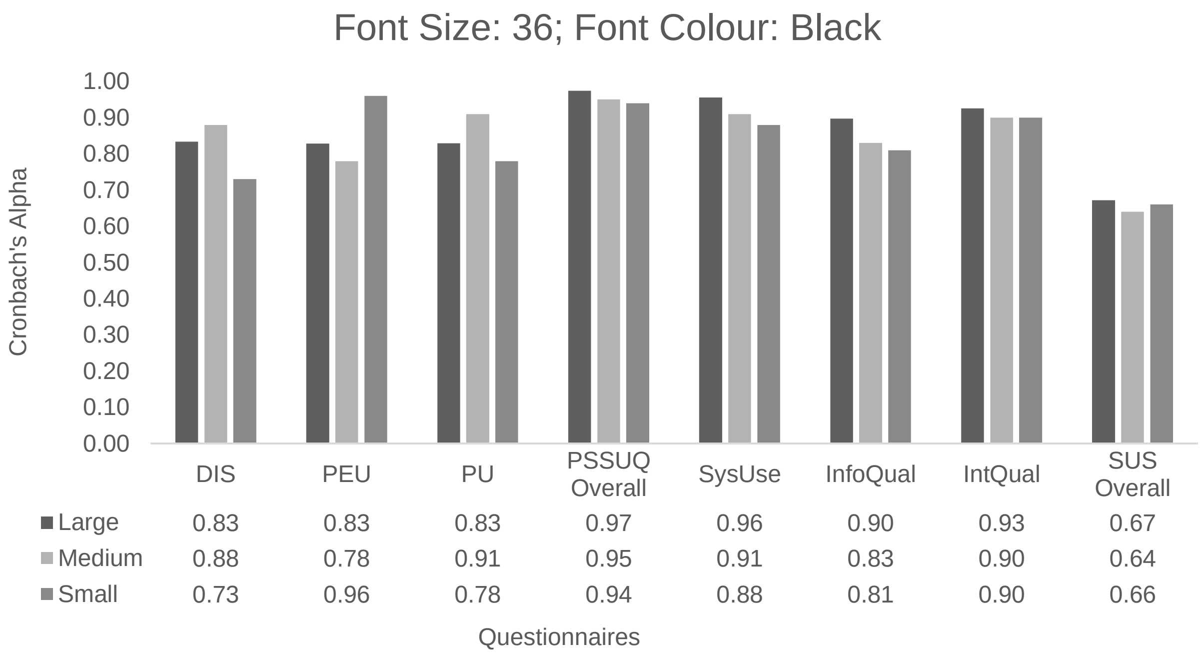


Figure 5.3. Reliability analysis (font size: 36 point; font colour: black) Study 1.

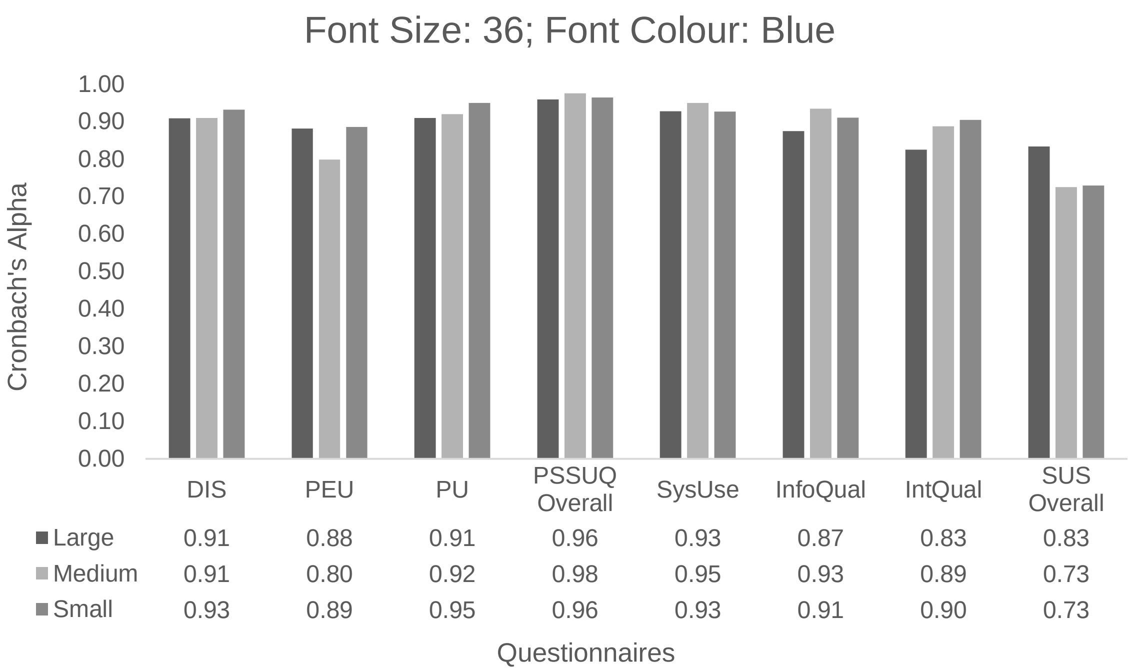
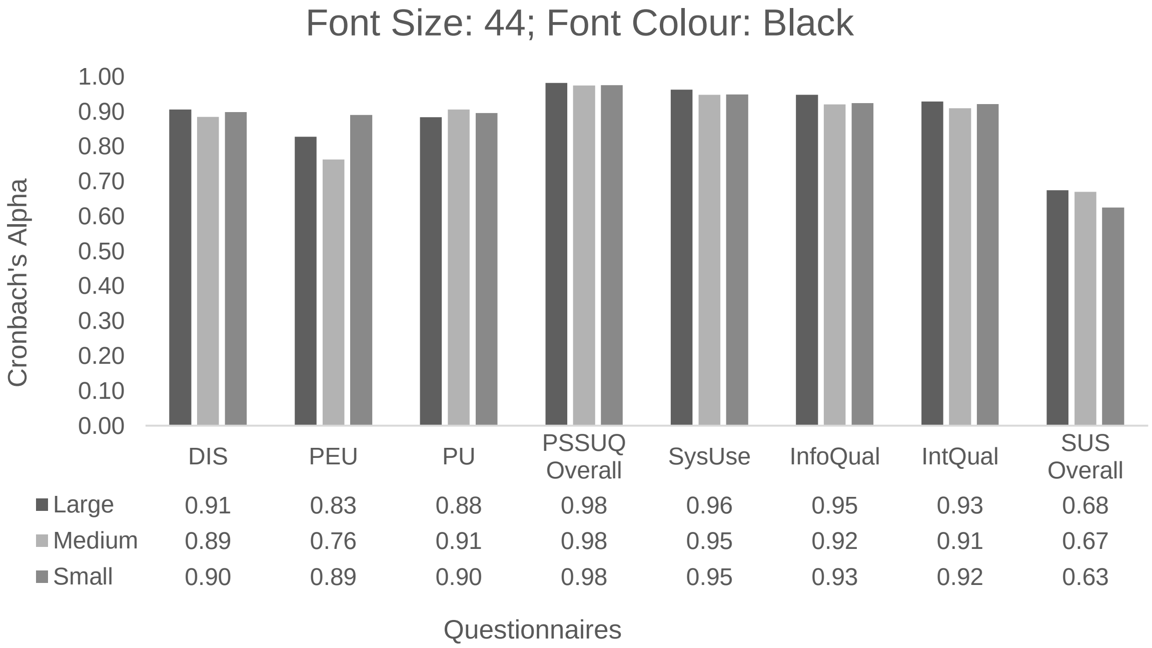
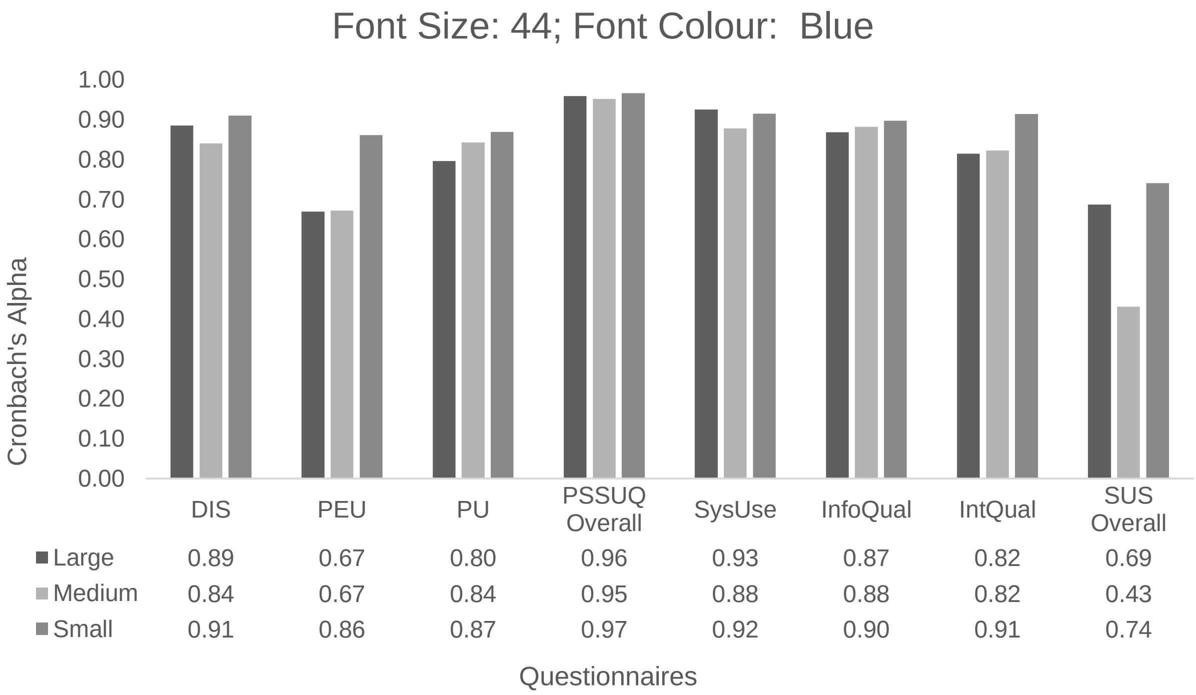


Figure 5.4. Reliability analysis (font size: 36 point; font colour: blue) Study 1.



	DIS	PEU	PU	PSSUQ Overall	SysUse	InfoQual	IntQual	SUS Overall
Large	0.91	0.83	0.88	0.98	0.96	0.95	0.93	0.68
Medium	0.89	0.76	0.91	0.98	0.95	0.92	0.91	0.67
Small	0.90	0.89	0.90	0.98	0.95	0.93	0.92	0.63

Figure 5.5. Reliability analysis (font size 44 point; font colour: black) Study 1.



	DIS	PEU	PU	PSSUQ Overall	SysUse	InfoQual	IntQual	SUS Overall
Large	0.89	0.67	0.80	0.96	0.93	0.87	0.82	0.69
Medium	0.84	0.67	0.84	0.95	0.88	0.88	0.82	0.43
Small	0.91	0.86	0.87	0.97	0.92	0.90	0.91	0.74

Figure 5.6. Reliability analysis (font size: 44 point; font colour: blue) Study 1.

The reliability analysis for the perceived enjoyment (PE) construct is presented in Figure 5.7. The clustered bar graph depicts the large, medium and small response target sizes for all combinations of font size and font colour. The scale possessed high reliability, consistent with the English version, for all combinations of the design parameters (font size, font colour and response target size).

5.4.1.1 Summary of Reliability analysis

In summary, the reliability for all the psychometric questionnaires (DIS, PEU, PU, PSSUQ and PE) was good, with the exception of SUS.

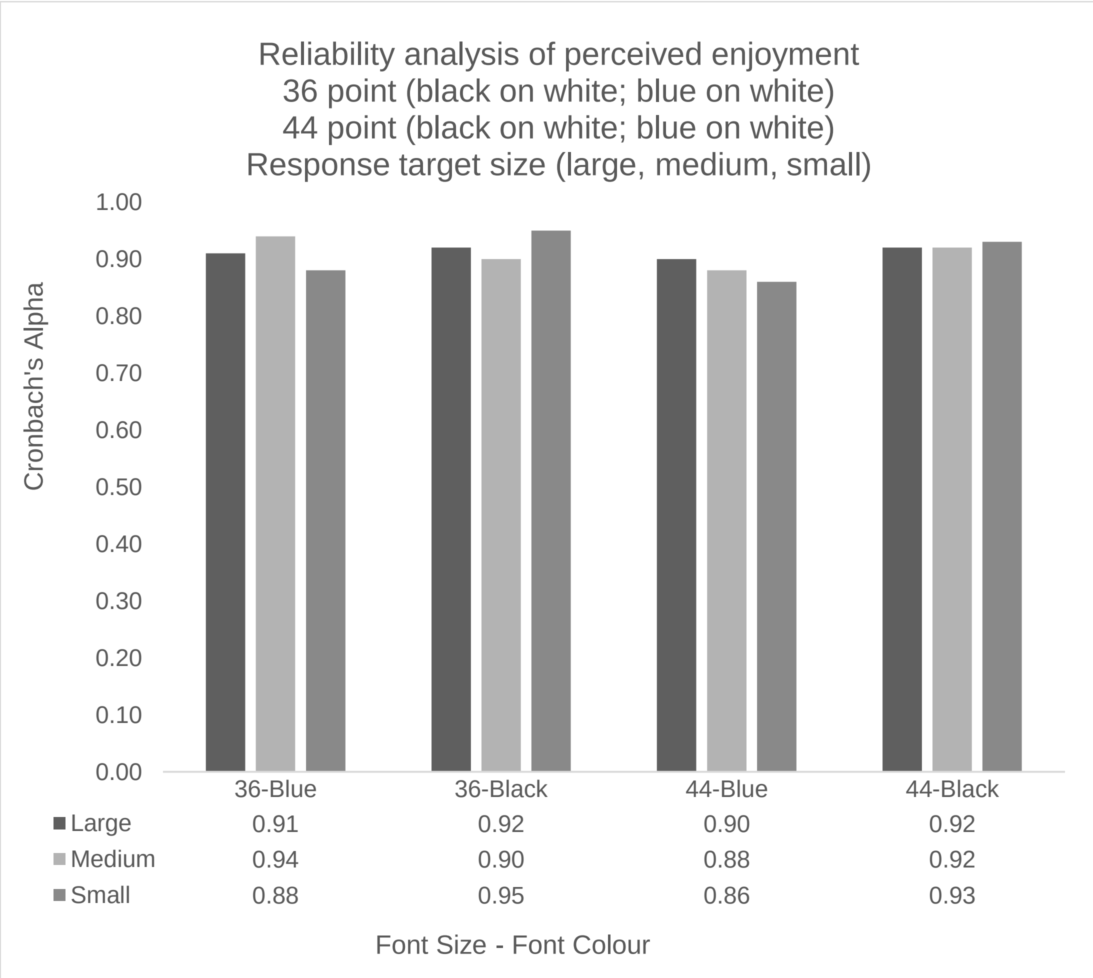


Figure 5.7. Reliability analysis of perceived enjoyment (all design parameter combinations) Study 1.

5.4.2. Validity

Pearson's correlation coefficient is used to assess criterion-related between the constructs disorientation, perceived ease of use, perceived usefulness, PSSUQ, the subscales of PSSUQ: SysUse, InfoQual, IntQual and SUS. The results are reported according to the response target size (see Tables 5.2, 5.3 and 5.4).

5.4.2.1 Large response target size

The correlation between the questionnaires for large response target size is presented in Table 5.2. Consistent with previous research (van Schaik and Ling, 2005), a moderate to strong positive correlation was found between: PU and PEU $r = .51$ $p < 0.01$. A significantly strong correlation was observed consistent with previous research (Lewis, 1995; 2002) between the subscales SysUse and InfoQual $r = .91$; SysUse and IntQual $r = .92$; InfoQual and IntQual $r = .86$ all $p < 0.01$;

Table 5.2. Correlations between constructs (response target size: large)

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	0.03	-.34**	-.17*	-.18*	-0.12	-.21*	-.33**
PU		.51**	.68**	.66**	.71**	.59**	.25**
PEU			.63**	.63**	.58**	.65**	.41**
PSSUQ				.98**	.96**	.95**	.35**
SYS					.91**	.92**	.37**
INFO						.86**	.30**
INT							.39**

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

5.4.2.2 Medium response target size

The correlation between the questionnaires for medium response target size is presented in Table 5.3. A moderately strong correlation consistent with previous research (van Schaik and Ling, 2005) was noted between: PU and PEU $r = .51$ $p < 0.01$. A significantly strong correlation was observed consistent with previous research (Lewis, 1995; 2002) between the subscales SysUse and InfoQual $r = .89$; SysUse and IntQual $r = .90$; InfoQual and IntQual $r = .85$; all $p < .01$).

Table 5.3. Correlations between constructs (response target size: medium)

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	-0.06	-0.13	-.25**	-.31**	-.17*	-.27**	-0.14
PU		.51**	.62**	.60**	.62**	.55**	.20*
PEU			.58**	.56**	.54**	.47**	.21*
PSSUQ				.97**	.96**	.93**	.17*
SYS					.89**	.90**	.23**
INFO						.85**	0.13
INT							.18*

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

5.4.2.3 Small response target size

The correlation between the questionnaires for small response target size is presented in Table 5.4. A moderately strong positive correlation consistent with previous research (van Schaik and Ling, 2005) was observed between: PU and PEU $r = .52$ $p < 0.01$. A significantly strong correlation was observed consistent with previous research (Lewis, 1995; 2002) between the subscales SysUse and InfoQual $r = .88$; SysUse and IntQual $r = .89$; InfoQual and IntQual $r = .84$; all $p < .01$).

Table 5.4. Correlations between constructs (response target size: small)

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	-0.08	-.27**	-0.15	-.19*	-0.13	-.20*	-.27**
PU		.52**	.62**	.61**	.61**	.55**	.20*
PEU			.67**	.67**	.63**	.63**	.30**
PSSUQ				.96**	.95**	.93**	.33**
SYS					.88**	.89**	.34**
INFO						.84**	.31**
INT							.34**

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

5.4.2.5 Summary of validity

In summary, for the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS), a high positive correlation $r > .50$ $p < 0.01$ was observed between PU and PEU across the three response target sizes consistent with previous research (van Schaik and Ling, 2005).

A strong correlation was observed between the three subscales: SysUse, InfoQual and IntQual of the PSSUQ construct across the three response target sizes, consistent with previous research by Lewis (1995; 2002).

5.4.3. Factor structure of PSSUQ

Lewis (1995, 2002) used principal axis factoring method with varimax rotation to determine the factor structure of PSSUQ. Close observation of the correlation values between the PSSUQ items, revealed moderately strong values within the PSSUQ subscales ranging from 0.54 to 0.75 and slightly lower correlation values 0.35 to 0.48 between the PSSUQ subscales.

This pattern of correlation provides some initial evidence for the factors of PSSUQ, with higher correlations among items within the subscales and lower correlations among items between the subscales. For the purpose of validating the PSSUQ construct administered in this research study, based on previous research, principal axis factoring method was used in this study and oblimin rotation was consistently used across all combination of design parameters to allow correlations between the factors. The KMO and Bartlett's test of sphericity values, indicated the suitability of data for factor extraction and is reported in Table 5.5.

Table 5.5. KMO and Bartlett's test of sphericity values (PSSUQ) across all design parameters.

Font size, text/background colour	Large	Medium	Small
36 point, black on white	*KMO* = .77; Bartlett: $\chi^2(171) = 611.82$	*KMO* = .78; Bartlett: $\chi^2(171) = 535.14$	*KMO* = .71; Bartlett: $\chi^2(171) = 517.01$
36 point, blue on white	*KMO* = .71; Bartlett: $\chi^2(171) = 632.86$	*KMO* = .64; Bartlett: $\chi^2(171) = 787.89$	*KMO* = .71; Bartlett: $\chi^2(171) = 625.89$
44 point, black on white	*KMO* = .92; Bartlett: $\chi^2(171) = 1346.3$	*KMO* = .94; Bartlett: $\chi^2(171) = 1097.8$	*KMO* = .92; Bartlett: $\chi^2(171) = 1157.4$
44 point, blue on white	*KMO* = .78; Bartlett: $\chi^2(171) = 789.61$	*KMO* = .79; Bartlett: $\chi^2(171) = 655.49$	*KMO* = .79; Bartlett: $\chi^2(171) = 959.7$

Note: all results of Bartlett's test of sphericity are significant, $p < 0.001$.

The factor analysis results are reported according to the design parameters font size and text/background colour, for all three response target sizes large, medium and small (Tables 5.7, 5.8, 5.9 and 5.10). It is to be noted that PSSUQ has a three-factor structure, as reported by Lewis (1995, 2002). The questionnaire items that load on the three factors obtained from previous studies are detailed in Table 5.6 (also see Section 3.5).

Table 5.6. Questionnaire items that load on the three factors of PSSUQ.

Factors	1995[a]	2002
SysUse	Items 1 to 7	Items 1 to 8 and 19
InfoQual	Items 9 to 15	Items 9 to 15
IntQual	Items 16 to 19	Items 16 to 18

Note: questionnaire items as reported by Lewis (2002).

[a] The first version of the PSSUQ (Lewis, 1992) did not contain questionnaire item 8.

Table 5.7. Factor analysis of PSSUQ (font size: 36 point; text/background colour: black on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	3	1	2	3	1	2	3
Q1	0.40	0.55	0.54	0.82				0.68	
Q2	0.41	0.73		0.73	0.53		0.56	0.68	
Q3	0.77	0.43		0.36	0.83		0.80		
Q4	0.65	0.41	0.45	0.86			0.67		0.43
Q5	0.78	0.46	0.30	0.88			0.56	0.58	
Q6	0.54	0.40	0.55		0.80		0.81		
Q7	0.67	0.50	0.39	0.77			0.60	0.62	
Q8	0.75		0.39		0.76	0.52			0.69
Q9	0.41	0.56				0.49			0.57
Q10			0.87		0.65	0.44	0.60		
Q11	0.72			0.70		0.35	0.59		0.40
Q12	0.48		0.78		0.68		0.33	0.73	
Q13		0.46	0.48			0.68		0.82	0.36
Q14	0.81	0.40	0.33	0.75		0.48	0.43	0.64	
Q15	0.67	0.48	0.37	0.82			0.71	0.37	0.43
Q16	0.48	0.56	0.44	0.69	0.50		0.80		
Q17		0.93		0.47	0.59		0.81		
Q18	0.65	0.54	0.36	0.62	0.53		0.61	0.63	
Q19	0.66	0.49	0.39	0.73	0.47		0.79	0.38	
Eigenvalues	13.20	1.29	1.04	10.21	2.41	1.66	9.89	2.13	1.76
% Variance	34.36	23.90	20.11	37.01	23.28	10.06	34.74	22.84	9.73
No. of items	10	5	4	11	6	2	9	8	2

Note: factor loadings < 0.3 were suppressed.

Table 5.8. Factor analysis of PSSUQ (font size: 36 point; text/background colour: blue on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	3	1	2	3	1	2	3
Q1	0.78			0.82			0.76		0.51
Q2	0.58	0.42	0.49	0.75			0.79	0.32	
Q3	0.57	0.45	0.33	0.58	0.68		0.32	0.76	
Q4	0.55	0.74		0.54	0.61	0.42	0.30	0.85	
Q5	0.43	0.73		0.38	0.73	0.42	0.30	0.78	
Q6	0.57	0.38	0.39	0.64	0.43	0.41	0.48	0.70	
Q7	0.55		0.70	0.95			0.74	0.43	
Q8	0.38	0.79		0.61	0.49			0.67	0.42
Q9		0.74		0.38		0.88		0.40	0.75
Q10		0.33	0.75		0.51	0.70	0.44		0.80
Q11		0.71	0.39		0.41	0.72		0.41	0.81
Q12	0.59		0.54	0.78	0.50		0.58		
Q13		0.77	0.31		0.75	0.37	0.58	0.41	0.42
Q14	0.70	0.53		0.31	0.69	0.53	0.50		0.69
Q15	0.72			0.69	0.43		0.66	0.49	0.33
Q16	0.93			0.47	0.58	0.49	0.66	0.60	
Q17	0.40	0.64	0.34	0.63	0.37	0.35	0.49	0.57	
Q18	0.51	0.46		0.34	0.77	0.33	0.62	0.35	0.37
Q19	0.60	0.47	0.36	0.80		0.35	0.80	0.31	
Eigenvalues	11.16	1.87	1.27	13.16	1.79	0.82	11.76	1.65	1.47
% Variance	29.46	27.91	13.63	34.41	26.48	19.31	29.47	25.70	19.50
No. of items	10	7	2	8	8	3	9	6	4

Note: factor loadings < 0.3 were suppressed.

Table 5.9. Factor analysis of PSSUQ (font size: 44 point; text/background colour: black on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	3	1	2	3	1	2	3
Q1	0.53	0.55		0.39	0.38	0.75	0.56	0.55	
Q2		0.77	0.35	0.67	0.41	0.42	0.66	0.43	0.32
Q3	0.64	0.52	0.31	0.45	0.49	0.49	0.47	0.54	0.47
Q4	0.50	0.68	0.34	0.69		0.42	0.64		0.63
Q5	0.73	0.45	0.30	0.71	0.47		0.41	0.63	0.49
Q6	0.55	0.63	0.36	0.35	0.34	0.76	0.59	0.32	0.45
Q7	0.47	0.66	0.40	0.40	0.75		0.78	0.38	
Q8	0.57	0.52	0.51	0.68	0.48	0.32	0.45	0.64	0.36
Q9	0.37	0.37	0.74	0.59	0.45	0.32		0.35	0.67
Q10	0.40	0.39	0.65	0.33	0.65	0.36	0.67		0.34
Q11			0.78	0.45	0.42	0.40	0.57		0.60
Q12	0.62	0.58	0.37	0.69	0.39	0.33	0.60	0.35	0.55
Q13	0.37	0.71	0.39	0.65	0.32	0.31	0.44	0.55	0.50
Q14	0.80		0.40	0.37	0.69	0.31	0.63	0.43	0.39
Q15	0.65	0.46	0.44	0.33	0.67	0.49	0.70	0.41	
Q16	0.63	0.46	0.40	0.43	0.68	0.36	0.36	0.48	0.66
Q17	0.53	0.38	0.56	0.32	0.45	0.69	0.33	0.65	0.42
Q18	0.70	0.38	0.45	0.61	0.31	0.47	0.39	0.48	0.71
Q19	0.65	0.46	0.40	0.58		0.66	0.37	0.77	0.32
Eigenvalues	14.37	0.82	0.67	13.08	0.89	0.83	13.32	0.94	0.75
% Variance	31.38	27.21	21.75	28.15	24.12	21.76	29.02	23.14	22.85
No. of items	10	5	4	9	6	4	9	6	4

Note: factor loadings < 0.3 were suppressed.

Table 5.10. Factor analysis of PSSUQ (font size: 44 point; text/background colour: blue on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	3	1	2	3	1	2	3
Q1	0.71		0.56	0.72	0.47		0.72	0.31	0.44
Q2	0.66		0.33	0.54		0.43	0.57	0.33	0.63
Q3	0.68	0.34			0.69		0.48	0.40	0.35
Q4	0.53	0.80				0.82	0.50	0.66	
Q5			0.83	0.77			0.53		0.61
Q6	0.70	0.44	0.33	0.46		0.51	0.61		0.46
Q7	0.38	0.53	0.48	0.30	0.81		0.88		
Q8	0.58	0.41	0.41	0.52	0.48	0.31	0.40	0.37	
Q9	0.31	0.41	0.37	0.72	0.31	0.33		0.72	0.30
Q10	0.79	0.43			0.54	0.56	0.32	0.85	
Q11	0.39	0.49		0.44				0.48	0.55
Q12	0.36	0.47	0.32		0.75	0.44	0.86		
Q13	0.44		0.52	0.31	0.53	0.58		0.83	0.44
Q14			0.72	0.38		0.58			0.88
Q15	0.64	0.33	0.38		0.59	0.40	0.89		
Q16	0.35		0.55	0.33	0.57		0.59	0.38	0.45
Q17		0.83		0.77	0.32		0.74	0.34	0.33
Q18		0.79		0.42	0.55	0.31	0.73	0.35	
Q19	0.66		0.55	0.67		0.39	0.55	0.45	0.57
Eigenvalues	11.06	1.41	1.15	10.24	1.41	1.34	11.98	1.67	1.10
% Variance	26.50	21.24	19.57	23.82	22.55	16.64	33.98	21.10	18.88
No. of items	8	7	4	8	6	5	10	4	5

Note: factor loadings < 0.3 were suppressed.

5.4.3.1 Summary of factor structure of PSSUQ

Considering all the design parameter combinations, the original factor structure of PSSUQ reported in research studies (Lewis, 1995 and 2002) could not be replicated. In addition, questionnaire items loaded differently on the three factors and no pattern could be identified. Hence, the factors could not be named.

5.4.4. Factor structure of SUS

The latest research reported by Lewis and Sauro (2017) presented a tone model for the factor structure of SUS, although it initially displayed a unidimensional structure (Brooke, 1996) (see Section 3.6). For the purpose of validating the SUS construct administered in this research study, the correlation values were examined. Within the items of each of the two SUS subscales (positive and negative), fairly moderate correlations with values between 0.39 and 0.49 were observed. However, the items between the two subscales displayed very low correlations values between 0.02 and 0.28. This pattern of correlation provides some initial evidence for the factors of SUS, with higher correlations among items within the subscales and lower correlations among items between the subscales. Previous research by Lewis and Sauro (2017) reported the use of all three extraction techniques such as principle components analysis (strictly not a factor analytic method, but commonly used), unweighted least squares and maximum likelihood extraction technique with varimax rotation. In the current research, principle axis factoring with oblimin rotation was initially used. However, with oblimin rotation, although the number of iterations were increased to 999, according to the SPSS results, the communality of a variable (not named in the results) exceeded 1.0 in iteration 25 for the medium response target size.

There were no results obtained for the medium response target size when oblimin rotation was used. Therefore, in order to consistently obtain results for small, medium and large response target sizes, principle axis factoring with varimax rotation was used. It should be noted that varimax rotation was also employed in previous research. The KMO and Bartlett's test of sphericity obtained, indicated the suitability of data for factor extraction and is reported in Table 5.11

Table 5.11. KMO and Bartlett's test of sphericity values (SUS) across all design parameters.

Font size, text/background colour	Large	Medium	Small
36 point, black on white	*KMO* = .47; Bartlett: $\chi^2(45) = 108.42$	*KMO* = .64; Bartlett: $\chi^2(36) = 87.96$	*KMO* = .47; Bartlett: $\chi^2(45) = 119.88$
36 point, blue on white	*KMO* = .72; Bartlett: $\chi^2(45) = 147.30$	*KMO* = .73; Bartlett: $\chi^2(36) = 142.66$	*KMO* = .70; Bartlett: $\chi^2(45) = 112.13$
44 point, black on white	*KMO* = .70; Bartlett: $\chi^2(45) = 159.23$	*KMO* = .71; Bartlett: $\chi^2(36) = 128.71$	*KMO* = .58; Bartlett: $\chi^2(45) = 130.11$
44 point, blue on white	*KMO* = .63; Bartlett: $\chi^2(45) = 183.01$	*KMO* = .58; Bartlett: $\chi^2(36) = 181.03$	*KMO* = .53; Bartlett: $\chi^2(45) = 227.06$

Note: all results of Bartlett's test of sphericity are significant, $p < 0.001$.

The question items with a positive tone (wording) load on one factor while the items with a negative tone (wording) load on the other. The questionnaire items that load on the two factors as reported in various studies are detailed in Table 5.12 (also see Section 3.6).

Table 5.12. Factors of SUS as reported in various studies.

Study	Factors	Questionnaire items
(Lewis, Brown, & Mayes, 2015)	Positive	1, 3, 5, 7, 9
	Negative	2, 4, 6, 8, 10
(Kortum & Sorber, 2015)	Factor 1	1, 3, 5, 7, 9
	Factor 2	2, 4, 6, 8, 10
(Kortum & Sorber, 2015)	Factor 1	1, 3, 5, 6, 7, 8, 9
	Factor 2	2, 4, 10
(Sauro & Lewis, 2011)	Factor 1	1, 2, 3, 5, 7, 9
	Factor 2	4, 6, 8, 10
(Borsci, Federici, Bacci, Gnaldi, & Bartolucci,	Factor 1	1,9
2015; Lewis, Utesch, et al., 2015)	Factor 2	2, 3, 4, 5, 6, 7, 8, 10

In the current study, the factor analysis results are reported according to the design parameters, font size and text/background colour, for all three large, medium and small response target sizes (Tables 5.13, 5.14, 5.15 and 5.16).

Table 5.13. Factor analysis of SUS (font size: 36 point; text/background colour: black on white).

	Large		Medium		Small	
	Factors		Factors		Factors	
Items	Negative	Positive	1	2	1	2
Q1		0.82	[a]	[a]	0.34	0.56
Q2	0.68			0.79	0.42	
Q3	0.30	0.50	0.40			0.69
Q4	0.63		-0.32	0.92	0.47	0.62
Q5		0.61	0.69			0.56
Q6	0.77		0.51		0.69	
Q7	0.35	0.49	0.78		-0.74	
Q8	0.62		-0.37			
Q9		0.63		0.65	0.64	0.43
Q10	0.31		0.75		0.65	
Eigenvalues	2.73	2.59	2.95	2.18	3.28	1.96
% Variance	21.92	20.65	25.67	21.98	23.66	18.32
No. of items	5	5	6	3	5	4

Note: factor loadings < 0.3 were suppressed.
[a] Item Q1 was removed for medium response target size due to high correlation.

Table 5.14. Factor analysis of SUS (font size: 36 point; text/background colour: blue on white).

	Large		Medium		Small	
	Factors		Factors		Factors	
Items	Positive	Negative	1	2	1	2
Q1	0.69		[a]	[a]	0.61	
Q2		0.85	0.45			0.55
Q3	0.52	0.35		0.84		0.44
Q4		0.80	0.96		0.70	
Q5	0.73			0.57		0.89
Q6		0.64	0.79		0.87	
Q7	0.64			0.75		0.56
Q8		0.70	0.69		0.75	
Q9	0.92		0.86			0.64
Q10	0.43	0.57		0.82	0.65	
Eigenvalues	4.07	2.32	3.39	2.67	3.16	2.51
% Variance	28	27.65	33.64	26.08	26.54	20.58
No. of items	5	5	5	4	5	5

Note: factor loadings < 0.3 were suppressed.
[a] Item Q1 was removed for medium response target size due to high correlation with Q9.

Table 5.15. Factor analysis of SUS (font size: 44 point; text/background colour: black on white).

	Large		Medium		Small	
	Factors		Factors		Factors	
Items	Negative	Positive	1	2	Negative	Positive
Q1		0.69	[a]	[a]		0.36
Q2	0.71		0.62		0.54	
Q3		0.52		0.73		0.50
Q4	0.66		0.63		0.71	
Q5		0.62		0.55		0.50
Q6	0.63		0.65		0.66	
Q7		0.69		0.58		0.58
Q8	0.83		0.61		0.55	
Q9		0.51	0.58			0.44
Q10	0.47			0.56	0.56	
Eigenvalues	2.77	2.55	2.67	2.14	2.59	1.99
% Variance	22.70	19.52	22.47	17.89	19.81	12.85
No. of items	5	5	5	4	5	5

Note: factor loadings < 0.3 were suppressed.
[a] Item Q1 was removed for medium response target size due to high correlation with Q9.

Table 5.16. Factor analysis of SUS (font size: 44 point; text/background colour: blue on white).

	Large		Medium		Small	
	Factors		Factors		Factors	
Items	Negative	Positive	1	2	Negative	Positive
Q1		0.65	[a]	[a]		0.68
Q2	0.81		0.73		0.80	
Q3		0.70		0.67		0.42
Q4	0.51		0.47		0.78	
Q5		0.51		0.53		0.59
Q6	0.49		0.64	0.32	0.81	
Q7		0.54		0.48		0.68
Q8	0.75		0.77		0.66	
Q9		0.79	0.74			0.79
Q10	0.85			0.88	0.70	
Eigenvalues	2.95	2.65	3.46	1.8	3.32	2.62
% Variance	24.42	21.68	27.03	20.88	28.90	20.93
No. of items	5	5	5	4	5	5

Note: factor loadings < 0.3 were suppressed.
[a] Item Q1 was removed for medium response target size due to high correlation with Q9.

5.4.4.1 Summary of factor structure of SUS

Considering the factor structure of SUS, a clear pattern of results was found consistent with the previous research (e.g. Lewis & Sauro, 2015; 2017). For the large response target size, the tone model (see Section 3.6) was evident across all combinations of the design parameters font size and text/background colour. A consistent pattern was also observed for the small response target size, for both text/background colour combinations and font size: 44 point. For the medium response target size, with one item removed, although a two-factor structure was evident, it did not follow the tone model consistent with the previous studies.

5.4.5. Factor structure of DIS, PEU and PU

For the purpose of validating the DIS, PEU and PU constructs administered in this research study, the correlation values were examined. Within the items of each of the scales, a moderately strong correlation was observed with values ranging from 0.39 to 0.69 for the DIS construct, 0.52 to 0.63 for the PU construct and 0.52 to 0.60 for the PEU construct. Weak correlation values were observed between the items of the DIS, PEU and PU constructs while a slight moderate correlation existed between the items of the PEU and PU constructs. This pattern of correlation provides some initial evidence for the factors of DIS, PEU and PU, with higher correlations among items within the subscales and lower correlations among items between the subscales. Similar to previous research reported by van Schaik and Ling (2003, 2005, and 2007), principle axis factoring with oblimin rotation was initially employed for small, medium and large response target sizes to detect the factor structure of the three questionnaires together. The KMO and Bartlett's test of sphericity values, indicated the suitability of data for factor extraction and is reported in Table 5.17.

Table 5.17. KMO and Bartlett's test of sphericity values (DIS, PEU and PU) across all design parameters.

Font size, text/background colour	Large	Medium	Small
36 point, black on white	*KMO* = .35; Bartlett: $\chi^2(91) = 250.89$	*KMO* = .63; Bartlett: $\chi^2(91) = 336.29$	*KMO* = .42; Bartlett: $\chi^2(91) = 286.03$
36 point, blue on white	*KMO* = .60; Bartlett: $\chi^2(91) = 336.71$,	*KMO* = .67; Bartlett: $\chi^2(91) = 282.57$	*KMO* = .62; Bartlett: $\chi^2(91) = 428.87$
44 point, black on white	*KMO* = .79; Bartlett: $\chi^2(91) = 519.05$	*KMO* = .77; Bartlett: $\chi^2(91) = 463.24$	*KMO* = .78; Bartlett: $\chi^2(91) = 551.05$
44 point, blue on white	*KMO* = .65; Bartlett: $\chi^2(91) = 331.63$	*KMO* = .65; Bartlett: $\chi^2(91) = 292.99$	*KMO* = .68; Bartlett: $\chi^2(91) = 445.54$

Note: all results of Bartlett's test of sphericity are significant, $p < 0.001$.

Based on literature studies, research by van Schaik and Ling (2003, 2005, and 2007) for online psychometrics, presented a three-factor solution: (1) disorientation, (2) perceived ease of use and (3) perceived usefulness (see Table 5.18, Section 5.4.5). In this study, results of factor analysis were obtained, reported and compared with the previous studies by van Schaik and Ling. The factor analysis results are reported according to the design parameters font size and text/background colour, for all three large, medium and small response target sizes (Tables 5.19, 5.20, 5.21 and 5.22). Research by Ahuja Webster (2001) and van Schaik and Ling (2003, 2005, and 2007) indicate that when the questionnaires are used together, disorientation items load on one factor; perceived ease of use items on a second factor and perceived usefulness items on a third factor. Table 5.18 summarizes the factor structure with the items when disorientation, perceived ease of use and perceived usefulness were used together in various studies.

Table 5.18. Factor structure of DIS, PEU, and PU as reported in various studies.

Study	Factors	Questionnaire items
(Ahuja & Webster, 2001)	DIS	Items 1 to 7
	PEU	Items 1 to 3
(Schaik & Ling, 2003)	DIS	Items 1 to 7
	PEU	Items 1 to 3
(Schaik & Ling, 2005b)	DIS	Items 1 to 7
	PEU	Items 1 to 3
	PU	Items 1 to 4
(van Schaik & Ling, 2007)	DIS	Items 1 to 7
	PEU	Items 1 to 3
	PU	Items 1 to 4

Table 5.19. Factor analysis of DIS, PEU and PU (font size: 36 point; text/background colour: black on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	3	1	2	3	PEU	DIS	PU
DIS_Q1	0.73		0.36		0.812			0.785	
DIS_Q2	0.74		0.36		0.874			0.714	
DIS_Q3	0.41				0.591				
DIS_Q4			0.54		0.422	0.519		0.528	0.311
DIS_Q5	0.51				0.899			0.543	
DIS_Q6			0.72	-0.402	0.642			0.428	
DIS_Q7	0.32		0.47		0.769			0.677	
PEU_Q1	-0.73			0.704			0.902		
PEU_Q2	-0.49	0.43	-0.41	0.439			0.948		
PEU_Q3	-0.93			0.979			0.977		
PU_Q1		0.80		0.522		0.640			0.519
PU_Q2		0.70	0.33	0.755					0.959
PU_Q3	-0.466	0.52	0.43	0.865					0.976
PU_Q4		0.96		0.409	-0.304	0.687			
Eigenvalues	4.74	3.02	1.50	5.71	3.19	1.43	3.28	2.70	2.49
% Variance	25.77	19.02	13.45	38.88	20.52	8.26	21.52	38.28	54.42
No. of items	7	4	3	5	6	3	3	6	3

Note: factor loadings < 0.3 are suppressed.

Table 5.20. Factor analysis of DIS, PEU and PU (font size: 36 point; text/background colour: blue on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	DIS	PU	PEU	DIS	PU	PEU	DIS	PU	PEU
DIS_Q1	0.53		-0.71	0.55		-0.34	0.90		
DIS_Q2	0.77		-0.43	0.69			0.86		
DIS_Q3	0.60			0.80			0.58		-0.46
DIS_Q4	0.56			0.79			0.50		
DIS_Q5	0.74		-0.39	0.87			0.85		
DIS_Q6	0.91			0.81			0.93		
DIS_Q7	0.69			0.78			0.85		
PEU_Q1	-0.35		0.65			0.75			0.83
PEU_Q2			0.96			0.90			0.82
PEU_Q3	-0.45		0.57			0.56			0.70
PU_Q1		0.92			0.97			0.80	
PU_Q2		0.69			0.86			0.84	
PU_Q3		0.93			0.89			0.92	
PU_Q4		0.81	0.32		0.72			0.98	
Eigenvalues	5.6	3.67	1.36	5.10	4.03	1.28	6.31	3.38	1.6
% Variance	27.02	22.66	19.95	33.87	60.83	67.86	43.36	66.30	76.04
No. of items	7	3	4	7	4	3	7	4	3

Note: factor loadings < 0.3 are suppressed.

Table 5.21. Factor analysis of DIS, PEU and PU (font size: 44 point; text/background colour: black on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	DIS	PEU	PU	1	DIS	3	1	DIS	3
DIS_Q1	0.81				0.74			0.66	
DIS_Q2	0.66	-0.35			0.63			0.67	
DIS_Q3	0.60				0.78			0.66	
DIS_Q4	0.67		-0.42		0.56			0.63	-0.38
DIS_Q5	0.72				0.75			0.82	
DIS_Q6	0.87				0.83			0.88	
DIS_Q7	0.66	-0.34			0.73			0.74	
PEU_Q1		0.91				0.81	0.75		
PEU_Q2		0.75		0.52			0.79		
PEU_Q3		0.50		0.59			0.82		0.35
PU_Q1		0.49	0.61	0.95			0.82		
PU_Q2			0.95	0.85			0.90		
PU_Q3		0.50	0.62	0.73			0.81		
PU_Q4		0.58	0.49	0.87			0.76		
Eigenvalues	6.64	2.24	1.27	5.65	2.86	1.21	6.32	3.03	1.05
% Variance	27.74	20.67	17.46	37.74	55.61	62.33	43.05	62.32	67.36
No. of items	7	3	4	6	7	1	7	7	

Note: factor loadings < 0.3 are suppressed.

Table 5.22. Factor analysis of DIS, PEU and PU (font size: 44 point; text/background colour: blue on white).

	Large			Medium			Small		
	Factor			Factor			Factor		
Items	1	2	PEU	PU	2	PEU	DIS	PU	PEU
DIS_Q1	0.79				-0.88		0.83		
DIS_Q2	0.63	0.41			-0.88		0.80		
DIS_Q3	0.35	0.68			-0.43		0.83		
DIS_Q4	0.81	0.31		0.37			0.73		
DIS_Q5	0.80				-0.57		0.82		
DIS_Q6	0.75				-0.47	0.43	0.81		
DIS_Q7	0.56				-0.72		0.75	0.31	
PEU_Q1			1.00			0.94			0.69
PEU_Q2			0.33			0.62			0.59
PEU_Q3			0.55						1.06
PU_Q1		0.65	0.49	0.82				0.84	
PU_Q2		0.85	0.33	0.82				0.83	
PU_Q3		0.57		0.55				0.64	
PU_Q4	0.47	0.36		0.89				0.83	
Eigenvalues	5.25	2.42	1.39	4.95	2.24	1.61	5.13	4.14	1.19
% Variance	26.38	17.45	13.48	32.28	45.90	54.78	34.32	27.48	6.85
No. of items	7	4	3	5	6	2	7	4	3

Note: factor loadings < 0.3 are suppressed.

5.4.5.1 Summary of factor structure of DIS, PEU, PU

In summary, the factor structure of DIS, PEU and PU, consistent with research reported by van Schaik and Ling (2002, 2005 and 2007), Ahuja and Webster (2001) and Davis (1989), was observed only for certain combinations of design parameters font size, text/background colour and response target size, in particular

(1) 36 point, blue on white, large/medium and small (see Table 5.20);

(2) 36 point, black on white, small (see Table 5.19);

(3) 44 point, blue on white, small (see Table 5.22);

(4) 44 point, black on white, large (see Table 5.21).

Although a three-factor structure was evident for all other combinations of the design parameters a clear pattern of result could not be replicated. Research by van Schaik and Ling (2005, 2007), reported validation of these questionnaires together with design parameters (questionnaire layout and rating scales). In this research study, the validation of the questionnaires (DIS, PEU and PU) with design parameters response target size, font size, font colour is reported.

5.4.6 Factor structure of PE

Principle axis factoring method was employed to determine the factor structure of the perceived enjoyment scale. The KMO and Bartlett's test of sphericity values indicated the suitability of data for factor extraction and is reported in Table 5.23.

Table 5.23. KMO and Bartlett's test of sphericity values (PE) across all design parameters.

Font size, text/background colour	Large	Medium	Small
36 point, black on white	*KMO* = .73; Bartlett: $\chi^2(3) = 61.48$	*KMO* = .68; Bartlett: $\chi^2(91) = 50.22$	*KMO* = .70; Bartlett: $\chi^2(3) = 96.81$
36 point, blue on white	*KMO* = .71; Bartlett: $\chi^2(91) = 48.55$	*KMO* = .73; Bartlett: $\chi^2(91) = 65.38$	*KMO* = .74; Bartlett: $\chi^2(91) = 34.37$
44 point, black on white	*KMO* = .75; Bartlett: $\chi^2(91) = 125.41$	*KMO* = .76; Bartlett: $\chi^2(91) = 117.63$	*KMO* = .74; Bartlett: $\chi^2(91) = 136.32$
44 point, blue on white	*KMO* = .72; Bartlett: $\chi^2(91) = 83.62$	*KMO* = .72; Bartlett: $\chi^2(91) = 66.43$	*KMO* = .70; Bartlett: $\chi^2(91) = 63.22$

Note: all results of Bartlett's test of sphericity are significant, $p < 0.001$.

In order to compare the results, factor analysis from the study by Davis (1992) is provided as a reference (Table 5.24). The factor analysis results are reported in Tables 5.25, 5.26 and 5.27 according to the design parameters: font size and text/background colour, for all three large, medium and small response target sizes.

Table 5.24. Factor loadings reported in previous studies.

Study	PE items factor loadings	
(Davis et al., 1992)	Item 1	0.84
	Item 2	0.84
	Item 3	0.94

Table 5.25. Factor analysis of PE (response target size: large).

	36 point		44 point	
Items	Black on white	Blue on white	Black on white	Blue on white
Q1	0.81	0.84	0.84	0.89
Q2	0.99	0.81	0.91	0.89
Q3	0.99	0.86	0.97	0.69
Eigenvalues	2.73	2.40	2.64	2.35
% Variance	87.22	70.23	82.39	68.72
No. of items	3	3	3	3

Table 5.26. Factor analysis of PE (response target size: medium).

Items	36 point		44 point	
	Black on white	Blue on white	Black on white	Blue on white
Q1	0.66	0.94	0.88	0.79
Q2	0.67	0.98	0.92	0.94
Q3	0.80	0.84	0.88	0.81
Eigenvalues	2.52	2.69	2.59	2.43
% Variance	84.03	85.09	79.59	72.08
No. of items	3	3	3	3

Table 5.27. Factor analysis of PE (response target size: small).

Items	36 point		44 point	
	Black on white	Blue on white	Black on white	Blue on white
Q1	0.96	0.76	0.88	0.79
Q2	0.94	0.92	0.95	0.95
Q3	0.79	0.96	0.86	0.86
Eigenvalues	2.60	2.54	2.61	2.50
% Variance	80.83	78.10	80.57	75.80
No. of items	3	3	3	3

5.4.6.1 Summary of factor structure of PE

In summary, the factor structure of the perceived enjoyment scale across all design parameters, was consistent and similar to the structure reported in other studies (e.g. Davis 1992).

5.4.7 Mixed ANOVA

A repeated-measures ANOVA was conducted to test the main and interaction effects of font size, text/background colour and response target size on the total completion time of the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS). In addition, repeated-measures ANOVA was conducted to test the main and interaction effects of font size, text/background colour and response target size on the total score of the perceived enjoyment construct that was administered immediately after the respondents completed their response to the psychometric questionnaires for each large, medium and small response target size.

5.4.7.1 Time to complete the questionnaires

Total completion time was heavily positively skewed. An inverse transformation of the total completion time for the five questionnaires (DIS, PEU, PU, PSSUQ and SUS) reduced the skew and improved normality of distribution. Table 5.28 details the mean, standard deviation and the confidence intervals of the mean for the transformed and the retransformed time in seconds. The means (seconds) presented in the Table 5.28 indicate that completion time was shorter for font size 44 point than font size 36 point across all the three large, medium and small response target sizes.

Table5.28. Descriptives for completion time of questionnaires (Study 1).

		Transformed (log(original time[a]))						Retransformed (exp[log(time)])					
Response target size		Large		Medium		Small		Large		Medium		Small	
Font size (point)		36	44	36	44	36	44	36	44	36	44	36	44
mean		5.22	4.83	5.32	4.87	5.28	4.84	185.60	125.01	204.18	130.36	195.93	126.54
Std. Deviation		0.72	0.70	0.73	0.64	0.68	0.62	2.05	2.01	2.08	1.89	1.97	1.86
95% Confidence Interval for mean	Lower Bound	5.04	4.69	5.13	4.74	5.10	4.72	154.02	108.73	168.79	114.72	164.24	111.68
	Upper Bound	5.41	4.97	5.51	5.00	5.45	4.97	223.66	143.73	247.00	148.12	233.74	143.37

[a] Seconds.

A mixed 2×2×(3) ANOVA (Table 5.29) revealed that the main effect of font size F (1, 153) = 45.57, η_p^2 = 0.23, p < 0.01 was significant. There were no interaction effects for font size and font colour at the different levels of response target sizes.

Table 5.29. Mixed ANOVA summary table for questionnaire completion time.

Source	*df*	*SS*	*MS*	*F*	*P*	η_p^2
Font colour	1	0.05	0.05	0.11	0.74	0.00
Font size	1	20.62	20.62	45.57	0.00	0.23
Font colour × Font size	1	0.68	0.68	1.51	0.22	0.01
Error (Font size and Font colour)	153	69.21	0.45			
Target size	2	0.40	0.20	0.44	0.65	0.00
Target size × Font colour	2	0.09	0.04	0.10	0.91	0.00
Target size × Font size	2	0.06	0.03	0.06	0.94	0.00
Target size × Font colour × Font size	2	0.53	0.27	0.58	0.56	0.00
Error (Target size)	306	140.43	0.46			

Note: Target size: large, medium and small.

5.4.7.2 Perceived enjoyment

Table 5.30 details the mean, standard deviation and confidence intervals of the perceived enjoyment scores. The scores represent the sum of the three questionnaire items of the perceived enjoyment construct.

Table 5.30. Descriptives of perceived enjoyment total scores

		Perceived enjoyment (sum of scores)					
Response target size		Large		Medium		Small	
Font size (point)		36	44	36	44	36	44
mean (SD)		7.23	6.99	6.84	7.16	7.11	6.96
Std. Deviation		3.12	3.31	3.09	2.99	3.29	3.39
95% Confidence Interval for mean	Lower Bound	6.40	6.29	6.02	6.53	6.23	6.24
	Upper Bound	8.06	7.69	7.66	7.79	7.98	7.67

A mixed 2×2×(3) ANOVA was used to test the effects of font size, font colour and response target size on respondents' enjoyment experience. The details of the mixed ANOVA are presented in Table 5.31. No main effect or interaction effect was observed for font size, font colour and response target size on the perceived enjoyment scores.

Table 5.31. Mixed ANOVA summary table for perceived enjoyment.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Font colour	1	0.84	0.84	0.18	0.67	0.00
Font size	1	2.19	2.19	0.47	0.50	0.00
Font colour × Font size	1	0.00	0.00	0.00	0.98	0.00
Error (Font size and Font colour)	137	644.95	4.71			
Target size	2	1.15	0.62	0.75	0.47	0.01
Target size × Font colour	2	0.88	0.47	0.57	0.55	0.00
Target size × Font size	2	1.59	0.86	1.03	0.35	0.01
Target size × Font colour × Font size	2	0.64	0.35	0.42	0.64	0.00
Error (Target size)	253	211.97	0.84			

Note: Target size: large, medium and small response target size.

5.4.7.3 Summary of Repeated-measures ANOVA

In summary, completion time was shorter for font size 44 point than for font size 36 point across all the three large, medium and small response target sizes. This result confirms Hypothesis 1: completion time decreases with increasing font size. No evidence was obtained with regards to Hypothesis 2 (font colour), Hypothesis 3a (response target size), Hypothesis 8 (perceived enjoyment for font size), Hypothesis 9 (perceived enjoyment for colour contrast) and Hypothesis 10a (perceived enjoyment for response target size).

Therefore, according to the results of this experiment, the design parameter (font size: 44 point) led to faster completion time and is therefore used in the next study for online psychometric experiments on small-screen devices.

5.5 Discussion

In summary, this chapter presented the method and analysis to test the effect of three design parameters (font size, font colour and response target size) in online psychometrics. The hypothesis for each of these design parameters were stated. The method of an experiment to test each of the hypotheses was presented. The psychometric properties were tested for each of the combinations of the design parameters. The hypotheses were tested with the mixed-measures ANOVA.

5.5.1. Psychometric analysis

A summary of the psychometric properties of the questionnaires translated into Arabic is presented.

5.5.1.1 Reliability analysis

The five scales DIS (disorientation), PEU (perceived ease of use), PU (perceived usefulness), PSSUQ and SUS were found to be reliable across all three target sizes, except for SUS (see Section 5.4.1). The perceived enjoyment questionnaire also possessed high reliability.

5.5.1.2 Validity

Discriminant validity of DIS and SUS was generally confirmed through low or moderate correlations with PU, PEU and PSSUQ and its subscales (Tables 5.32, 5.33 and 5.34). Convergent validity was demonstrated through a substantial correlation between PU and PEU and between PSSUQ and its subscales of PSSUQ.

Table 5.32. Correlations between constructs (response target size: large).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS		-.34**					-.33**
PU		.51**	.68**	.66**	.71**	.59**	
PEU			.63**	.63**	.58**	.65**	.41**
PSSUQ				.98**	.96**	.95**	.35**
SYS					.91**	.92**	.37**
INFO						.86**	
INT							.39**

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.

** Correlation is significant at the 0.01 level (2-tailed).

Table 5.33. Correlations between constructs (response target size: medium).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS							
PU		.51**	.62**	.60**	.62**	.55**	
PEU			.58**	.56**	.54**	.47**	
PSSUQ				.97**	.96**	.93**	
SYS					.89**	.90**	
INFO						.85**	
INT							

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.

** Correlation is significant at the 0.01 level (2-tailed).

Table 5.34. Correlations between constructs (response target size: small).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS							
PU		.52**	.62**	.61**	.61**	.55**	
PEU			.67**	.67**	.63**	.63**	
PSSUQ				.96**	.95**	.93**	.33**
SYS					.88**	.89**	.34**
INFO						.84**	
INT							.34**

Note: PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual.
** Correlation is significant at the 0.01 level (2-tailed).

5.5.1.3 Factor structure

It should be noted that at the time of this research, there was no available translated version into Arabic of the six questionnaires DIS, PEU, PU, PSSUQ, SUS and perceived enjoyment. Therefore, the first attempt of the translated version of these questionnaires has been adopted. As far as the author is aware, the factor structure of these questionnaires has not been validated in any research with manipulation of design parameters. More experimental studies and reporting of results will help in establishing the factor structure of the five questionnaires.

5.5.2. Repeated-measures ANOVA

5.5.2.1 Hypothesis 1 and 8

Repeated-measures design ANOVA revealed a significant main effect for font size thus validating Hypothesis 1: completion time decreases with increasing font size. This was evident as the completion time (in seconds) for font size: 44 point had the smaller mean value across all response target sizes (large: mean = 4.83, medium: mean = 4.87, small: mean = 4.84) compared to the font size: 36 point (large: mean = 5.22, medium: mean = 5.32, small: mean = 5.28). No significant main effect was evident to indicate the effect of font size for perceived enjoyment, so Hypothesis 8 was not supported (perceived enjoyment increases with increasing font size).

5.5.2.2 Hypothesis 2 and 9

No significant main effect for response target size was observed in the repeated-measures ANOVA indicating that response target size did not have an effect on completion time (Hypothesis 2) or perceived enjoyment (Hypothesis 9).

5.5.2.3 Hypothesis 3a and 10a

No significant main effect was observed for font colour in the mixed-measures ANOVA indicating that colour contrast did not have an effect on the completion time (Hypothesis 3) or perceived enjoyment (Hypothesis 10a).

5.6 Limitations

A limitation observed in this study, is with regards to the devices used by the participants. The respondents, interacted with the questionnaires in the experiment using their own mobile devices. Thus the environment was not strictly controlled.

A future recommendation would be to provide devices with exactly the same specifications to achieve standardisation of screen size and screen presentation. Given the sample size of 171 participants, a post hoc power analysis indicated a 100% chance (power = 1 – β = 1.0) of detecting a large and medium effect size (f = 0.40, f = 0.25) and 82% chance (power = 1 – β = 0.82) of detecting a small effect size (f = 0.10). Therefore, the sample size was deemed to be sufficient for the repeated measures ANOVA tests. However, the KMO values in the factor analysis for particular design parameter combinations indicated that the data were not suitable for factor analysis. This could be due to insufficient data collection or insufficient comprehension by the participants of the questions translated into Arabic. However, the rigorous process of translation and back-translation suggests that this is unlikely. Nevertheless, the translation and comprehension of the translated questionnaire could have limited the factor replication of the previously well-established psychometric questionnaires. With regards to sample size, Nunnally (1978) recommended having 10 times as many participants as variables; therefore the required sample size would need to be 10 × 10 variables = 100 for the SUS, 10 × 14 variables = 140 for the disorientation scale, perceived usefulness scale and perceived ease of use scale combined, and 10 × 19 variables = 190 for the PSSUQ. In addition, Tabachnick & Fidell (pg. 613, 2013) agree that 'it is comforting to have at least 300 cases for factor analysis'. However, the number of participants was only 171 indicating insufficient sample size for factor analysis for PSSUQ according to Nunnally and for all the scales according to Tabachnick and Fidell (2013). Therefore, recommendations for future work include increased sample size and potential refinement of the Arabic-language translation of the questionnaires.

Chapter 6

Study 2-A

Contents

Chapter 6: study 2-A

6.1 Overview

Regarding questionnaire design for mobile technologies, many features have been imported directly from the web- and paper-based procedures without considering the potential impact on respondents' experience. In the study of mobile web survey design Peytchev and Hill (2010) stated that it is common for best practices in design to be used from similar existing technologies (e.g. desktop) before rigorous testing is conducted on a new technology (e.g. small-screen devices). The authors report the results for a series of experiments comparing various aspects of questionnaire design and layout, including horizontal scrolling, number of questions per screen, direction of response options, impact of embedded images, and the use of open-ended options using a Samsung Blackjack smartphone. In the context of online psychometrics for both desktop and small-screen devices, currently, there is little research evidence to inform design guidelines for web-based administration of psychometric questionnaires. The few studies that exist conducted by van Schaik and Ling (2003, 2005a, 2005b, 2007) and van Schaik et al. (2015) report results from experiments conducted on desktop computers. This study examines, if design parameters such as font size, text/background polarity and response format affect respondents' completion time of the questionnaires, mental workload and perceived enjoyment on two platforms: mobile devices (current Chapter 6) and desktop computers (see Chapter 7). In addition, the psychometric properties of the questionnaires when these design parameters are manipulated are also investigated. This chapter begins with the statement of the hypotheses, followed by the description of the design and method of the experiment. The analysis sections follow and finally the results and discussion of the Study are presented.

6.2 Outline of Study 2-A experiments

Desktop and mobile versions of web-based questionnaires differ in terms of fundamental human-computer interaction design. The regular web layout that is designed for desktop computers supposes large screens with mouse-handling. Conversely, on small-screen devices like mobiles, the layout is designed for fingertip navigation on touch-screens. Often web-based questionnaires designed for desktop computers are administered on mobile devices, and this results in suboptimal questionnaire presentation (de Bruijne & Wijnant, 2013b). According to van Schaik et al. (2015), the design of online psychometrics has become increasingly important to ensure good measurement properties and future research should be directed at online psychometrics on both small and large displays. By their nature, smartphones have small screens, and this is one of their limitations (Motamedi & Choe, 2015). Because of the difference in screen sizes between desktop computers and mobile devices, a site, when viewed on a desktop computer, will look and behave differently from that same site when viewed on a smartphone. Unlike the traditional way of designing a questionnaire for desktop computers and then administering it on mobile devices, a more appropriate approach will be to formulate a design guided by relevant research and apply it for small-screen devices. Therefore, in Study 2, experiments were designed for small-screen devices (Study 2-A) and desktop computers (Study 2-B see Chapter 7). A new group of participants in each of the two studies responded to psychometric questionnaires and rated the quality of their experience with the system. The design parameters font size, text/background polarity and response format of the questionnaires were manipulated to analyse the effect on the time taken to complete the questionnaires, mental workload and perceived enjoyment.

6.3 Objectives of Study 2-A

Study 2-A was conducted on mobile devices and addresses the following objectives. First, to test the effect of design parameters font size on questionnaire completion time. Second, to test the effect of design parameters text/background polarity on questionnaire completion time. Third, to test and compare the effect of response format (Likert scale using radio button vs Likert scale using visual analogue scale) on questionnaire completion time. In this study, the author reports the experiment investigating three parameters of questionnaire design:

1 font size (44 point vs 64 point);
2 response format (Likert Scale [using radio button] vs Likert scale [using Visual analogue scale]);
3 text/background polarity (black on white vs white on black).

The psychometric properties of questionnaires were also analysed. The current study was conducted with Arabic speakers and, for this study, Arabic versions of the questionnaires were required. As mentioned in Chapter 3, since Arabic version of established questionnaires to measure usability (such as PSSUQ by Lewis, 1995; and SUS by Brooke, 1996), perceived disorientation (Ahuja & Webster, 2001), perceived ease of use, perceived usefulness (Davis et al 1989), perceived enjoyment (Davis 1992), and mental workload (NASA-TLX, Hart & Staveland, 1988) did not exist before this research started, forward-and back translations of the questionnaires were conducted and reviewed by an expert committee (see Section 3.8.2).

6.4 Statement of hypotheses

As described in Section 3.3, the following Hypotheses are presented for Study 2A.

Hypothesis 1: completion time decreases with increasing font size;

Hypothesis 3b: positive text/background polarity decreases completion time;

Hypothesis 4: completion time of Likert scales is shorter with radio button format than visual analogue scale format;

Hypothesis 5: workload decreases with increasing font size;

Hypothesis 6: positive text/background polarity decreases workload;

Hypothesis 7: workload is lower with radio button format than with visual analogue scale format;

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 10b: positive text/background polarity enhances perceived enjoyment;

Hypothesis 11: perceived enjoyment is higher for Likert scale using radio button format than with visual analogue scale format.

6.5 Method

6.5.1 Design

The experiment used a 2×2×(2) experimental design with three measures. The first, second and third independent-measures, variables were:

(1) font size: (44 point [also used in Study 1] or 64 point);

(2) text/background polarity: (black on white [also used in Study 1] or white on black);

(3) response format: (Likert scale using radio button format and Likert scale using visual analogue scale format).

The dependent variables were completion time for the set of five psychometric questionnaires, perceived enjoyment and mental workload.

6.5.1.1 Research design

The experiment was carried out on mobile devices. Eight versions of the questionnaires were created. The experiment was organised into four groups: according the font size, text/background polarity and response format. The response format was the repeated measure. The details are as follows:

Group 1: 44 point, black on white, radio button/visual analogue

Order/version 1: 44 point, black on white, radio button/visual analogue

Order/version 2: 44 point, black on white, visual analogue/radio button

Group 2: 44 point, white on black, radio button/visual analogue

Order/version 3: 44 point, white on black, radio button/visual analogue

Order/version 4: 44 point, white on black, visual analogue/radio button

Group 3: 64 point, black on white, radio button/visual analogue

Order/version 5: 64 point, black on white, radio button/visual analogue

Order/version 6: 64 point, black on white, visual analogue/radio button

Group 4: 64 point, white on black, radio button/visual analogue

Order/version 7: 64 point, white on black, radio button/visual analogue

Order/version 8: 64 point, white on black, visual analogue/radio button

Similar to Study 1 (see Section 5.3.1), the participants in each group responded to five psychometric questionnaires administered in a fixed order (PSSUQ, SUS, DIS, PEU and PU). A brief introduction to these questionnaires is provided in Section 3.4. In Study 2, participants responded to these psychometric questionnaires by rating their experience with the university’s Moodle-based learning management system website. This experiment was a retrospective assessment of the Moodle-based learning management system. Participants were not required to use the system within this experiment. It is to be noted that course materials, class-work, quizzes and assignments are maintained within the learning management system as per the university rules and regulations. Thus, the last time the students used the system will most likely have been one week previously at most. Immediately after the participants completed their responses to the five psychometric questionnaires, they completed the perceived enjoyment (PE) scale developed by Davis (1992); specifically they rated their experience of responding to the psychometric questionnaires they just completed. Following this, the participants responded to the workload questionnaire NASA-TLX (Hart & Staveland, 1998); specifically, they rated their experience of responding to the five psychometric questionnaires. Each participant completed the psychometric questionnaire twice: once with each response format (radio button and visual analogue scale).

6.5.2 Participants for Study 2-A

A total of one hundred ten university students from Kuwait (50 female; 60 male) took part. All the participants were native Arabic-speakers. They were enrolled for 095/096 Math foundation courses (Beginning Algebra and Intermediate Algebra).

6.5.3 Materials and equipment

The experiment was administered through the Online Psychometric Questionnaire Design tool OnPQDT developed for this research purpose (see Chapter 4). All participants who took part in Study 2-A used their own mobile devices. Among the participants 5% used Android based devices such as Samsung: Note 5, S8 and J5 (2017) while 95% of the participants used iOS-based models such as iPhone 6s, iPhone 6s plus, iPhone 7, and iPhone 7plus (see Table 6.1). The specifications of the mobile devices are detailed in Table 6.1.

Table 6.1. Specifications of mobile devices used in the experiment Study 2-A.

	Numbers of each device type	Screen Size (inches) diagonally	Resolution (pixels)	Pixels per inch (ppi)	Dimension (mm)
iPhone 6s	11	4.7	750×1334	326	138.30×67.10×7.10
iPhone 6s plus	23	5.5	1080×1920	401	158.20×77.90×7.30
iPhone 7	42	4.7	750×1334	326	138.30×67.10×7.10
iPhone 7 plus	28	5.5	1080×1920	401	158.20×77.90×7.30
Samsung Note 5	2	5.8	1440×2960	570	148.90×68.10×8.00
Samsung S8	3	5.7	1440×2560	518	153.20×76.10×7.60
Samsung J5	1	5	720×1280	294	142.00×73.00×7.90

6.5.4 Procedure

Research ethics, electronic informed consent, verbal briefing on the reasons for participation prior to the experiment, incentives and voluntary participation, and confidentiality of data collected were highly similar to those in Study 1 (see Section 5.3.4). Data collection took place between December 2017 and April 2018. The sessions were conducted in classrooms within the campus of a private university at all the sessions. The author of this thesis was present, along with the course leader. The sessions were led by the author. Every participant in each session received a URL link from the author.

The URL link was protected by an Access code and the Access code (see Section 4.5.2.4) was provided to the participants to prevent unauthorized access to the questionnaire otherwise. When accessed, the URL link first provided the participants with an online consent form. Thereafter, brief online instructions (see Section 4.5.2.3) in Arabic were provided, on how to complete, the five psychometric questionnaires. At the end of the responses to the five psychometric questionnaires, a brief instruction page was administered online on how to complete the perceived enjoyment scale based on their experience of responding to the psychometric questionnaires. Thereafter, the NASA-TLX workload questionnaire was administered with an online instruction page guiding the participants to respond to the workload questionnaire, based on their experience than responding to the psychometric questionnaires. The experiment automatically repeated with the second response format (see Section 6.5.1). The font size and the text/background polarity remained uniform within the groups assigned for the sessions. At the end of the experiment, a 'Thank you' page was presented to every participant in appreciation for their time and involvement. The author, along with the module leader, was present till all the students completed the experiment that lasted for about 35 minutes. The data collected for Study 2-A were analysed using SPSS version 23 and the output will be provided by the author upon request.

6.6. Analysis of Study 2-A

6.6.1 Reliability

Figures 6.1, 6.2, 6.3 and 6.4 present the reliability coefficients for all five psychometric scales (DIS, PEU, PU, PSSUQ and SUS). The clustered bar graph depicts the response format (Likert scale using visual analogue scale and Likert scale using radio button) for all questionnaires according to font size and text/background polarity. All the scales were reliable, consistent with the English versions except for the perceived ease of use (PEU) construct. The low value of Cronbach's alpha α varied between 0.51 and 0.57 for certain combinations of design parameters as follows:

(1) visual analogue scale format:

(α = 0.57) font size: 44 point; polarity: black on white;

(α = 0.55) font size: 64 point; polarity: white on black

(2) radio button format:

(α = 0.51) font size: 64 point; polarity: black on white;

The reliability analysis for the perceived enjoyment (PE) construct is presented in Figure 6.5. The clustered bar graph depicts the response format (Likert scale using visual analogue scale and Likert scale using radio button) for all combinations of font size and polarity. The scale possessed high reliability, consistent with the English version for all combinations of the design parameters (font size, polarity and response format).

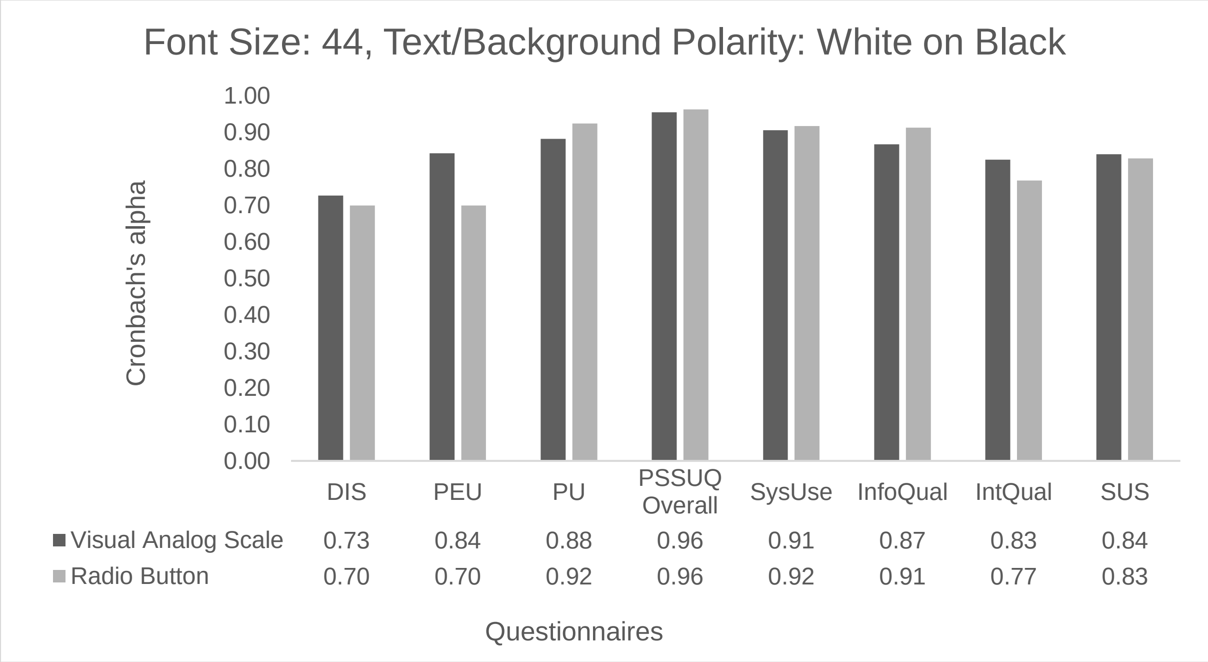


Figure 6.1. Reliability analysis Study 2-A (44 point; white on black).

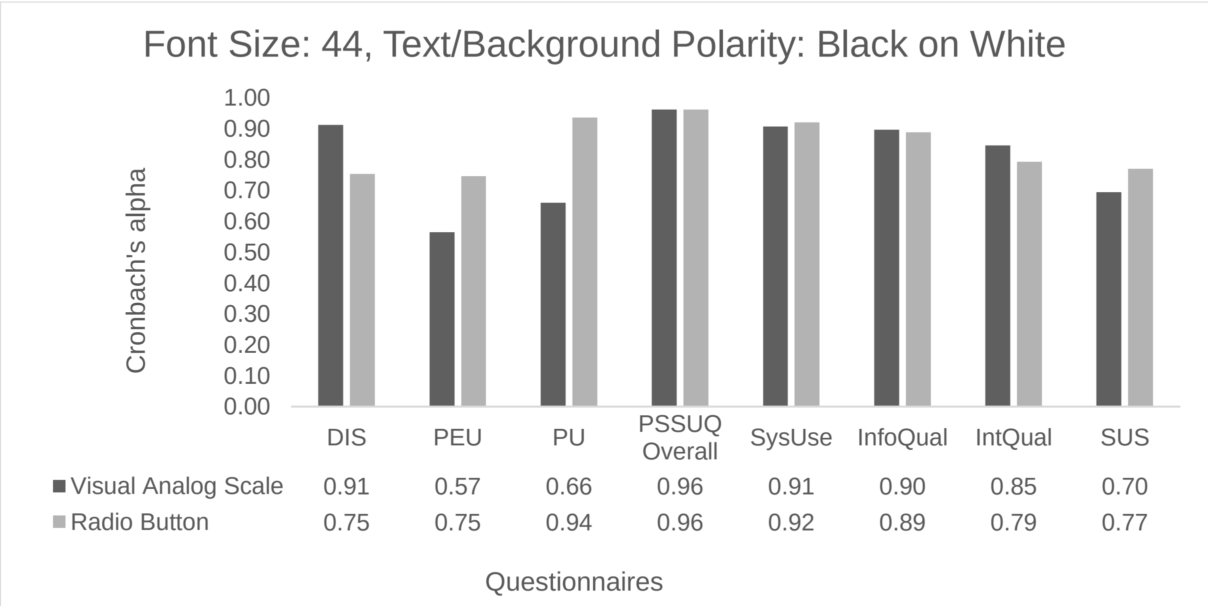


Figure 6.2. Reliability analysis Study 2-A (44 point; black on white).

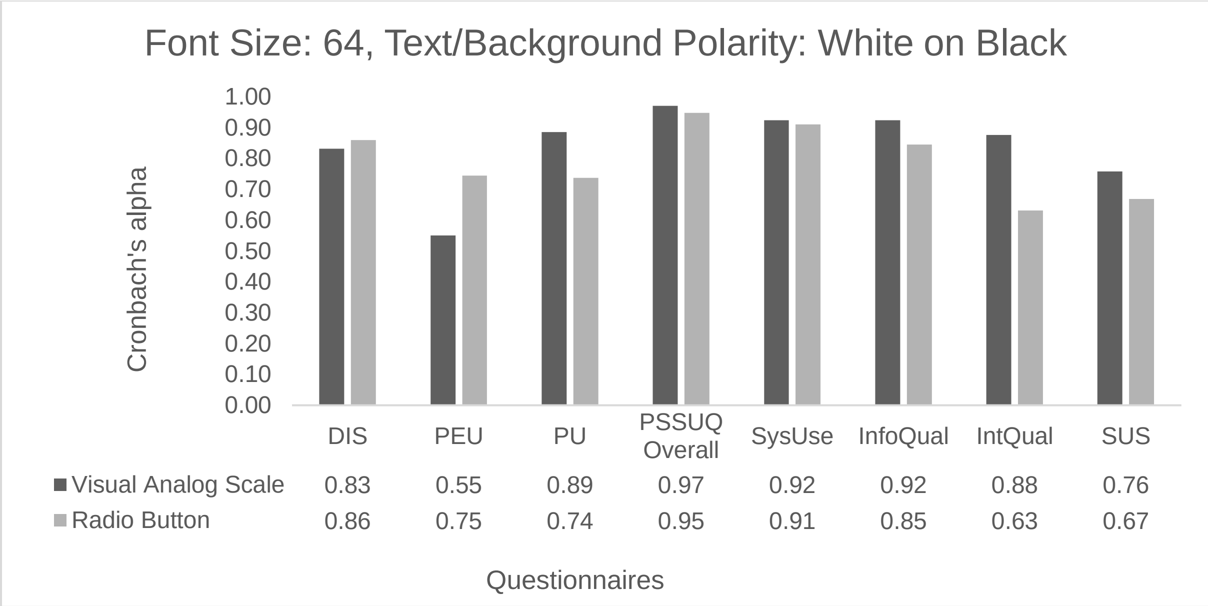


Figure 6.3. Reliability analysis Study 2-A (64 point; white on black).

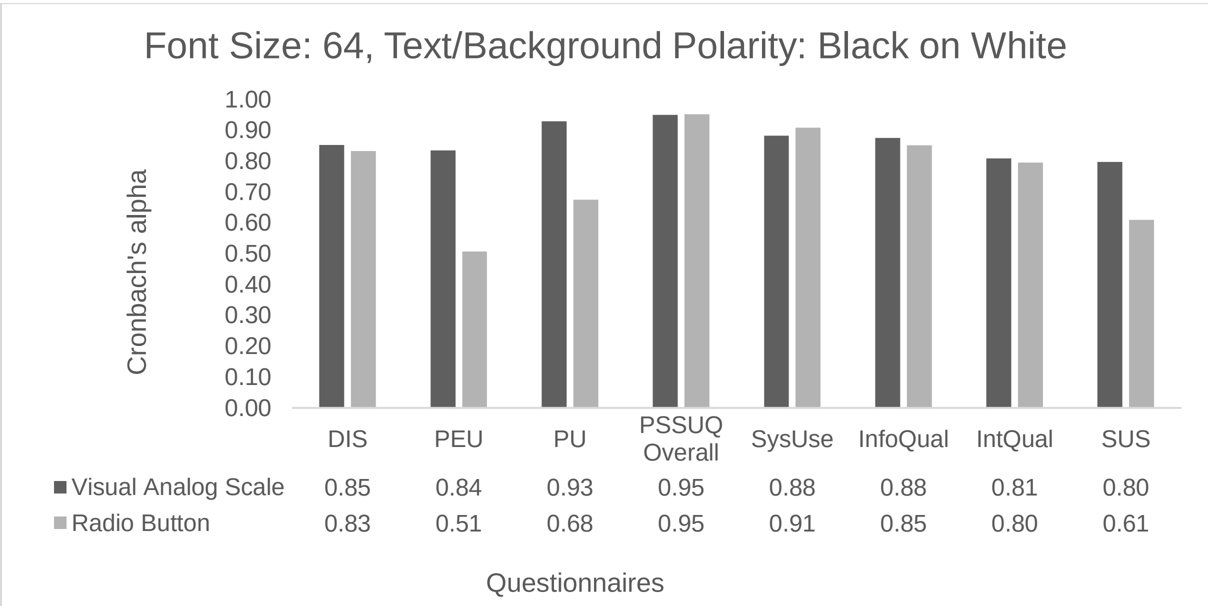


Figure 6.4. Reliability analysis Study 2-A (64 point; black on white).

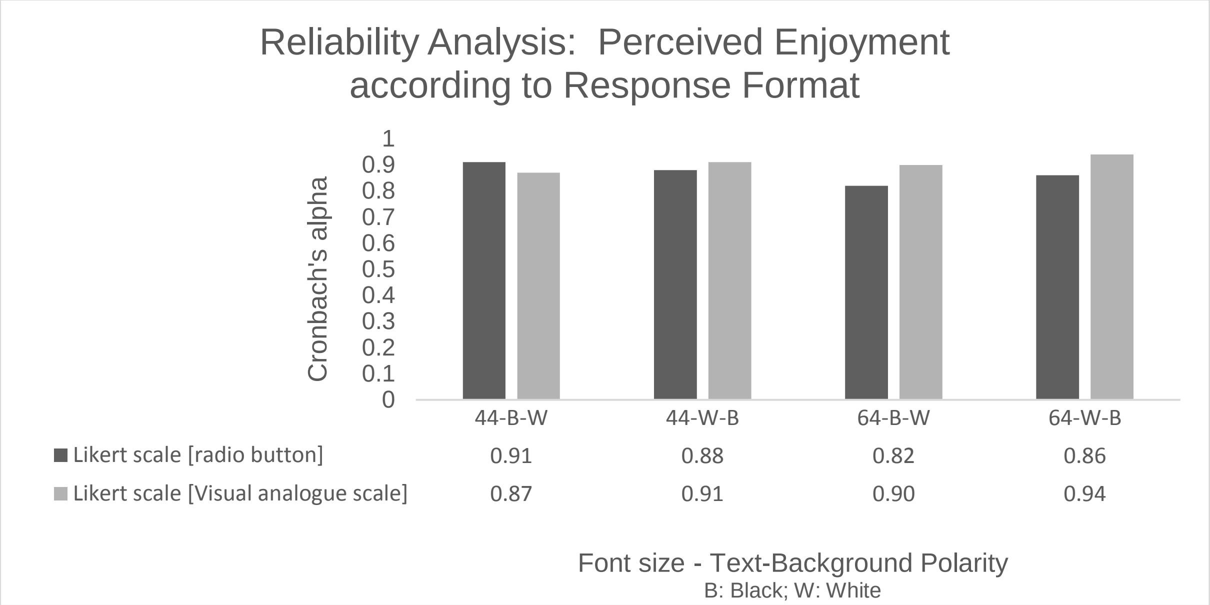


Figure 6.5. Reliability analysis of perceived enjoyment (Study 2-A).

6.6.1.1 Summary of Reliability analysis

In summary, the reliability for all the psychometric questionnaires (DIS, PU, PSSUQ, SUS and PE) was good, with the exception of PEU for certain combinations of design parameters (see Section 6.6.1).

6.6.2 Validity

Pearson’s correlation coefficient is used to assess criterion-related between the constructs disorientation, perceived ease of use, perceived usefulness, PSSUQ, the subscales of PSSUQ: SysUse, InfoQual, IntQual and SUS. The results are reported according to response format (Tables 6.2 and 6.3).

Table 6.2. Correlations between constructs (response format: visual analogue scale format).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	-.07	-.02	-.02	-.03	-.04	.01	-.05
PU		.18	-.22*	-.21*	-.19	-.12	.10
PEU			.02	.03	.03	.12	.15
PSSUQ				.96**	.96**	.90**	-.05
SYS					.87**	.82**	-.06
INFO						.85**	-.04
INT							-.02

PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 6.3. Correlations between constructs (response format: radio button format).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	.08	-.24*	-.03	.02	.00	-.02	-.08
PU		.03	-.04	.01	-.02	-.04	-.26**
PEU			-.03	-.01	-.09	-.08	.00
PSSUQ				.93**	.94**	.90**	.07
SYS					.79**	.81**	-.05
INFO						.86**	.16
INT							.09

PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual

* Correlation is significant at the 0.05 level (2-tailed).

6.6.2.1 Response format: visual analogue scale

The correlation between the questionnaires for the response format: visual analogue scale is presented in Table 6.2. A significantly strong correlation was observed between the subscales SysUse and InfoQual $r = .87$; SysUse and IntQual $r = .82$; InfoQual and IntQual $r = .85$; all $p < .01$.

6.6.2.2 Response format: radio button

The correlation between the questionnaires for the response format: radio button is presented in Table 6.3. A significantly strong correlation was observed between the subscales SysUse and InfoQual $r = .79$; SysUse and IntQual $r = .81$; InfoQual and IntQual $r = .86$; all $p < .01$.

6.6.2.3 Summary of validity

In summary, for the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS), a high positive correlation $r > .50$ $p < 0.01$ was observed between the PSSUQ overall and its subscales: SysUse, InfoQual and IntQual across both the response formats. This observation is consistent with the previous research by Lewis (1995; 2002).

6.6.3 Factor structure of PSSUQ

Based on literature studies (Lewis, 1995, 2002), the factor structure of PSSUQ consisted of three subscales SysUse, InfoQual and IntQual. Close observation of the correlation values between the PSSUQ items, revealed moderately strong values within the PSSUQ subscales ranging from 0.61 to 0.71 and slightly lower correlation values 0.37 to 0.50 between the PSSUQ subscales.

This pattern of correlation provides some initial evidence for the factors of PSSUQ, with higher correlations among items within the subscales and lower correlations among items between the subscales. Previous research (Lewis, 1995, 2002) reported principal axis factoring as the extraction method, with varimax rotation, in order to determine the factor structure of PSSUQ. Therefore, principle axis extraction technique with oblimin rotation was conducted to assess the factor structure for the 19 items of the usability questionnaire PSSUQ. However, no result was obtained, because SPSS reported the problem of a non-positive definite matrix. The unweighted least squares extraction method with oblimin rotation produced results and was then applied for all combinations of the design parameters (font size and polarity). Factor loadings < 0.3 were suppressed (Field, 2013) and three factors were explicitly requested. According to Kaiser (1974), the KMO value and Bartlett's test of sphericity values reported in Table 6.4, indicated the suitability of the data for structure detection for all design parameters except for the following combinations of design parameters:

(1) visual analogue scale format:

(*KMO* = 0.38; .32) font size: 44 point; polarity: black on white; white on black;

(2) radio button format:

(*KMO* = 0.24; .33) font size: 44 point; polarity: black on white; white on black;

Table 6.4. KMO and Bartlett's test of sphericity values (PSSUQ).

Font size, text/background polarity	Visual analogue scale	Radio button format
44 point, black on white	*KMO* = .38; Bartlett: $\chi^2(110) = 328.07$	*KMO* = .32; Bartlett: $\chi^2(110) = 354.57$
44 point, white on black	*KMO* = .24; Bartlett: $\chi^2(110) = 343.54$	*KMO* = .33; Bartlett: $\chi^2(110) = 277.12$
64 point, black on white	*KMO* = .70; Bartlett: $\chi^2(110) = 526.44$	*KMO* = .69; Bartlett: $\chi^2(110) = 569.31$
64 point, white on black	*KMO* = .75; Bartlett: $\chi^2(110) = 754.202$	*KMO* = .74; Bartlett: $\chi^2(110) = 427.83$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research, factor analysis was still conducted for all combinations of design parameters and the results are reported in Tables 6.5, 6.6, 6.7 and 6.8 according to the combinations of the design parameters font size and polarity. The results are compared with the established results in research studies by Lewis (1995, 2002) (see Table 5.6, Section 5.4.3).

Table 6.5. Factor analysis of PSSUQ (44 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.93			.82	-.44	
Q2	.81	-.33		.76	-.57	
Q3	.76			.74	.40	
Q4	.43	.50	.77	.56		
Q5	.77			.75		
Q6	.87			.81		
Q7	.82	-.47		.92		
Q8	.68	.42		.78		.41
Q9	.69			.50	.41	
Q10	.65			.69		.61
Q11	.58	.53		.70	.54	
Q12	.80	.38	-.36	.87	-.45	
Q13	.81			.84		-.32
Q14	.88			.79	.42	-.31
Q15	.91			.86		-.33
Q16	.75			.68		
Q17	.83		.35	.91	-.33	
Q18	.87			.78	.50	
Q19	.81			.75		-.44
Eigenvalues	11.57	1.68	1.15	11.25	2.12	1.27
% Variance	60.88	8.84	6.05	59.19	11.15	6.70
No. of items	18	0	1	19	0	0

Note: factor loadings < 0.3 are suppressed.

Table 6.6. Factor analysis of PSSUQ (44 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.83			.74		
Q2	.75		-.32	.77	-.47	
Q3	.73	.45		.85		
Q4	.67		-.57	.67		
Q5	.55	.42		.82	.39	
Q6	.77			.83	-.34	
Q7	.78			.67		
Q8	.79	.45		.76		-.35
Q9	.89			.78		
Q10	.71	-.41		.66		
Q11	.36	.47		.71	.50	
Q12	.82			.90	.31	
Q13	.65	-.43		.81		
Q14	.63	-.56	.32	.81		-.32
Q15	.74	.32	.51	.73		.50
Q16	.77	.33		.73	-.35	.52
Q17	.84	-.38		.90	-.31	
Q18	.75	-.31		.72	.36	
Q19	.90			.75		
Eigenvalues	10.50	2.12	1.16	11.32	1.47	1.26
% Variance	55.24	11.15	6.12	59.56	7.73	6.62
No. of items	11	1	0	19	0	0

Note: factor loadings < 0.3 were suppressed.

Table 6.7. Factor analysis of PSSUQ (64 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.66	.33		.79		
Q2	.77		.47	.85		
Q3	.69			.82		
Q4	.67	-.35		.67		
Q5	.54			.78		
Q6	.74			.82		.32
Q7	.77	-.44		.79	-.38	
Q8	.78			.47		
Q9	.77				.51	
Q10	.65		.50	.82		
Q11	.70	.50	.33	.63		
Q12	.81			.79		
Q13	.62	.56		.61		.58
Q14	.75			.83		
Q15	.69			.79		
Q16	.79			.76		
Q17	.66	-.30		.67		
Q18	.81			.89		
Q19[b]	.84			.78		
Eigenvalues	9.99	1.04	.94	10.48	.87	.86
% Variance	52.58	7.58	4.96	55.14	4.60	4.53
No. of items	19	0	0	18	1	0

Note: factor loadings < 0.3 were suppressed.

Table 6.8. Factor analysis of PSSUQ (64 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.80	.39		.85	-.31	
Q2	.80		.38	.81		
Q3	.83			.74	-.35	
Q4	.51	.53		.48		
Q5	.83		.38	.70		
Q6	.89			.72		
Q7	.78	-.34	.32	.80		
Q8	.78			.84		
Q9	.87			.54	.38	
Q10	.64			.62		
Q11	.81			.71	.47	
Q12	.67			.63	-.31	-.32
Q13	.85			.76	.34	
Q14	.88	-.30		.69	.37	
Q15	.85			.75		
Q16[b]	.89			.53	.41	-.48
Q17	.88	.34		.70		.44
Q18	.82			.80	.42	
Q19	.91			.81		-.30
Eigenvalues	12.49	1.04	.74	9.76	1.66	1.15
% Variance	65.70	5.46	3.88	51.37	8.76	6.05
No. of items	19	0	0	19	0	0

Note: factor loadings < 0.3 are suppressed.

6.6.3.1 Summary of the factor structure of PSSUQ

In summary, across all the design parameter combinations, the original factor structure of PSSUQ reported in research studies (Lewis, 1995 and 2002) could not be replicated. There was no pattern that could be identified and hence, the factors could not be named.

6.6.4 Factor structure of SUS

The latest research reported by Lewis and Sauro (2017) presented a tone model for the factor structure of SUS, although it initially displayed a unidimensional structure (Brooke, 1996) (see Section 3.6). For the purpose of validating the SUS construct administered in this research study, the correlation values were examined. Within the items of each of the two SUS subscales (positive and negative), a fairly moderate correlation with values between 0.34 and 0.54 were observed. However, the items between the two subscales displayed very low correlations values between 0.05 and 0.44. This pattern of correlation provides some initial evidence for the factors of SUS, with higher correlations among items within the subscales and lower correlations among items between the subscales. Previous research by Lewis and Sauro (2017) reported the use of all three extraction techniques such as principle components analysis (strictly not a factor analytic method, but commonly used), unweighted least squares and maximum likelihood extraction technique with varimax rotation. In the current research, principle axis factoring with oblimin rotation was initially used for all combinations of design parameters (font size and text/background polarity) and both response formats (radio button and visual analogue scale) to detect the factor structure. However, a solution could not be obtained consistently across all combinations of the design parameters and response formats due to the error, communality of a variable (not reported in SPSS version 23) that exceeded 1.0. Therefore, two methods with different rotations had to be used for the two response formats in particular:

(1) visual analogue scale: unweighted least squares with oblimin rotation;

(2) radio button: maximum likelihood extraction with varimax rotation.

Factor loadings < 0.3 were suppressed. According to Kaiser (1974), the KMO value and Bartlett's test of sphericity values reported in Table 6.9, indicated the suitability of the data for structure detection for all design parameters except for the following combination of design parameters:

(1) visual analogue scale format:

(*KMO* = 0.32) font size: 44 point; polarity: black on white;

(2) radio button format:

(*KMO* = 0.43) font size: 44 point; polarity: black on white;

(*KMO* = 0.49) font size: 44 point; polarity: white on black;

(*KMO* = 0.49) font size: 64 point; polarity: white on black.

Table 6.9. KMO and Bartlett's test of sphericity values (SUS)-Study 2-A.

Font size, text/background polarity	Visual analogue scale	Radio button format
44 point, black on white	*KMO* = .32; Bartlett: $\chi^2(45) = 111.60$	*KMO* = .43; Bartlett: $\chi^2(45) = 104.49$
44 point, white on black	*KMO* = .51; Bartlett: $\chi^2(45) = 109.91$	*KMO* = .49; Bartlett: $\chi^2(45) = 105.70$
64 point, black on white	*KMO* = .59; Bartlett: $\chi^2(45) = 162.94$	*KMO* = .72; Bartlett: $\chi^2(45) = 93.92$
64 point, white on black	*KMO* = .71; Bartlett: $\chi^2(45) = 155.85$	*KMO* = .49; Bartlett: $\chi^2(45) = 116.03$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research, factor analysis was carried out for all combinations of design parameters across both the response formats and the results are reported in Tables 6.10, 6.11, 6.12 and 6.13. Thus the results of the current research, were compared with the established results (see Table 5.12, Section 5.4.4) and reported.

Table 6.10. Factor analysis of SUS (44 point; black on white).

	Visual analogue scale		Radio button	
	Factors[a]		Factors[b]	
Items	1	2	1	2
Q1		0.94	0.67	
Q2	0.53		0.46	
Q3		0.77		
Q4	0.81			0.84
Q5	0.48	-0.54	0.98	
Q6	0.56	0.34	0.63	0.57
Q7				0.40
Q8	0.94		0.31	0.82
Q9	0.49	0.36		0.41
Q10		0.60	0.77	
Eigenvalues	3.07	2.83	3.95	1.76
% Variance	26.85	24.51	27.93	22.22
No. of items	6	3	5	4

Note: factor loadings < 0.3 were suppressed.

[a] unweighted least squares with oblimin rotation.

[b] maximum likelihood extraction with varimax rotation.

Table 6.11. Factor analysis of SUS (44 point; white on black).

Items	Visual analogue scale		Radio button	
	Factors[a]		Factors[b]	
	1	2	1	2
Q1				0.62
Q2	0.56			
Q3		0.63	0.78	0.63
Q4	0.76		0.90	
Q5	0.86		0.62	
Q6	0.66	0.34	0.87	-0.47
Q7	0.75		0.56	0.33
Q8	0.83		0.75	
Q9		0.93	0.59	
Q10	0.47	0.33		0.77
Eigenvalues	4.52	1.62	4.28	2.41
% Variance	41.6	12.47	38.24	19.45
No. of items	7	3	7	2

Note: factor loadings < 0.3 were suppressed.

[a] unweighted least squares with oblimin rotation.

[b] maximum likelihood extraction with varimax rotation.

Table 6.12. Factor analysis of SUS (64 point; black on white).

Items	Visual analogue scale		Radio button	
	Factors[a]		Factors[b]	
	1	2	Negative	Positive
Q1		0.68		0.57
Q2	0.78		0.80	
Q3		0.89	-0.35	0.81
Q4	0.63		0.86	
Q5	0.74			0.56
Q6	0.58		0.62	
Q7	-0.32	0.34		0.53
Q8	0.95		0.56	
Q9		0.82		0.51
Q10			0.60	
Eigenvalues	3.91	2.14	3.28	2.19
% Variance	35.32	17.51	25.85	18.76
No. of items	5	4	5	5

Note: factor loadings < 0.3 were suppressed.

[a] unweighted least squares with oblimin rotation.

[b] maximum likelihood extraction with varimax rotation.

Table 6.13. Factor analysis of SUS (64 point; white on black).

	Visual analogue scale		Radio button	
	Factors[a]		Factors[b]	
Items	1	2	1	2
Q1		0.49	0.75	
Q2	0.64			1.00
Q3		1.01	0.99	
Q4	0.70			0.52
Q5	0.81			
Q6	0.61			
Q7	0.43		0.76	
Q8	0.75			0.65
Q9		0.62		
Q10	0.64		0.39	
Eigenvalues	3.66	2.08	2.91	2.1
% Variance	31.68	17.46	23.48	18.06
No. of items	7	3	4	3

Note: factor loadings < 0.3 were suppressed.

[a] unweighted least squares with oblimin rotation.

[b] maximum likelihood extraction with varimax rotation.

6.6.4.1 Summary of the factor structure of the SUS

Considering the factor structure of the SUS in this study, across all design parameters, the tone model reported by Lewis and Sauro (2015, 2017) was evident only for one combination of design parameter response format radio button, font size 64 point, polarity black on white. For all other instances, the factor structure established previously in various studies (e.g. Lewis & Sauro, 2015, 2017) could not be replicated.

6.6.5 Factor structure of DIS, PEU and PU

With regards to the three questionnaires disorientation, perceived ease of use and perceived usefulness, research by van Schaik and Ling (2003, 2005, and 2007), presented a three-factor solution: (1) disorientation, (2) perceived ease of use and (3) perceived usefulness (see Table 5.18, Section 5.4.5). In this study, results of factor analysis were obtained, reported and compared with the previous studies by van Schaik and Ling. For the purpose of validating the DIS, PEU and PU constructs administered in this research study, the correlation values were examined. Within the items of each of the scales, a moderate correlation was observed with values ranging from 0.22 to 0.54 for the DIS construct, 0.29 to 0.59 for the PEU construct and 0.46 to 0.64 for the PU construct. Low correlation values were observed between the items of the DIS, PEU and PU constructs while a moderate correlation existed between the items of the PEU and PU constructs. This pattern of correlation provides some initial evidence for the factors of DIS, PEU and PU, with higher correlations among items within the subscales and lower correlations among items between the subscales. Similar to previous research reported by van Schaik and Ling (2003, 2005, and 2007), principle axis factoring with oblimin rotation was initially employed for all combinations of design parameters (font size and text/background polarity) and both response formats (radio button and visual analogue scale) to detect the factor structure of the three questionnaires together. However, a solution could not be obtained consistently across all combinations of the design parameters and response formats due to the error, communality of a variable (not reported in SPSS version 23) that exceeded 1.0. Therefore, two methods had to be used for the two response formats as follows:

(1) visual analogue scale: maximum likelihood extraction with oblimin rotation;

(2) radio button: unweighted least squares with oblimin rotation.

Factor loadings < 0.3 were suppressed. According to Kaiser (1974), the KMO value and Bartlett's test of sphericity values reported in Table 6.14, indicated the suitability of the data for structure detection for all design parameters except for the following combination of design parameters:

(1) visual analogue scale format:

(*KMO* = 0.44) font size: 44 point; polarity: black on white;

(*KMO* = 0.11) font size: 44 point; polarity: white on black;

(2) radio button format:

(*KMO* = 0.38) font size: 44 point; polarity: black on white;

(*KMO* = 0.14) font size: 44 point; polarity: white on black;

Table 6.14. KMO and Bartlett's test of sphericity values (DIS, PEU and PU) Study 2-A.

Font size, text/background polarity	Visual analogue scale	Radio button format
44 point, black on white	*KMO* = .44; Bartlett: $\chi^2(91) = 209.91$	*KMO* = .38; Bartlett: $\chi^2(91) = 201.55$
44 point, white on black	*KMO* = .11; Bartlett: $\chi^2(91) = 161.61$	*KMO* = .14; Bartlett: $\chi^2(91) = 295.14$
64 point, black on white	*KMO* = .62; Bartlett: $\chi^2(91) = 312.59$	*KMO* = .53; Bartlett: $\chi^2(91) = 190.33$
64 point, white on black	*KMO* = .66; Bartlett: $\chi^2(91) = 230.39$	*KMO* = .65; Bartlett: $\chi^2(91) = 230.05$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research (see Table 5.18, Section 5.4.5), factor analysis was carried out for all combinations of design parameters across both the response formats and the results are reported in Tables 6.15, 6.16, 6.17 and 6.18.

Table 6.15. Factor analysis of DIS, PEU and PU (44 point; black on white).

	Visual analogue scale			Radio button format		
	Factors[a]			Factors[b]		
Items	1	2	PU	PU	DIS	PEU
DIS_Q1	.91				.83	
DIS_Q2	.86				.55	
DIS_Q3	.71			-.42	.59	
DIS_Q4	.67		.30	.44	.60	
DIS_Q5	.70		-.34			
DIS_Q6	.62	-.71			.95	.33
DIS_Q7	.80		.31		.46	
PEU_Q1	-.54					.51
PEU_Q2		.35				.57
PEU_Q3	-.36					.96
PU_Q1			.75	.91		
PU_Q2		.43	.64	.86		
PU_Q3		.67		.75		
PU_Q4			.48	.93		
Eigenvalues	5.52	2.33	1.44	4.69	2.78	1.98
% Variance	31.46	15.46	11.06	31.69	17.39	12.15
No. of items	8	3	3	4	7	3

Note: factor loadings < 0.3 are suppressed.

[a] maximum likelihood extraction with oblimin rotation.

[b] unweighted least squares extraction with oblimin rotation.

Table 6.16. Factor analysis of DIS, PEU and PU (44 point; white on black).

	Visual analogue scale			Radio button format		
	Factors[a]			Factors[b]		
Items	PEU	2	PU	1	2	3
DIS_Q1		.68		.31		.76
DIS_Q2		.80		.33	.32	.60
DIS_Q3		.67				.61
DIS_Q4		.41	.52		.64	.43
DIS_Q5		.99			.40	.55
DIS_Q6				-.62	.33	
DIS_Q7	-.50					.35
PEU_Q1	-1.00				-.83	
PEU_Q2	-.65			.53	-.44	
PEU_Q3	-.79				-.69	
PU_Q1			.92	.77		
PU_Q2	-.38		.66	.64	.72	
PU_Q3			.91	.96		
PU_Q4			.82	.97		
Eigenvalues	3.76	2.92	2.76	4.15	3.46	2.08
% Variance	18.41	20.07	22.40	28.23	21.93	11.73
No. of items	4	5	4	4	5	5

Note: factor loadings < 0.3 are suppressed.

[a] maximum likelihood extraction with oblimin rotation.

[b] unweighted least squares extraction with oblimin rotation.

Table 6.17. Factor analysis of DIS, PEU and PU (64 point; black on white).

	Visual analogue scale			Radio button format		
	Factors[a]			Factors[b]		
Items	PU	PEU	DIS	DIS	PU	3
DIS_Q1			.43	.86		
DIS_Q2			.78	.41		
DIS_Q3		-.34	.63	.49		
DIS_Q4			.54	.45		
DIS_Q5			.86	.81		
DIS_Q6			.74	.92		
DIS_Q7			.76	.52	.37	-.32
PEU_Q1		.78				1.01
PEU_Q2		.62			.47	
PEU_Q3		.95				.44
PU_Q1	.73				.83	
PU_Q2	.93				.38	
PU_Q3	.90				.72	
PU_Q4	.96					
Eigenvalues	3.85	3.43	2.51	3.73	2.36	1.71
% Variance	23.05	15.72	23.75	23.77	13.73	9.91
No. of items	4	3	7	7	4	2

Note: factor loadings < 0.3 are suppressed.

[a] maximum likelihood extraction with oblimin rotation.

[b] unweighted least squares extraction with oblimin rotation.

Table 6.18. Factor analysis of DIS, PEU and PU (64 point; white on black).

	Visual analogue scale			Radio button format		
	Factors[a]			Factors[b]		
Items	1	DIS	PU	DIS	2	PEU
DIS_Q1		.31		.68	-.46	
DIS_Q2		.79		.74		
DIS_Q3		.64		.47		
DIS_Q4		.58		.74		
DIS_Q5		.66		.75		
DIS_Q6		.68		.51		
DIS_Q7	.33	.79		.84		
PEU_Q1	.90		.33			.77
PEU_Q2	.33		.40			.47
PEU_Q3	.55		-.32			.98
PU_Q1			.70	.50		
PU_Q2			.84		.82	
PU_Q3			.88		.75	
PU_Q4		-.30	.70		.77	
Eigenvalues	4.65	2.70	1.72	4.38	2.51	2.10
% Variance	14.91	29.47	12.60	28.47	15.17	12.55
No. of items	2	7	5	8	3	3

Note: factor loadings < 0.3 are suppressed.

[a] maximum likelihood extraction with oblimin rotation.

[b] unweighted least squares extraction with oblimin rotation.

6.6.5.1 Summary of the factor structure of DIS, PEU, PU

In summary, considering the factor structure of DIS, PEU and PU the following response format and design parameter combinations exhibited a clear three-factor solution similar to previous studies by van Schaik and Ling (2003, 2005, and 2007) and Ahuja and Webster (2001).

(1) Clear pattern of DIS, PEU and PU as three distinct factors:

(DIS, PEU and PU) visual analogue scale; 64 point, black on white;

(DIS, PEU and PU) radio button format; 44 point, black on white.

(2) At least one or more distinct factors with either DIS, PEU or PU:

(PEU and PU): visual analogue scale; 44 point; white on black;

(DIS and PU): visual analogue scale; 64 point; white on black;

(DIS and PU): radio button; 64 point; black on white;

(DIS and PEU): radio button; 64 point; white on black;

(PU): visual analogue scale; 44 point; black on white;

It should be noted that research by van Schaik and Ling (2003, 2005, and 2007) was done on desktop computers, while in this research, Study 2-A was conducted on mobile phones. The results of van Schaik and Ling were replicated only for certain combinations as reported earlier.

6.6.6 Factor structure of PE

Unweighted least squares factoring method was employed to determine the factor structure of the perceived enjoyment scale. In order to compare the results, factor analysis from the study by Davis (1992) is provided as a reference (see Section 5.4.6 Table 5.24). The KMO and Bartlett's test of sphericity values indicated the suitability of data for factor extraction and is reported in Table 6.19.

Table 6.19. KMO and Bartlett's test of sphericity values for PE Study 2-A.

Font size, text/background polarity	Visual analogue scale	Radio button format
44 point, black on white	*KMO* = .73; Bartlett: $\chi^2(91) = 25.69$	*KMO* = .60; Bartlett: $\chi^2(91) = 54.60$
44 point, white on black	*KMO* = .71; Bartlett: $\chi^2(91) = 39.08$	*KMO* = .65; Bartlett: $\chi^2(91) = 30.81$
64 point, black on white	*KMO* = .70; Bartlett: $\chi^2(91) = 89.75$	*KMO* = .66; Bartlett: $\chi^2(91) = 40.24$
64 point, white on black	*KMO* = .73; Bartlett: $\chi^2(91) = 63.62$	*KMO* = .63; Bartlett: $\chi^2(91) = 54.10$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

The factor analysis results are reported in Tables 6.20 and 6.21 according to the response formats (Likert scale using radio button and visual analogue scale) for all design parameters: font size and text/background polarity.

Table 6.20. Factor analysis of PE (response format: radio button format).

	44 point		64 point	
Items	Black on white	White on black	Black on white	White on black
Q1	0.92	1.00	0.61	0.82
Q2	1.01	0.85	0.96	1.01
Q3	0.71	0.70	0.81	0.64
Eigenvalues	2.54	2.44	2.24	2.34
% Variance	78.73	73.85	64.63	69.57
No. of items	3	3	3	3

Table 6.21. Factor analysis of PE (response format: visual analogue scale format).

	44 point		64 point	
Items	Black on white	White on black	Black on white	White on black
Q1	0.79	0.76	0.88	0.94
Q2	0.88	0.93	0.87	0.78
Q3	0.82	0.96	1.00	0.88
Eigenvalues	2.38	2.56	2.68	2.50
% Variance	69.22	78.96	84.60	75.70
No. of items	3	3	3	3

6.6.6.1 Summary of the factor structure of PE

In summary, the factor structure of the perceived enjoyment scale across all design parameters, was consistent and similar to the structure reported in other studies (e.g. Davis 1992).

6.6.7 Mixed ANOVA

A mixed-measures ANOVA was conducted to test the main and interaction effects of font size, text/background polarity and response format on the total completion time of the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS). In addition, mixed-measures ANOVA was conducted to test the main and interaction effects of font size, text/background polarity and response format on the perceived enjoyment and the workload (NASA-TLX) that respondents rated, immediately after their response to the psychometric questionnaires.

6.6.7.1 Time to complete the questionnaires

Total completion time was heavily positively skewed. An inverse transformation of the total completion time for the five questionnaires (DIS, PEU, PU, PSSUQ and SUS) reduced the skew and improved normality of distribution. Table 6.22 details the mean, standard deviation and the confidence intervals of the mean for the original, transformed and the retransformed time in seconds. From Table 6.22, we see that font size 44 point had faster completion time when the response format was radio button.

Table 6.22. Descriptives for completion time of questionnaires.

		Original (time[a], all items)				Transformed (log(time))				Retransformed (exp[log(time)])			
Response format		RB		VAS		RB		VAS		RB		VAS	
Font size		44	64	44	64	44	64	44	64	44	64	44	64
mean		185.90	243.88	204.90	207.62	4.98	5.37	5.24	5.29	185.90	243.88	204.90	207.62
Std. Deviation		112.41	108.56	84.46	61.21	0.82	0.57	0.41	0.32	112.41	108.56	84.46	61.21
95% Confidence Interval for mean	Lower Bound	150.88	217.61	178.59	192.80	4.73	5.24	5.11	5.21	150.88	217.61	178.59	192.80
	Upper Bound	220.93	270.16	231.22	222.43	5.24	5.51	5.37	5.38	220.93	270.16	231.22	222.43

Note: RB: Likert scale using radio button; VAS: visual analogue scale.

[a] Seconds.

A mixed 2×2×(2) ANOVA (Table 6.23) demonstrated that the interaction effect of response format and font size was significant, F (1, 106) = 6.70, $\eta_p^2 = 0.06$, $p < 0.05$.

Table 6.23. Mixed ANOVA summary table for Study 2-A.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity	1	1.10	1.10	2.89	0.09	0.03
Font size	1	2.56	2.56	6.70	0.01	0.06
Polarity × Font size	1	0.00	0.00	0.00	0.98	0.00
Error (Font size and Polarity)	106	40.46	0.38			
Response format	1	0.12	0.12	0.62	0.43	0.01
Response format × Polarity	1	0.70	0.70	3.69	0.06	0.03
Response format × Font size	1	1.56	1.56	8.21	0.01	0.07
Response format × Polarity × Font size	1	0.12	0.12	0.64	0.43	0.01
Error (Response format)	106	20.08	0.19			

Further univariate 2×2 ANOVAs (Tables 6.24, 6.25) for each response format (radio button format and visual analogue scale) showed that, the main effect of font size F (1, 106) = 9.16, $\eta_p^2 = 0.08$, $p < 0.01$ was significant for the response format radio button (see Table 6.24).

Table 6.24. Univariate ANOVA summary table (response format: radio button).

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity	1	1.54	1.54	3.47	0.07	0.03
Font size	1	4.06	4.06	9.16	0.00	0.08
Polarity × Font size	1	0.05	0.05	0.12	0.73	0.00
Error (Font size and Polarity)	106	47.02	0.44			

Table 6.25. Univariate ANOVA summary table (response format: visual analogue scale)

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity	1	0.04	0.043	0.338	0.562	0.003
Font size	1	0.06	0.061	0.475	0.492	0.004
Polarity × Font size	1	0.066	0.066	0.518	0.473	0.005
Error (Font size and Polarity)	106	13.527	0.128			

Furthermore, simple-effects test (Tables 6.26, 6.27) showed that main effect of response format was significant for font sizes 44 point, with faster responses when radio buttons were used. The interaction effect of response format and polarity was significant for font size 64 point, F (1, 66) = 5.97, $\eta_p^2 = 0.08$, $p < 0.05$ (Table 6.27) and the main effect of response format was significant for font size 44 point, F (1, 40) = 4.79, $\eta_p^2 = 0.11$, $p < 0.05$ (Table 6.27).

Table 6.26. Simple effect analysis font size 44 point.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Response format	1	1.40	1.40	4.79	0.03	0.11
Polarity	1	0.44	0.44	0.79	0.38	0.02
Response format × Polarity	1	0.06	0.06	0.20	0.65	0.01
Error (Response format)	40	11.67	0.29			
Error (Polarity)	40	22.23	0.56			

Table 6.27. Simple effect analysis font size 64 point.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Response format	1	0.25	0.25	2.13	0.15	0.03
Polarity	1	0.66	0.66	2.41	0.13	0.04
Response format × Polarity	1	0.76	0.76	5.97	0.02	0.08
Error (Response format)	66	8.41	0.13			
Error (Polarity)	66	18.24	0.28			

A paired samples t-test (Table 6.28) showed a significant effect of response format for polarity black on white, t (32) = 3.40, $p < 0.005$, with slower responses when radio buttons were used.

Table 6.28. Paired samples t-test (response formats: radio button, visual analogue scale; font size 64 point.

Polarity	*df*	*mean (SD)* RB	VAS	*t*	*p*	*r*
Black on white	32	5.52 (0.40)	5.28 (0.30)	3.40	0.002	0.51
White on black	34	5.23 (0.66)	5.29 (0.33)	-0.61	0.545	0.01

Note: RB: Likert scale using radio button; VAS: visual analogue scale.

An independent samples t-test (Table 6.29) showed significant effect on completion time with response format radio button and no significant effect with response format visual analogue scale,

(1) response format radio button for positive polarity (M = 5.52, SD = 0.40) and negative polarity (M = 5.23, SD = 0.66) conditions; t (56) = 2.20, $r = 0.3$, $p = 0.003$;

(2) response format visual analogue scale for positive polarity (M = 5.28, SD = 0.30) and negative polarity (M = 5.29, SD = 0.33) conditions; t (66) = -0.13, $r = 0.0$, $p = 0.900$.

Table 6.29. Independent samples t-test (response formats: radio button, visual analogue scale; font size 64 point.

Response format	*df*	*mean (SD)* Black on white	White on black	*t*	*p*	*r*
RB	56	5.52 (0.40)	5.23 (0.66)	2.17	0.03	0.3
VAS	66	5.28 (0.30)	5.29 (0.33)	-0.13	0.900	0.0

6.6.7.2 Perceived enjoyment

Table 6.30 details the means, standard deviations and the confidence intervals of the perceived enjoyment scores. The scores represent the sum of the three questionnaire items of the perceived enjoyment construct. A higher score was observed consistently for the visual analogue scale format than for radio buttons as the response format.

Table 6.30. Descriptives of perceived enjoyment scores.

		Perceived enjoyment (sum of items)			
Response format		RB		VAS	
Font size		44	64	44	64
mean		8.00	8.58	14.58	14.82
Std. Deviation		4.58	4.28	4.76	4.53
95% Confidence Interval for mean	Lower Bound	6.49	7.54	13.01	13.72
	Upper Bound	9.51	9.63	16.14	15.93

Note: RB: radio button format; VAS: visual analogue scale format.

A mixed 2×2×(2) ANOVA (Table 6.31) revealed that the main effect of response format F (1, 101) = 87.11, η_p^2 = 0.46, p < 0.01 was significant.

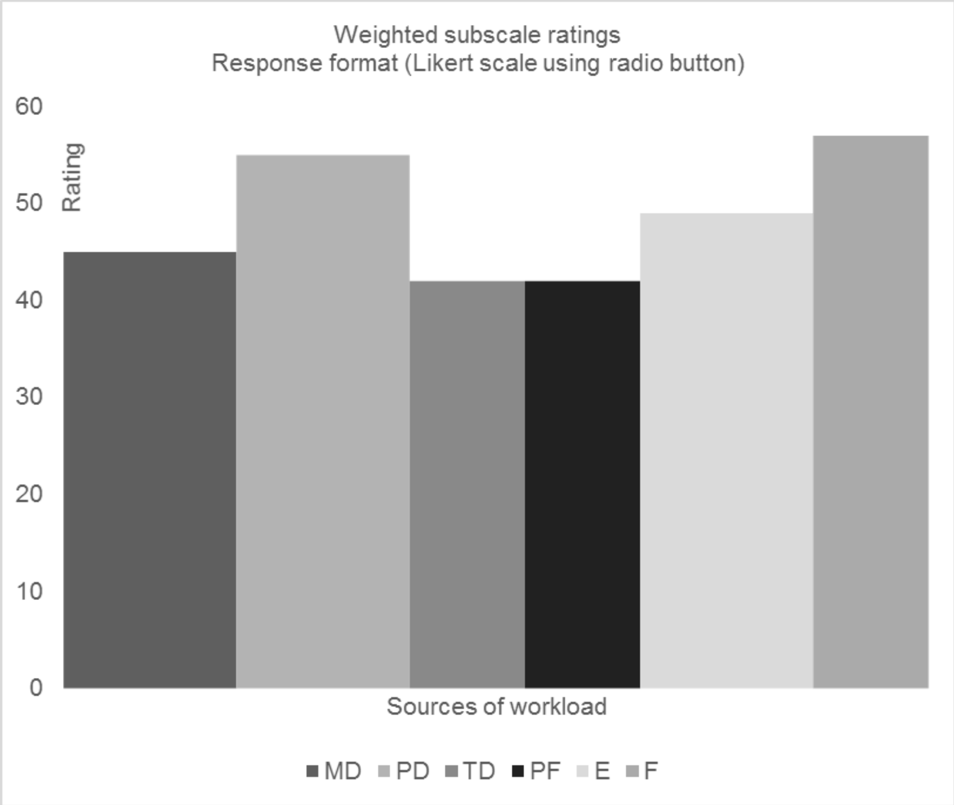
Table 6.31. Mixed ANOVA summary table for perceived enjoyment.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity[a]	1	12.04	12.04	0.72	0.40	0.01
Font size	1	8.88	8.88	0.53	0.47	0.01
Polarity × Font size	1	0.15	0.15	0.01	0.92	0.00
Error (Font size and Polarity)	101	1681.51	16.65			
Response format	1	2124.88	2124.88	87.11	0.00	0.46
Response format × Polarity	1	12.44	12.44	0.51	0.48	0.01
Response format × Font size	1	1.14	1.14	0.05	0.83	0.00
Response format × Polarity × Font size	1	9.65	9.65	0.40	0.53	0.00
Error (Response format)	101	2463.63	24.39			

[a] text/background polarity

6.6.7.3 Workload

A graphic representation of the weighted subscale rating for the NASA-TLX workload questionnaire is shown in Figure 6.6 for both response formats (radio button and visual analogue scale). The X-axis represents the weight of the six subscales: mental demand (MD), physical demand (PD), temporal demand (TD), performance (PF), effort (E) and frustration (F). The Y-axis represents the workload rating of these subscales.



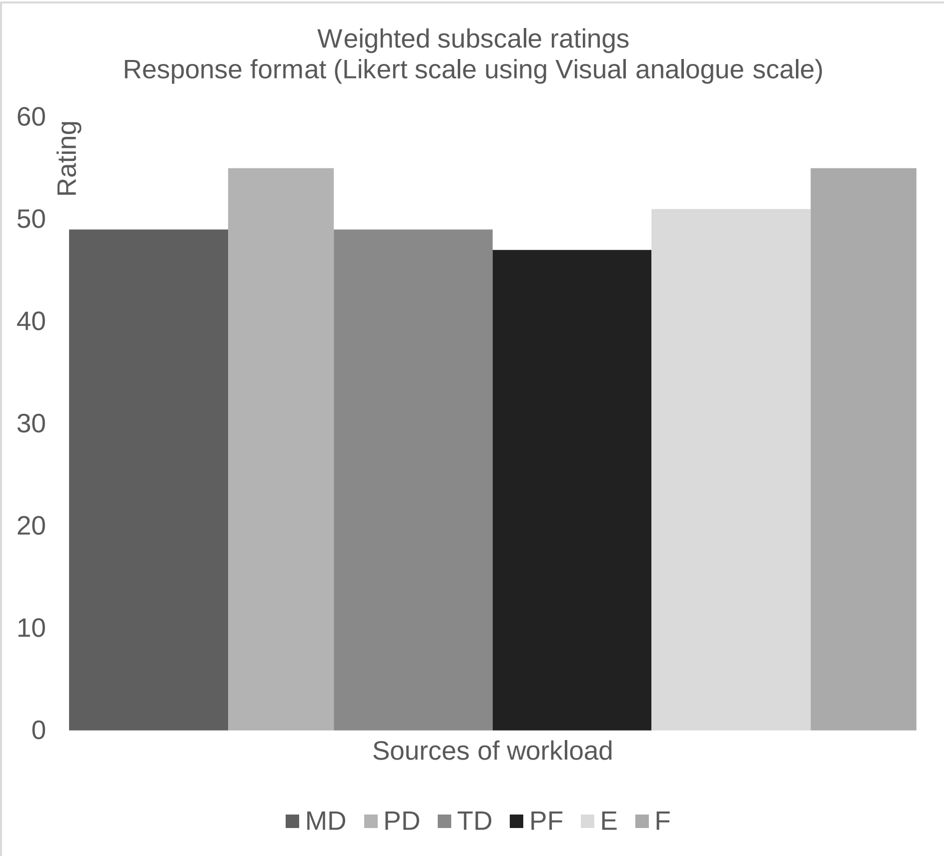


Figure 6.6. Graphic representation of weighted subscale ratings.

The descriptive statistics for the raw scores of the subscales and the weighted overall workload score is described in Table 6.32. It is observed that the overall weighted workload score was just below the midpoint of 50 for all design parameters except,

(1) font size 44 point, text/background polarity black on white, response format visual analogue scale; mean (SD) = 55.24 (16.82),

(2) font size 64 point, text/background polarity black on white, response format visual analogue scale; mean (SD) = 49.07 (18.05).

According to the research by Grier (2015), the workload score of 55 is above the midpoint but not remarkable so. Considering the frequency distribution of scores reported in this study collected from various other research studies for all tasks, the mean 55.24 is higher than 60% of the scores. Similarly, scores in the range of 45 - 49 is higher than 40% of the scores and scores in the range of 44 to 45 is higher than 30% of the scores. Although the workload scores are somewhat high, they fall in the range of acceptable workload scores.

Table 6.32. Descriptive statistics of NASA-TLX subscales according to design parameters font size, text/background polarity and response format.

	mean (SD)							
	RB				VAS			
	44		64		44		64	
Subscales	BW	WB	BW	WB	BW	WB	BW	WB
Mental Demand	53.33 (21.96)	46.31 (35.76)	36.54 (22.81)	48.15 (25.80)	51.47 (23.10)	46.67 (25.39)	52.95 (18.90)	44.10 (21.06)
Physical Demand	50.13 (20.90)	55.06 (32.12)	51.25 (27.22)	60.85 (31.73)	53.60 (29.00)	59.73 (25.30)	52.95 (26.13)	55.20 (25.81)
Temporal Demand	39.47 (18.88)	43.48 (26.64)	46.92 (21.09)	37.47 (27.28)	58.40 (26.76)	50.67 (18.28)	47.54 (21.80)	43.95 (23.83)
Performance	49.33 (26.23)	40.09 (32.06)	46.30 (23.24)	35.15 (25.65)	51.73 (22.61)	48.00 (29.95)	51.72 (22.37)	40.65 (23.98)
Effort	56.27 (21.07)	49.13 (28.93)	44.13 (21.52)	48.46 (24.61)	53.07 (25.05)	45.87 (25.46)	58.53 (18.85)	46.05 (27.34)
Frustration	62.40 (21.08)	55.91 (32.11)	55.59 (22.02)	57.14 (28.84)	65.60 (24.64)	49.07 (22.37)	52.03 (18.68)	54.00 (28.49)
Weighted Overall workload	45.56 (18.68)	45.72 (24.00)	44.44 (15.08)	46.48 (21.47)	55.24 (16.82)	44.01 (20.71)	49.07 (18.05)	44.97 (18.84)

Note: RB: Radio button; VAS: Visual analogue scale; SD: Standard deviation; BW: Black on white; WB: White on black.

A mixed 2×2×(2) ANOVA was used to test the effects of font size, polarity and response format on the respondents' workload experience. The details of the mixed ANOVA are presented in Table 6.33. No main effect or interaction effect was observed for font size, text/background polarity and response format on the weighted subscale scores.

Table 6.33. Mixed ANOVA summary table for NASA-TLX.

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity[a]	1	598.12	598.12	1.57	0.21	0.02
Font size	1	156.24	156.24	0.41	0.52	0.00
Polarity × Font size	1	586.49	586.49	1.54	0.22	0.02
Error (Font size and Polarity)	96	36513.07	380.34			
Response format	1	339.53	339.53	1.23	0.27	0.01
Response format × Polarity	1	449.68	449.68	1.63	0.21	0.02
Response format × Font size	1	34.81	34.81	0.13	0.72	0.00
Response format × Polarity × Font size	1	0.42	0.42	0.00	0.97	0.00
Error (Response format)	96	26517.36	276.22			

[a] text/background polarity

6.6.7.4 Summary of mixed ANOVA

In summary, for completion time, font size 64 point was faster than 44 point only in one condition for the design parameters negative polarity and response format visual analogue scale. Font size 44 point had faster completion times for response format radio button. For perceived enjoyment, the mean score indicated higher enjoyment for the visual analogue scale format. This was evident with the significant results obtained for the visual analogue scale response format.

For workload, no significant results were observed for the effects of font size, text/background polarity and response format.

6.8 Discussion

In summary, this chapter presented the method and analysis to test the effect of three design parameters (font size, text/background polarity and response format) in online psychometrics on small-screen devices. The hypothesis for each of these design parameters, perceived enjoyment and workload were stated. The method of the experiment to test each of the hypotheses was presented. The psychometric properties of the questionnaires were tested for each of the combinations of the design parameters. The hypotheses were tested with the mixed-measures ANOVA.

6.8.1 Psychometric analysis

6.8.1.1 Reliability analysis

The reported reliability analysis of all questionnaires was good, consistent with the English versions except for the perceived ease of use construct (see Section 6.6.1).

6.8.1.2 Validity

The three subscales of PSSUQ exhibited strong correlation consistent with previous research (Lewis, 1995, 2002) across both response formats, thereby showing evidence of convergent validity. The remaining correlations were low, showing evidence of discriminant validity.

6.8.1.3 Factor structure

In addition to the first attempt of the translated versions of all the questionnaires being adopted, this is also the first attempt of validating the questionnaires with two response formats along with the manipulation of design parameters such as font size and text/background polarity. For the usability questionnaire PSSUQ, the factor structure established earlier by Lewis (1995, 2002) could not be replicated. The factor structure of the SUS across certain design parameter combinations (see Section 6.6.4) was consistent with previous established studies (e.g. Sauro and Lewis, 2017). The factor structure of disorientation, perceived ease of use and perceived usefulness for certain combinations of design parameters (see Section 6.6.5) was consistent with previous studies by van Schaik and Ling (2005, 2007). The factor structure of PE was consistent with previous research by Davis (1992) for all combinations of design parameters and response formats. A summary of the factor extraction and the rotation method for all the questionnaires is detailed in Table 6.34

Table 6.34. Factor extraction and rotation methods

Questionnaire	Factor extraction method		Rotation method	
	RB	VAS	RB	VAS
PSSUQ	ULS	ULS	Oblimin	Oblimin
SUS	ML	ULS	Varimax	Oblimin
DIS-PEU-PU	ULS	ML	Oblimin	Varimax

Note: RB: Likert scale using radio button; VAS: visual analogue scale; ULS: unweighted least squares; ML: maximum likelihood.

6.8.2 Mixed ANOVA

6.8.2.1 Completion time

The following hypotheses were developed in Chapter 3 and tested here.

Hypothesis 1: completion time decreases with increasing font size;

Hypothesis 3b: positive text/background polarity decreases completion time.

Hypothesis 4: completion time of Likert scales is shorter with radio button format than visual analogue scale format.

Hypothesis 1 is rejected because, completion time for font size 44 point was faster than font size 64. Hypothesis 3b is partially supported, for only one specific combination of design parameters, as with negative polarity faster completion time was observed only with font size 64 point and response format visual analogue scale. Hypothesis 4 is partially supported because with response format radio buttons completion time was faster, only when font size 44 point was used.

6.8.2.2 Perceived enjoyment

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 10b: positive text/background polarity enhances perceived enjoyment;

Hypothesis 11: perceived enjoyment is higher for Likert scale using radio button format than with visual analogue scale format.

No significant effect of font size and text/background polarity was evident, thus rejecting Hypothesis 8 and 10b. A significant main effect of response format on perceived enjoyment was observed. The mean scores for perceived enjoyment were higher for the visual analogue scale format thus rejecting Hypothesis 11.

6.8.2.3 Workload

Hypothesis 5: workload decreases with increasing font size;

Hypothesis 6: positive text/background polarity decreases workload;

Hypothesis 7: workload is lower with radio button format than with visual analogue scale format;

There was no evident significant main effect of font size, text/background polarity and response format on the workload scores, thus rejecting Hypothesis 5, 6 and 7.

6.9 Limitations

A limitation observed in Study 2-A, is with regards to the devices used by the participants. The respondents interacted with the questionnaires in the experiment using their own mobile devices. Thus the environment was not strictly controlled. A future recommendation would be to provide devices with exactly the same specifications to achieve standardisation of screen size and screen presentation.

Given the sample size 110 participants, a post hoc power analysis indicated 100% chance (power = 1 – β = 1.0) of detecting a large and medium effect size (f = 0.40, f = 0.25) and 50% chance (power = 1 – β = 0.55) of detecting a small effect size (f = 0.10). Therefore, the sample size was deemed to be sufficient for the repeated measures ANOVA tests. The KMO values in the factor analysis of certain psychometric scales indicated that the data were not suitable for factor analysis for certain combinations of design parameters. This could be due to insufficient data collection or insufficient comprehension of the questions due to inappropriate Arabic translation. However, the rigorous process of translation and back-translation suggests that the latter is unlikely.

Nevertheless, the translation and comprehension of the translated questionnaire could have limited the factor replication of the previously well-established psychometric questionnaires. With regards to sample size, Nunnally (1978) recommended having 10 times as many participants as variables; therefore the required sample size would need to be 10 × 10 variables = 100 for the SUS, 10 × 14 variables = 140 for the disorientation scale, perceived usefulness scale and perceived ease of use scale combined, and 10 × 19 variables = 190 for the PSSUQ. In addition, Tabachnick & Fidell (pg. 613, 2013) agree that ‘it is comforting to have at least 300 cases for factor analysis’. However, the number of participants was only 110 indicating sufficient sample size for factor analysis of SUS and insufficient for all other scales. Therefore, recommendations for future work include increased sample size for factor analysis (Tabachnick & Fidell, 2013) and refinement of the Arabic-language translation of the questionnaires.

Chapter 7

Study 2-B

Contents

Chapter 7: study 2-B

7.1 Overview

When actively designing questionnaires for multiple devices such as desktop computers and handheld mobile devices, it is common to take the computer-web design as the template and adapt it to mobile devices (de Bruijne & Wijnant, 2013a). However, with the rise in smartphone use, it is important to have questionnaires designed in a way that can be adapted on both smartphones and desktop computers. Because of the difference in screen sizes between smartphones and desktop computers, questionnaires viewed on a smartphone, will look and behave differently when viewed on a desktop. In addition design parameters chosen for smartphones may not be adaptable for desktop computers (e.g. font size). In the context of online psychometrics for both desktop and small-screen devices, currently there is little research evidence to inform design guidelines for web-based administration of psychometric questionnaires on desktop computers and no research evidence as far as the author is aware for small-screen devices. The few studies that exist conducted by van Schaik and Ling (2003, 2005a, 2005b, 2007) and van Schaik et al. (2015) report results from experiments conducted on desktop computers with regards to design parameters for rating scales and questionnaire layout for online psychometrics. The current research examines, if design parameters such as font size, text/background polarity and response format affect respondents' completion time of the questionnaires, mental workload and perceived enjoyment on two platforms: mobile devices (see Chapter 6) and desktop computers (current Chapter 7). In addition, the psychometric properties of the questionnaires when these design parameters are manipulated are also investigated.

Chapter 7: study 2-B

This chapter begins with the objectives of the study, followed by the statement of the hypotheses and the description of the design and method of the experiment. The analysis sections follow and finally the results and discussion of the Study are presented.

7.2 Objectives of Study 2-B

Study 2-B was conducted on desktop computers and addressess the following objectives. First, to test the effect of design parameters font size on questionnaire completion time. Second, to test the effect of design parameters text/background polarity on questionnaire completion time. Third, to test and compare the effect of response format (Likert scale using radio button vs Likert scale using visual analogue scale) also on questionnaire completion time. In this study, the author reports two experiments investigating three parameters of questionnaire design:

(1) font size (12 point vs 20 point);
(2) response format (Likert Scale [using radio button] vs Likert scale [using Visual analogue scale]);
(3) text/background polarity (black on white vs white on black).

The psychometric properties of questionnaires were also analysed. The current study was conducted with Arabic speakers and, for this study, Arabic versions of the questionnaires were required. As mentioned in Chapter 3, since Arabic version of established questionnaires to measure usability (such as PSSUQ by Lewis, 1995; and SUS by Brooke, 1996), perceived disorientation (Ahuja & Webster, 2001), perceived ease of use, perceived usefulness (Davis et al 1989), perceived enjoyment (Davis 1992), and mental workload (NASA-TLX, Hart & Staveland, 1988) did not exist before this research started, forward-and back translations of the questionnaires were conducted and reviewed by an expert committee (see Section 3.8.2).

7.3 Statement of hypotheses

As described in Sections 3.3 and 6.4, the following Hypotheses are presented for Study 2B.

Hypothesis 1: completion time decreases with increasing font size;

Hypothesis 3b: positive text/background polarity decreases completion time;

Hypothesis 4: completion time of Likert scales is shorter with radio button format than visual analogue scale format;

Hypothesis 5: workload decreases with increasing font size;

Hypothesis 6: positive text/background polarity decreases workload;

Hypothesis 7: workload is lower with radio button format than with visual analogue scale format;

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 10b: positive text/background polarity enhances perceived enjoyment;

Hypothesis 11: perceived enjoyment is higher for Likert scale using radio button format than with visual analogue scale format.

7.4 Method

7.4.1 Design

The experiment designed for Study 2-B used a 2×2×(2) experimental design. The first, second and third independent-measures, variables were:

(1) font size: (12 point or 20 point);

(2) text/background polarity: (black on white [also used in Study 1 and Study 2-A] or white on black [used in Study 2-A]);

(3) response format: (Likert scale using radio button format and Likert scale using visual analogue scale format).

The dependent variable was completion time for the set of five psychometric questionnaires, perceived enjoyment and mental workload.

7.4.1.1 Research design

The experiment was carried out on desktop computers. Eight versions of the questionnaires were created. The experiment was organised into four groups: according the font size, text/background polarity and response format. The response format was the repeated-measure. The details are as follows:

Group 1: 12 point, black on white, radio button/visual analogue

Order/version 1: 12 point, black on white, radio button/visual analogue

Order/version 2: 12 point, black on white, visual analogue/radio button

Group 2: 12 point, white on black, radio button/visual analogue

Order/version 3: 12 point, white on black, radio button/visual analogue

Order/version 4: 12 point, white on black, visual analogue/radio button

Group 3: 20 point, black on white, radio button/visual analogue

Order/version 5: 20 point, black on white, radio button/visual analogue

Order/version 6: 20 point, black on white, visual analogue/radio button

Group 4: 20 point, white on black, radio button/visual analogue

Order/version 7: 20 point, white on black, radio button/visual analogue

Order/version 8: 20 point, white on black, visual analogue/radio button

Similar to Study 1 (see Section 5.3.1) the participants in each group responded to five psychometric questionnaires administered in a fixed order (PSSUQ, SUS, DIS, PEU and PU). A brief introduction to these questionnaires is provided in Section 3.4. Similar to Study 2-A (see Section 6.5.1), this experiment was a retrospective assessment of the Moodle-based learning management system and the participants were not required to use the system within this experiment. The participants responded to the psychometric questionnaires by rating their experience with the university's Moodle-based learning management system website. Immediately after the participants completed their responses to the five psychometric questionnaires, they completed the perceived enjoyment (PE) scale developed by Davis (1992); specifically they rated their experience of responding to the psychometric questionnaires they just completed. Following this, the participants responded to the workload questionnaire NASA-TLX (Hart & Staveland, 1998); specifically, they rated their experience of responding to the five psychometric questionnaires. Each participant completed the psychometric questionnaires twice: once with each response format (radio button and visual analogue).

7.4.2 Participants

A total of one hundred six university students from Kuwait (56 female; 50 male) took part. All the participants were native Arabic-speakers. They were enrolled for the course CSC 102 Computer Programming for Beginners.

7.4.3 Materials and equipment

The experiment was administered through the Online Psychometric Questionnaire Design tool OnPQDT developed for this research purpose (see Chapter 4).

7.4.3.1 Equipment

All participants of Study 2-B used the desktop computers in a computer lab within the university campus. Table 7.1 details the specifications of the desktop computers used in the lab for the experiment.

Table 7.1. Dimensions of Desktop computer used in the experiment Study 2-B.

	Screen Size (inches) diagonally	Resolution (pixels)	Pixels per inch (ppi)	Processor Type	Dimension (cm)
Lenovo All-in-one PC	21.5	1920×1080	102.46	Intel Core i3-5005U	53×5.1×36.1

7.4.4 Procedure

Research ethics, electronic informed consent, verbal briefing on the reasons for participation prior to the experiment, incentives and voluntary participation, and confidentiality of data collected were highly similar to those in Study 1 (see Section 5.3.4). Data collection took place between December 2017 and April 2018. The experimental sessions took place in a computer lab within the campus of a private university and each session lasted for about 20 minutes. The author was present along with the course leader at all the sessions. The sessions were led by the author. Every participant in each session received a URL link from the author. The URL link was protected by an Access code and the Access code (see Section 4.5.2.4) was provided to the participants to prevent unauthorized access to the questionnaire otherwise. When accessed, the URL link first provided the participants with an online consent form.

Thereafter, brief online instructions (see Section 4.5.2.3) in Arabic were provided, on how to complete, the five psychometric questionnaires. At the end of the responses to the five psychometric questionnaires, a brief instruction page was administered online on how to complete the perceived enjoyment scale based on their experience of responding to the psychometric questionnaires. Thereafter, the NASA-TLX workload questionnaire was administered with an online instruction page guiding the participants to respond to the workload questionnaire, based on their experience of responding to the psychometric questionnaires. The experiment automatically ran two times for the two response formats (see Section 7.4.1). The font size and the text/background polarity remained uniform according to the groups assigned for the sessions. At the end of the experiment, a ‘Thank you’ page was presented to every participant in appreciation for their time and involvement. The author, along with the module leader, was present till all the students completed the experiment. Figure 7.1 shows an experimental session in progress captured by the author. The data collected for the Study 2-B were analysed using SPSS version 23 and the output will be provided by the author upon request.

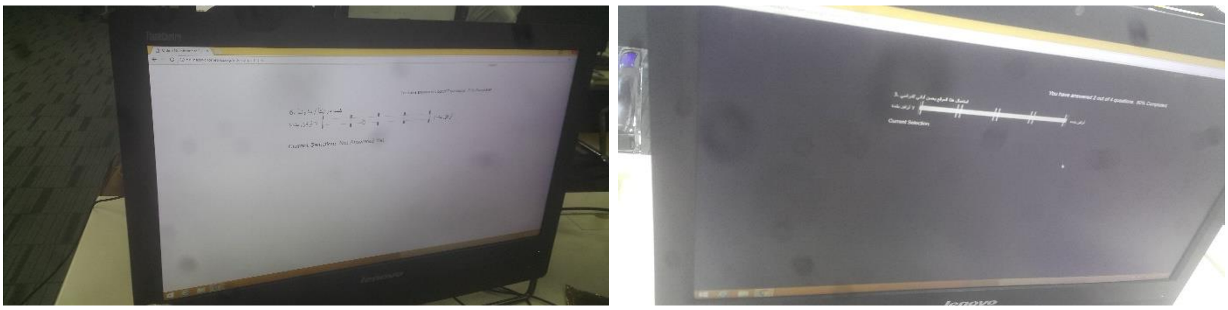


Figure 7.1. Study 2-B text/background polarity (black on white; white on black).

7.5 Analysis of Study-2B

7.5.1 Reliability

Figures 7.2, 7.3, 7.4 and 7.5 illustrate the reliability coefficients for all scales (DIS, PEU, PU, PSSUQ and SUS). The clustered bar graph depicts the response format (visual analogue scale and radio button) for all questionnaires according to font size and text/background polarity. All the scales were reliable, consistent with the English versions and no exceptions were observed.

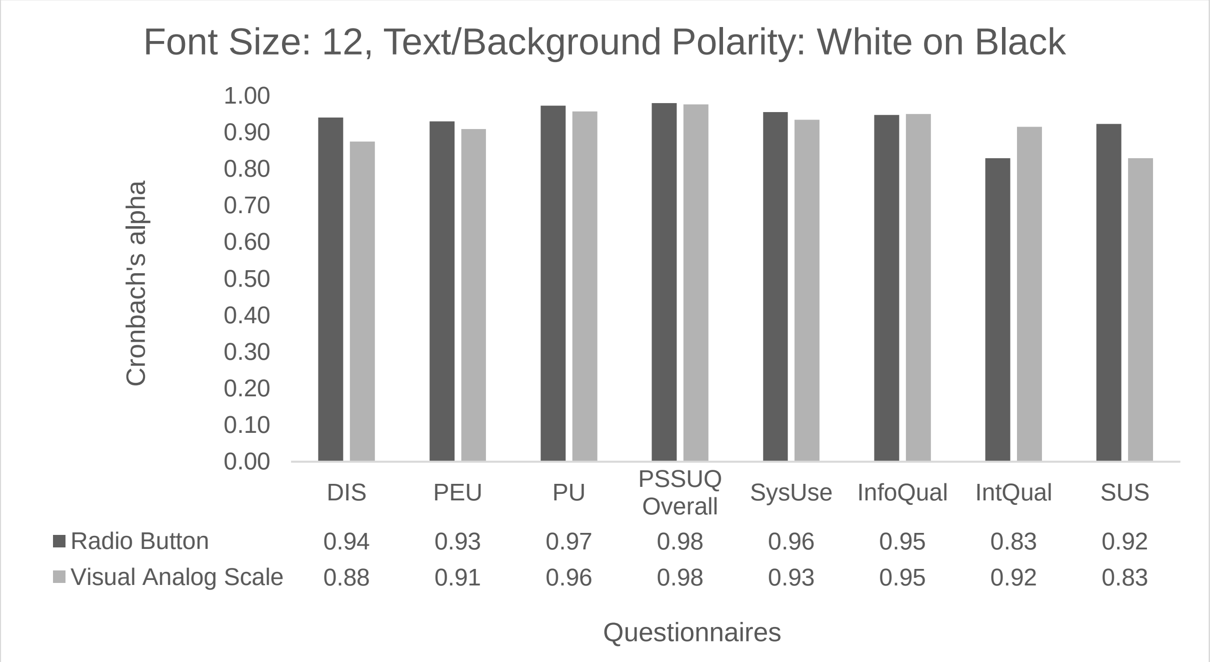


Figure 7.2. Reliability analysis Study 2-B (12 point; white on black).

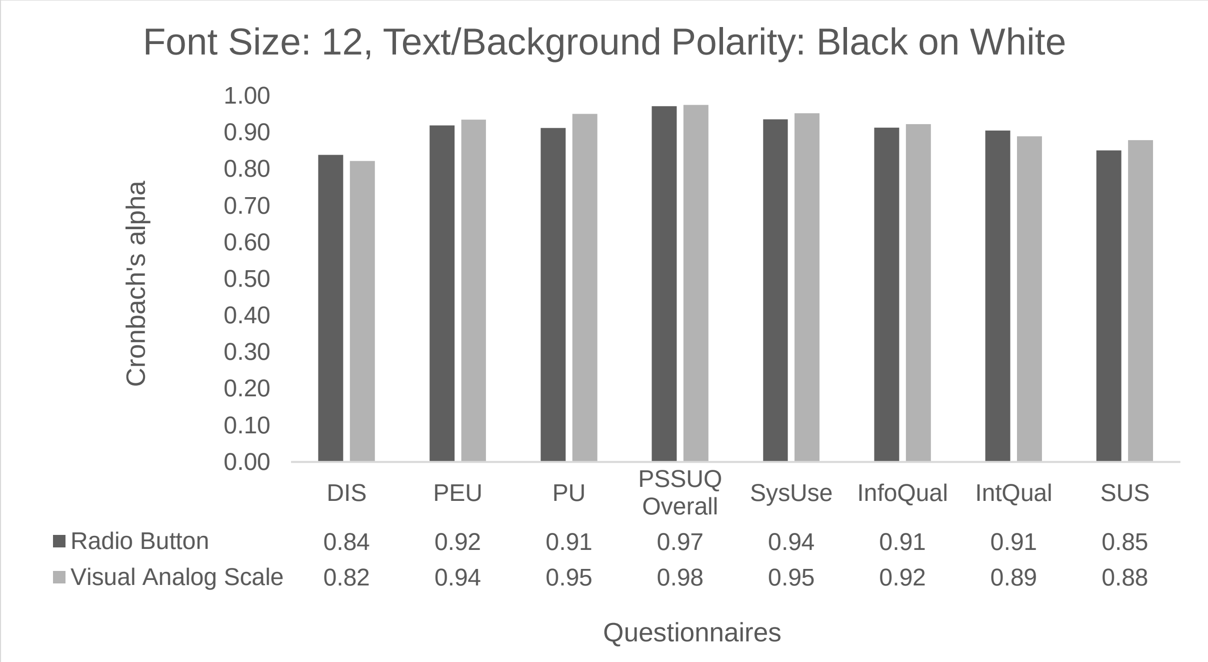


Figure 7.3. Reliability analysis Study 2-B (12 point; black on white).

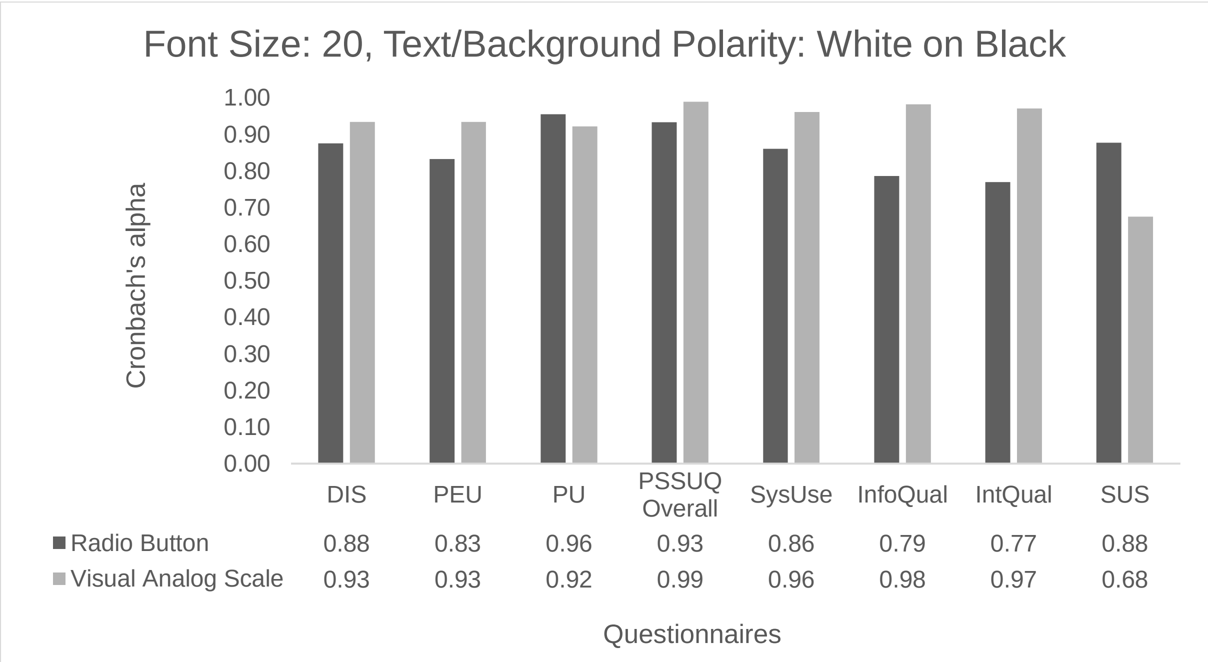


Figure 7.4. Reliability analysis Study 2-B (20 point; white on black).

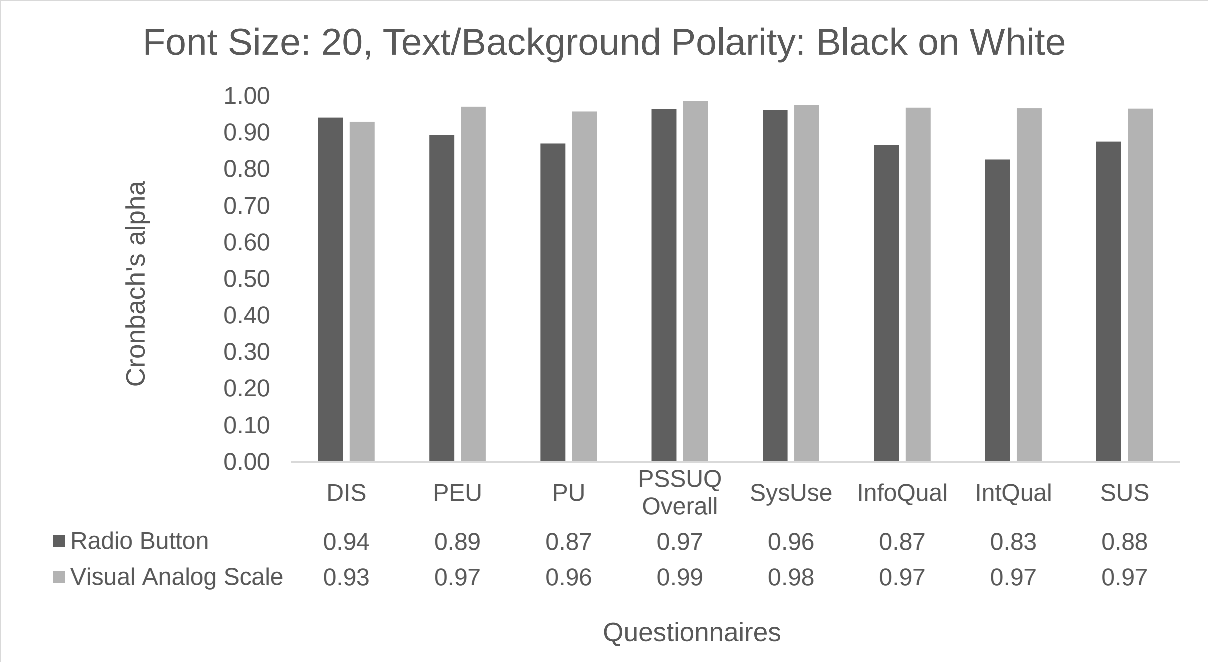


Figure 7.5. Reliability analysis Study 2-B (20 point; black on white).

The reliability analysis for the perceived enjoyment (PE) construct is presented in Figure 7.6. The clustered bar graph depicts the response format (Likert scale using visual analogue scale and Likert scale using radio button) for all combinations of font size and text/background polarity. The scale possessed high reliability, consistent with the English version for all combinations of the design parameters (font size, text/background polarity and response format).

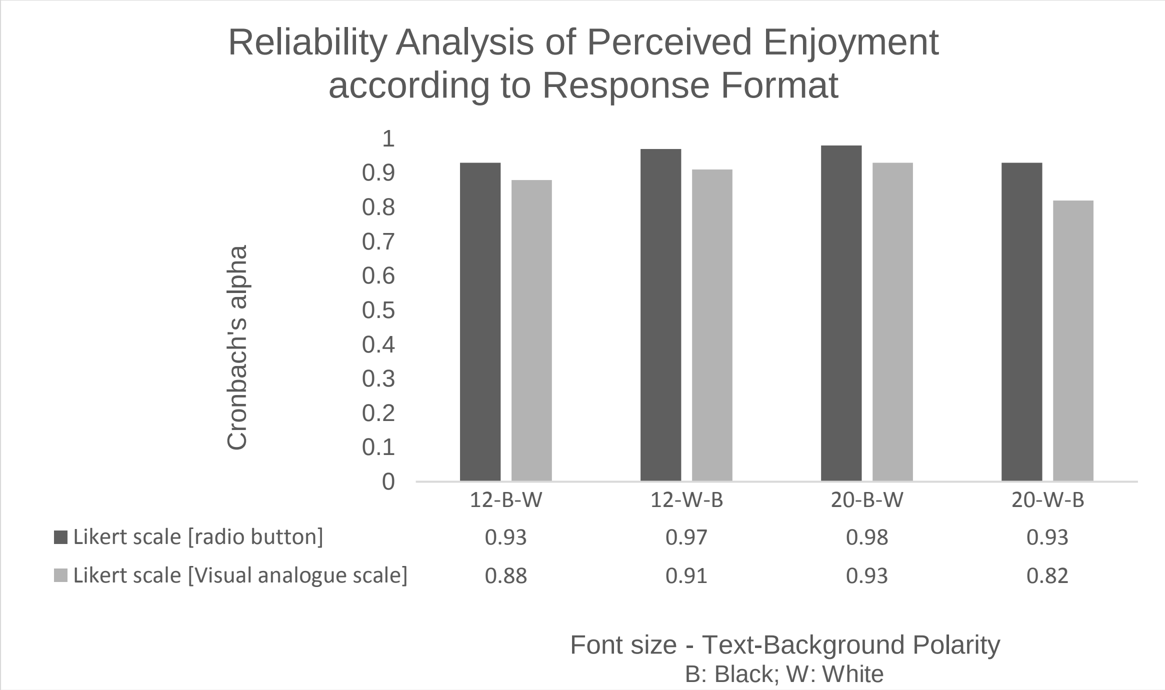


Figure 7.6. Reliability analysis of perceived enjoyment (Study 2-B).

7.5.1.1 Summary of Reliability analysis

In summary, the results for all the psychometric questionnaires (DIS, PEU, PU, PSSUQ, SUS and PE) showed they possessed good reliability for all combinations of design parameters and response formats.

7.5.2 Validity

Pearson's correlation coefficient is used to assess criterion-related validity between the constructs disorientation, perceived ease of use, perceived usefulness, PSSUQ, the subscales of PSSUQ: SysUse, InfoQual, IntQual and SUS. The results are reported according to the response format (Tables 7.2 and 7.3).

7.5.2.1 Response format: visual analogue scale

The correlation between the questionnaires for the response format: visual analogue scale is presented in Table 7.2. A significantly strong correlation was observed between the subscales SysUse and InfoQual $r = .93$; SysUse and IntQual $r = .93$; InfoQual and IntQual $r = .86$; all $p < .01$.

Table 7.2. Correlations (Response Format: Visual Analogue Scale).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	-.05	-.10	-.04	.00	-.09	.00	-.28**
PU		-.03	-.04	-.06	-.02	-.06	.06
PEU			.19	.16	.16	.20	.15
PSSUQ				.99**	.96**	.94**	.10
SYS					.93**	.93**	.08
INFO						.86**	.13
INT							.06

PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual

**. Correlation is significant at the 0.01 level (2-tailed).

7.5.2.2 Response format: radio button

The correlation between the questionnaires for the response format: radio button is presented in Table 7.3. A significantly strong correlation was observed consistent with previous research (Lewis, 1995; 2002) between the subscales SysUse and InfoQual $r = .90$; SysUse and IntQual $r = .83$; InfoQual and IntQual $r = .92$; all $p < .01$).

Table 7.3. Correlations (Response Format: Radio Button).

	PU	PEU	PSSUQ	SYS	INFO	INT	SUS
DIS	.17	.22*	-.09	-.16	-.10	-.09	-.22*
PU		.16	.03	.03	.00	-.01	-.08
PEU			-.06	-.12	.04	.03	.08
PSSUQ				.97**	.97**	.91**	.08
SYS					.90**	.83**	.02
INFO						.92**	.09
INT							.16

PSSUQ (overall score); SYS: SysUse; INFO: InfoQual; INT: IntQual

*. Correlation is significant at the 0.05 level (2-tailed).

7.5.2.3 Summary of validity

A strong positive correlation was observed between the PSSUQ overall and its subscales: SysUse, InfoQual and IntQual. This observation is consistent with the previous research by Lewis (1995; 2002).

7.5.3 Factor structure of PSSUQ

Based on literature studies (Lewis, 1995, 2002), the factor structure of PSSUQ had subscales SysUse, InfoQual and IntQual (see Table 5.6, Section 5.4.3). Close observation of the correlation values between the PSSUQ items, revealed moderately strong values within the PSSUQ subscales ranging from 0.46 to 0.93 and fairly lower correlation values 0.25 to 0.45 between the PSSUQ subscales. This pattern of correlation provides some initial evidence for the factors of PSSUQ, with higher correlations among items within the subscales and lower correlations among items between the subscales. Previous research (Lewis, 1995, 2002) reported principal axis factoring as the extraction method, with varimax rotation, in order to determine the factor structure of PSSUQ.

In the current study, principle axis extraction technique with oblimin rotation was initially used for all combinations of design parameters font size and text/background polarity to detect the factor structure of the usability questionnaire PSSUQ. However, a solution could not be obtained consistently across all combinations of the design parameters because in Iteration 999 the communality of a variable exceeded 1.0 and SPSS terminated the extraction. Similarly, maximum likelihood method too failed to present a consistent solution for all combinations of design parameters due to termination of extraction. However, the unweighted least squares extraction with oblimin rotation presented a consistent solution across all combinations of the design parameters and response formats. Thus, this method was used to assess the factor structure for the 19 items of the PSSUQ.

Factor loadings < 0.3 were suppressed and three factors were explicitly requested. Three question items Q5, Q6 and Q13 were removed due to high correlation (the lowest correlation value observed was 0.51 and the highest was 0.93) for the response format visual analogue scale, for all combinations of the design parameter (font size and polarity). According to Kaiser (1974), the KMO value and Bartlett's test of sphericity values reported in Table 7.4 indicated the suitability of the data for structure detection for all design parameters except for the following combination of design parameters:

(1) radio button format:

(*KMO* =.36) font size: 12 point; polarity: black on white;

Table 7.4. KMO and Bartlett's test of sphericity values (PSSUQ)-Study 2-B.

Font size, text/background polarity	Visual analogue scale	Radio button format
12 point, black on white	*KMO* = .82; Bartlett: $\chi^2(36) = 302.80$	*KMO* = .86; Bartlett: $\chi^2(28) = 245.35$
20 point, white on black	*KMO* = .79; Bartlett: $\chi^2(36) = 268.56$	*KMO* = .87; Bartlett: $\chi^2(28) = 293.18$
12 point, black on white	*KMO* = .50; Bartlett: $\chi^2(36) = 137.87$	*KMO* = .36; Bartlett: $\chi^2(28) = 77.73$
20 point, white on black	*KMO* = .87; Bartlett: $\chi^2(36) = 346.71$	*KMO* = .67; Bartlett: $\chi^2(28) = 87.24$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research, factor analysis was still conducted for all combinations of design parameters and the results are reported in Tables 7.5, 7.6, 7.7 and 7.8 according to the combinations of the design parameters font size and polarity.

Table 7.5. Factor analysis of PSSUQ (12 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1		.68		.58	.33	
Q2	.83		-.30	.75		
Q3	.89			.66		
Q4	.92			.95		
Q5	[a]	[a]	[a]	.52		
Q6	[a]	[a]	[a]	.91		
Q7		.80			.77	
Q8	.88			.75		
Q9	.70				.68	.48
Q10		.90			.86	
Q11		.87		.79		
Q12	.87			.69		
Q13	[a]	[a]	[a]	.76		
Q14		.83			.76	
Q15		.73			.69	
Q16	.61				.89	
Q17	.86				.85	
Q18	.93			.78		
Q19	.80			.95		
Eigenvalues	11.13	1.45	0.76	12.71	1.60	0.89
% Variance	68.36	7.72	3.43	65.71	7.18	3.15
No. of items	10	6	0	12	7	0

Note: factor loadings < 0.3 were suppressed.

[a] question items removed due to high correlation.

Table 7.6. Factor analysis of PSSUQ (12 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1		-.72		.88		
Q2	.85			.93		
Q3	.88				.72	-.33
Q4	.62			.94		
Q5	[a]	[a]	[a]	.35[b]	.35[b]	
Q6	[a]	[a]	[a]	1.02		
Q7		-.68	-.53		.96	
Q8	.83			.97		
Q9	.85			.66		.45
Q10		-.98			.88	
Q11		-.84			.86	
Q12	.90				.89	
Q13	[a]	[a]	[a]		1.04	
Q14		-.93		.76		
Q15		-.81		.94		
Q16	.75				.62	
Q17	.87			.87		
Q18	.79				.93	
Q19	.78			.34	.66	
Eigenvalues	11.26	1.64	0.74	14.08	1.64	0.72
% Variance	69.35	9.17	3.14	73.36	7.89	2.31
No. of items	10	6	0	10	9	0

Note: factor loadings < 0.3 were suppressed.

[a] question items removed due to high correlation.

[b] question item with equal factor loadings on the two factors excluded from the count for the number of items.

Table 7.7. Factor analysis of PSSUQ (20 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.98			.87		
Q2	.38	.39	.46	.82		
Q3	.71		.38	.82		
Q4			.92	.90		
Q5	[a]	[a]	[a]	.94	-.32	
Q6	[a]	[a]	[a]	.80		
Q7	.85					-.86
Q8			.94	.87		
Q9	.65	.48				-.49
Q10	.45	.32	.42			-.92
Q11	.84			.91		
Q12	.58	.39		.87		
Q13	[a]	[a]	[a]	.47	.84	
Q14	.89				.83	-.42
Q15	.94					-.80
Q16	1.04				.74	-.54
Q17	.84					-.86
Q18	.84			.50	.69	
Q19		.52	.64	.89		
Eigenvalues	13.13	1.07	0.87	12.09	2.94	1.93
% Variance	81.67	6.28	5.02	63.02	15.03	9.53
No. of items	12	0	4	9	6	4

Note: factor loadings < 0.3 were suppressed.

[a] question items removed due to high correlation.

Table 7.8. Factor analysis of PSSUQ (20 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	1	2	3	1	2	3
Q1	.74	.32		.66		
Q2	.82			.76		
Q3		.65			.71	
Q4		.43			.55	
Q5	[a]	[a]	[a]	.82		
Q6	[a]	[a]	[a]	.66		.52
Q7	.84				1.00	-.32
Q8	.77				.43	.66
Q9	1.10	-.30		.32		.68
Q10	.93				.69	
Q11	.73	.33		.92		
Q12	.88			.73		
Q13	[a]	[a]	[a]	.83		
Q14	.97				.68	
Q15	.89					.66
Q16	.85			.81		
Q17	1.02				.49	.73
Q18	.81		-.32		.76	
Q19	.83		.38	.93		
Eigenvalues	13.40	0.88	0.62	8.87	3.30	2.14
% Variance	83.26	3.84	1.95	45.29	15.99	9.9
No. of items	14	2	0	9	6	4

Note: factor loadings < 0.3 were suppressed.

[a] question items removed due to high correlation.

7.5.3.1 Summary of the factor structure of PSSUQ

In summary, across all the design parameter combinations, the original factor structure of PSSUQ reported in research studies (Lewis, 1995 and 2002) could not be replicated. The Eigenvalues across the design parameters were not consistently above the suggested minimum value of 1. Although three factors were extracted, loadings were evident on two factors consistently for the response format visual analogue scale for all combinations of this design parameter with font size and polarity. A similar structure was evident for the response format radio button for only the following two combinations of the design parameter (font size and polarity):

(a) 12 point; black on white;

(b) 12 point; white on black.

Due to a lack in the pattern of factor loadings, common factor names could not be suggested.

7.5.4 Factor structure of SUS

Based on literature studies the latest research reported by Lewis and Sauro (2017) presented a tone model for the factor structure of SUS (see Table 5.12, Section 5.4.4). For the purpose of validating the SUS construct administered in this research study, the correlation values were examined. Within the items of each of the two SUS subscales (positive and negative), a moderately strong correlation with values between 0.37 and 0.84 were observed. However, the items between the two subscales displayed very low correlations values between 0.09 and 0.47.

This pattern of correlation provides some initial evidence for the factors of SUS, with higher correlations among items within the subscales and lower correlations among items between the subscales. Previous research by Lewis and Sauro (2017) reported the use of all three extraction techniques such as principle components analysis (strictly not a factor analytic method, but commonly used), unweighted least squares and maximum likelihood extraction technique with varimax rotation. In the current research principle axis factoring with oblimin rotation was initially used for all combinations of design parameters font size and text/background polarity and response formats radio button format and visual analogue scale, to detect the factor structure of the usability questionnaire SUS. However, a solution could not be obtained consistently across all combinations of the design parameters and response formats because in Iteration 999 the communality of a variable exceeded 1.0 and SPSS terminated the extraction. Therefore, to obtain a consistent factor solution for all combinations of design parameters, unweighted least squares method was used with oblimin rotation.

Factor loadings < 0.3 were suppressed. Four question items Q2, Q3, Q7 and Q10 were removed due to high correlation (lowest correlation value observed was 0.68 and the highest was 0.84) for the response format visual analogue scale. One question item (Q6) was removed also due to high correlation for the response format radio button.

The KMO and Bartlett's test of sphericity values are reported in Table 7.9. According to Kaiser (1974), the KMO values that are reported were suitable for data analysis except for the following instance:

(1) radio button format: (*KMO* =.44) font size: 12 point; polarity: black on white;

Table 7.9. KMO and Bartlett's test of sphericity values (SUS)-Study 2-B.

Font size, text/background polarity	Visual analogue scale	Radio button format
12 point, black on white	*KMO* = .62; Bartlett: $\chi^2(15) = 104.90$	*KMO* = .70; Bartlett: $\chi^2(36) = 154.81$
20 point, white on black	*KMO* = .57; Bartlett: $\chi^2(15) = 39.95$	*KMO* = .74; Bartlett: $\chi^2(36) = 179.87$
12 point, black on white	*KMO* = .63; Bartlett: $\chi^2(15) = 66.43$	*KMO* = .44; Bartlett: $\chi^2(36) = 100.73$
20 point, white on black	*KMO* = .62; Bartlett: $\chi^2(15) = 27.39$	*KMO* = .64; Bartlett: $\chi^2(36) = 102.79$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research (see Table 5.12, Section 5.4.4), factor analysis was carried out for all combinations of design parameters across both the response formats and the results are reported in Tables 7.10, 7.11, 7.12 and 7.13.

Table 7.10. Factor analysis of SUS (12 point; black on white).

	Visual analogue scale		Radio button	
	Factors		Factors	
Items	1	2	Negative	Positive
Q1	0.33	0.37		0.50
Q2	[a]	[a]	0.90	
Q3	[a]	[a]		0.58
Q4	0.79		0.43	
Q5		1.07		0.64
Q6	0.98		[b]	[b]
Q7	[a]	[a]		0.73
Q8	0.57		0.73	
Q9		0.59		0.62
Q10	[a]	[a]	0.81	
Eigenvalues	3.04	1.25	3.92	1.55
% Variance	44.79	17.94	38.36	12.15
No. of items	3	3	4	5

Note: factor loadings < 0.3 were suppressed.

[a] question items were removed due to high correlation

[b] question items were removed due to high correlation

Table 7.11. Factor analysis of SUS (12 point; white on black).

Items	Visual analogue scale Factors 1	Visual analogue scale Factors 2	Radio button Factors 1	Radio button Factors 2
Q1	0.34			0.84
Q2	[a]	[a]	0.81	
Q3	[a]	[a]		0.43
Q4	1.05		0.61	
Q5		-0.78		0.98
Q6			[b]	[b]
Q7	[a]	[a]		0.54
Q8	0.51		0.98	
Q9		-0.92	0.97	
Q10	[a]	[a]	0.77	
Eigenvalues	2.57	1.12	5.34	1.34
% Variance	37.38	13.11	56.25	11.93
No. of items	3	2	5	4

Note: factor loadings < 0.3 were suppressed.

[a] question items were removed due to high correlation

[b] question item was removed due to high correlation

Table 7.12. Factor analysis of SUS (20 point; black on white).

	Visual analogue scale		Radio button	
	Factors		Factors	
Items	1	2	Positive	Negative
Q1		0.66	0.77	
Q2	[a]	[a]		0.88
Q3	[a]	[a]	0.91	
Q4	0.83			0.85
Q5	0.81	0.36	0.92	
Q6	0.89		[b]	[b]
Q7	[a]	[a]	0.80	
Q8	1.10	-0.31		0.88
Q9	0.70		0.47	0.33
Q10	[a]	[a]		0.64
Eigenvalues	4.49	0.89	4.16	2.38
% Variance	72.42	8.58	42.67	23.18
No. of items	5	1	5	4

Note: factor loadings < 0.3 were suppressed.

[a] question items were removed due to high correlation

[b] question item were removed due to high correlation

Table 7.13. Factor analysis of SUS (20 point; white on black).

	Visual analogue scale		Radio button	
	Factors		Factors	
Items	Positive	Negative	1	2
Q1	0.60		0.42	0.39
Q2	[a]	[a]	0.95	
Q3	[a]	[a]		0.74
Q4	0.46	0.64	0.63	
Q5	0.78		0.74	
Q6		0.59	[b]	[b]
Q7	[a]	[a]		0.80
Q8		0.75	0.65	
Q9	0.65			0.81
Q10	[a]	[a]		
Eigenvalues	2.19	1.79	4.27	1.68
% Variance	27.94	22.36	42.76	14.41
No. of items	3	3	5	3

Note: factor loadings < 0.3 were suppressed.

[a] question items were removed due to high correlation

[b] question items were removed due to high correlation

7.5.4.1 Summary of the factor structure of SUS

Considering the factor structure of SUS in this study, across all design parameters, the tone model reported by Lewis and Sauro (2015, 2017) was evident for the following response format and combinations of design parameters:

(1) response format: visual analogue scale; 20 point, white on black;

(2) response format: radio button; 12 point, black on white;

20 point, black on white.

For all other cases, the multi-dimensional factor structure established earlier in various studies (e.g. Lewis & Sauro, 2015, 2017) could not be replicated.

7.5.5 Factor structure of DIS, PEU and PU

Based on literature studies, research by van Schaik and Ling (2003, 2005, and 2007) for online psychometrics, presented a three-factor solution: (1) disorientation, (2) perceived ease of use and (3) perceived usefulness (see Table 5.18, Section 5.4.5). In this study, results of factor analysis were obtained, reported and compared with the previous studies by van Schaik and Ling. For the purpose of validating the DIS, PEU and PU constructs administered in this research study, the correlation values were examined. Within the items of each of the scales, a moderately strong correlation was observed with values ranging from 0.45 to 0.74 for the DIS construct, 0.66 to 0.75 for the PEU construct and 0.73 to 0.89 for the PU construct. Low correlation values were observed between the items of the DIS, PEU and PU constructs while a moderate correlation existed between the items of the PEU and PU constructs.

This pattern of correlation provides some initial evidence for the factors of DIS, PEU and PU, with higher correlations among items within the subscales and lower correlations among items between the subscales. Similar to previous research reported by van Schaik and Ling (2003, 2005, and 2007), principle axis factoring with oblimin rotation was initially employed for all combinations of design parameters (font size and text/background polarity) and both response formats (radio button and visual analogue scale) to detect the factor structure of the three questionnaires together. However, a solution could not be obtained consistently across all combinations of the design parameters and response formats because in Iteration 999 the communality of a variable (not reported in SPSS version 23) exceeded 1.0. However, unweighted least squares method with oblimin rotation was used and consistent results were obtained for all combinations of design parameters font size and text/background polarity and both response formats radio button and visual analogue scale.

Factor loadings < 0.3 were suppressed. According to Kaiser (1974), the KMO value and Bartlett's test of sphericity values reported in Table 7.14 indicated the suitability of the data for structure detection for all design parameters except for the following combination of design parameters:

(1) radio button format:

(*KMO* =.35) font size: 20 point; polarity: black on white;

(*KMO* =.46) font size: 20 point; polarity: white on black;

(2) visual analogue scale format:

(*KMO* =.32) font size: 20 point; polarity: black on white;

(*KMO* =.47) font size: 20 point; polarity: white on black;

Table 7.14. KMO and Bartlett's test of sphericity values (DIS, PEU and PU).

Font size, text/background polarity	Visual analogue scale	Radio button format
12 point, black on white	*KMO* = .68; Bartlett: $\chi^2(91) = 252.77$	*KMO* = .69; Bartlett: $\chi^2(91) = 213.41$
12 point, white on black	*KMO* = .52; Bartlett: $\chi^2(91) = 234.32$	*KMO* = .79; Bartlett: $\chi^2(91) = 292.05$
20 point, black on white	*KMO* = .32; Bartlett: $\chi^2(91) = 157.58$	*KMO* = .35; Bartlett: $\chi^2(91) = 186.50$
20 point, white on black	*KMO* = .47; Bartlett: $\chi^2(91) = 206.37$	*KMO* = .46; Bartlett: $\chi^2(91) = 271.05$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Nevertheless, for comparison with previous research (see Table 5.18, Section 5.4.5), factor analysis was carried out for all combinations of design parameters across both the response formats and the results are reported in Tables 7.15, 7.16, 7.17 and 7.18.

Table 7.15. Factor analysis of DIS, PEU and PU (12 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	PU	DIS	PEU	DIS	PU	PEU
DIS_Q1		.58		.61		
DIS_Q2		.60		.82		
DIS_Q3		.61		.55		
DIS_Q4		.83		.50		
DIS_Q5		.55		.86		
DIS_Q6		.68				
DIS_Q7		.63		.89		
PEU_Q1			.85			.90
PEU_Q2			.98			.95
PEU_Q3			.92			.86
PU_Q1	.91				.84	
PU_Q2	.92				.85	
PU_Q3	.99				.81	
PU_Q4	.87				.91	
Eigenvalues	4.11	3.29	2.55	4.31	3.62	1.97
% Variance	27.69	20.73	16.76	28.21	23.87	12.34
No. of items	4	7	3	7	4	3

Note: factor loadings < 0.3 are suppressed.

Table 7.16. Factor analysis of DIS, PEU and PU (12 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	DIS	PU	PEU	DIS	PU	PEU
DIS_Q1	.74			.79		
DIS_Q2	.71	-.38	-.45	1.01		
DIS_Q3	.77			.82		
DIS_Q4	.52		.33	.91		
DIS_Q5	.68	-.44		.81		
DIS_Q6	.88			.92		
DIS_Q7	.84			.63	.37	-.35
PEU_Q1			.84			.68
PEU_Q2			.78			.94
PEU_Q3			.91			.80
PU_Q1		.99			.85	
PU_Q2		.96			.94	
PU_Q3		.87			.90	
PU_Q4		.86			.95	
Eigenvalues	4.24	4.03	2.94	6.54	4.10	1.37
% Variance	28.41	27.54	19.11	45.56	28.15	1.22
No. of items	7	4	3	7	4	3

Note: factor loadings < 0.3 are suppressed.

Table 7.17. Factor analysis of DIS, PEU and PU (20 point; black on white).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	DIS	PU	PEU	DIS	PU	PEU
DIS_Q1	.88			.95		
DIS_Q2	.96			.88	-.34	
DIS_Q3	.41		-.60	.93		
DIS_Q4	.99			.37	.35	
DIS_Q5	.46		-.44	1.00		
DIS_Q6	.55	-.32		.96		
DIS_Q7	.63			.82		
PEU_Q1			1.05			1.05
PEU_Q2			1.05			.70
PEU_Q3	-.32		.64			.78
PU_Q1		1.00			.87	
PU_Q2		.89		.43	.53	
PU_Q3		.88			.90	
PU_Q4		.92			.92	
Eigenvalues	7.15	3.83	.99	6.10	3.80	1.79
% Variance	49.91	26.41	5.98	42.49	25.74	11.88
No. of items	7	4	3	7	4	3

Note: factor loadings < 0.3 are suppressed.

Table 7.18. Factor analysis of DIS, PEU and PU (20 point; white on black).

	Visual analogue scale			Radio button format		
	Factors			Factors		
Items	DIS	PEU	PU	DIS	PU	PEU
DIS_Q1	.84			.86		
DIS_Q2	.94			.44		
DIS_Q3	.44			.51		
DIS_Q4	.88			.48		
DIS_Q5	.85			.51		
DIS_Q6	.90			.96		
DIS_Q7	.77			.93		
PEU_Q1		.98				.86
PEU_Q2		.86				.69
PEU_Q3		.88				.99
PU_Q1			.60		-1.04	
PU_Q2			.92		-.82	
PU_Q3			.80		-.87	
PU_Q4			.98		-.95	
Eigenvalues	5.69	3.27	2.21	5.06	3.01	2.36
% Variance	39.18	22.03	14.33	34.19	19.96	15.25
No. of items	7	3	4	7	4	3

Note: factor loadings < 0.3 are suppressed.

7.5.5.1 Summary of the factor structure of DIS, PEU, PU

In summary, considering the factor structure of DIS, PEU and PU a clear three-factor solution similar to previous studies by van Schaik and Ling (2003, 2005, and 2007) and Ahuja and Webster (2001) was evident across all design parameters with no exceptions. The results of van Schaik and Ling (2005, 2007) were successfully replicated for all combinations of design parameters and response format.

7.5.6 Factor structure of PE

For the purpose of validating the PE construct administered in this research study, principle axis factoring with oblimin rotation was initially used for all combinations of design parameters font size and text/background polarity for both response formats radio button and visual analogue scale to detect the factor structure. However, a solution could not be obtained consistently across both response formats. Therefore, unweighted least squares factoring method was employed to determine the factor structure of the perceived enjoyment scale. The KMO and Bartlett's test of sphericity values indicated the suitability of data for factor extraction and is reported in Table 7.19.

Table 7.19. KMO and Bartlett's test of sphericity values for PE Study 2-B.

Font size, text/background polarity	Visual analogue scale	Radio button format
12 point, black on white	KMO = .64; Bartlett: $\chi 2(91) = 70.40$	KMO = .70; Bartlett: $\chi 2(91) = 128.84$
12 point, white on black	KMO = .66; Bartlett: $\chi 2(91) = 73.69$	KMO = .76; Bartlett: $\chi 2(91) = 90.86$
20 point, black on white	KMO = .60; Bartlett: $\chi 2(91) = 40.79$	KMO = .78; Bartlett: $\chi 2(91) = 40.89$
20 point, white on black	KMO = .67; Bartlett: $\chi 2(91) = 24.68$	KMO = .71; Bartlett: $\chi 2(91) = 50.15$

Note: all results of Bartlett's test of sphericity are significant $p < 0.001$.

Factor analysis from the study by Davis (1992) is provided as a reference (see Table 5.24 Section 5.4.6). The factor analysis results are reported in Tables 7.20 and 7.21 according to the response formats Likert scale using radio button and visual analogue scale for all design parameters font size and text/background polarity.

Table 7.20. Factor analysis of PE (response format: radio button format).

	12 point		20 point	
Items	Black on white	White on black	Black on white	White on black
Q1	0.99	0.96	0.97	0.86
Q2	0.98	0.98	0.98	0.99
Q3	0.76	0.92	0.95	0.86
Eigenvalues	2.64	2.82	2.87	2.63
% Variance	83.56	90.95	93.50	82.21
No. of items	3	3	3	3

Table 7.21. Factor analysis of PE (response format: visual analogue scale format).

	12 point		20 point	
Items	Black on white	White on black	Black on white	White on black
Q1	1.01	0.69	0.75	0.69
Q2	0.74	0.96	1.01	0.96
Q3	0.79	1.00	0.96	0.72
Eigenvalues	2.42	2.56	2.65	2.23
% Variance	72.68	80.25	83.82	63.69
No. of items	3	3	3	3

7.5.6.1 Summary of the factor structure of PE

In summary, the factor structure of the perceived enjoyment scale across all design parameters, was consistent and showed the same pattern as the structure reported in other studies (e.g. Davis 1992).

7.5.7 Mixed ANOVA

A mixed-measures ANOVA was conducted to test the main and interaction effects of font size, text/background polarity and response format on the total completion time of the five psychometric questionnaires (DIS, PEU, PU, PSSUQ and SUS). In addition, mixed-measures ANOVA was conducted to test the main and interaction effects of font size, text/background polarity and response format on perceived enjoyment and the workload (NASA-TLX) that respondents rated, immediately after their response to the psychometric questionnaires.

7.5.7.1 Time to complete the questionnaires

Total completion time was heavily positively skewed. A logarithmic transformation of the total completion time for the five questionnaires (DIS, PEU, PU, PSSUQ and SUS) reduced the skew and improved normality of distribution. Table 7.22 details the mean, standard deviation and the confidence intervals of the mean for the original, transformed and the retransformed time in seconds. It was observed that the mean completion time for visual analogue scale was faster than the radio button format.

Table 7.22. Descriptives for completion time of questionnaires.

	Original (time[a], all items)				Transformed (log(time))				Retransformed (exp[log(time)])			
Response format	RB		VAS		RB		VAS		RB		VAS	
Font size	12	20	12	20	12	20	12	20	12	20	12	20
mean	188.52	193.66	167.19	165.60	5.16	5.16	5.05	5.03	188.52	193.66	167.19	165.60
Std. Deviation	76.00	85.78	61.55	65.25	0.42	0.48	0.37	0.42	76.00	85.78	61.55	65.25
95% Confidence	169.99	164.19	152.18	143.19	5.05	5.00	4.97	4.88	169.99	164.19	152.18	143.19
Interval for mean	207.06	223.12	182.21	188.01	5.26	5.33	5.14	5.17	207.06	223.12	182.21	188.01

Note: RB: Likert scale using radio button; VAS: visual analogue scale; bold text: $p < 0.05$.

[a] Seconds.

A mixed 2×2×(2) ANOVA (Table 7.23) revealed that, the main effect of response format F (1, 98) = 4.68, $\eta_p^2 = 0.05$, $p < 0.05$ was significant. There were no significant other main effects or interaction effects.

Table 7.23. Mixed ANOVA summary table for completion time (Study 2-B).

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity[a]	1	0.02	0.02	0.08	0.78	0.00
Font size	1	0.01	0.01	0.04	0.84	0.00
Polarity × Font size	1	0.03	0.03	0.15	0.70	0.00
Error (Font size and Polarity)	98	20.54	0.21			
Response format	1	0.65	0.65	4.68	0.03	0.05
Response format × Polarity	1	0.00	0.00	0.01	0.91	0.00
Response format × Font size	1	0.01	0.01	0.07	0.79	0.00
Response format × Polarity × Font size	1	0.02	0.02	0.18	0.68	0.00
Error (Response format)	98	13.58	0.14			

[a] text/background polarity

7.5.7.2 Perceived enjoyment

Table 7.24 details the mean, standard deviation and the confidence intervals of the perceived enjoyment scores. The scores represent the sum of the three questionnaire items of the perceived enjoyment construct. A higher score was observed consistently for the visual analogue scale format compared to the radio button.

Table 7.24. Descriptives of perceived enjoyment scores (Study 2-B).

		Perceived enjoyment (sum of items)			
Response format		RB		VAS	
Font size		12	20	12	20
mean		10.95	11.41	14.33	15.35
Std. Deviation		6.48	7.02	4.65	4.51
95% Confidence Interval for mean	Lower Bound	9.33	8.96	13.17	13.78
	Upper Bound	12.57	13.86	15.49	16.92

Note: RB: Likert scale using radio button; VAS: visual analogue scale

A mixed 2×2×(2) ANOVA (Table 7.25) showed that the main effect of response format F (1, 94) = 17.43, $\eta_p^2 = 0.16$, $p < 0.01$ was significant. There were no other significant main effects or interaction effects.

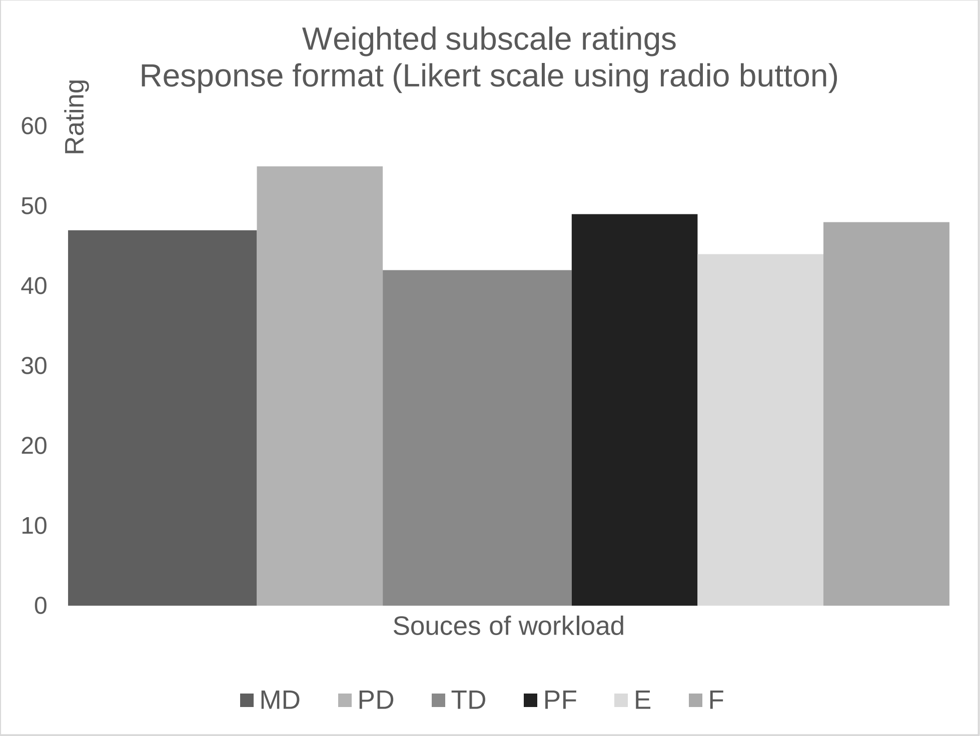
Table 7.25. Mixed ANOVA summary table for perceived enjoyment (Study 2-B).

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity[a]	1	16.52	16.52	0.54	0.47	0.01
Font size	1	13.83	13.83	0.45	0.50	0.00
Polarity × Font size	1	0.32	0.32	0.01	0.92	0.00
Error (Font size and Polarity)	94	2899.68	30.85			
Response format	1	625.00	625.00	17.43	0.00	0.16
Response format × Polarity	1	9.42	9.42	0.26	0.61	0.00
Response format × Font size	1	6.85	6.85	0.19	0.66	0.00
Response format × Polarity × Font size	1	6.82	6.82	0.19	0.66	0.00
Error (Response format)	94	3370.20	35.85			

[a] text/background polarity

7.5.7.3 Workload

A graphic representation of the weighted subscale rating for the NASA-TLX workload questionnaire is shown in Figure 7.7 for both the response formats (radio button and visual analogue scale). The X-axis represents the weight of the six subscales: mental demand (MD), physical demand (PD), temporal demand (TD), performance (PF), effort (E) and frustration (F). The Y-axis represents the workload rating of these subscales.



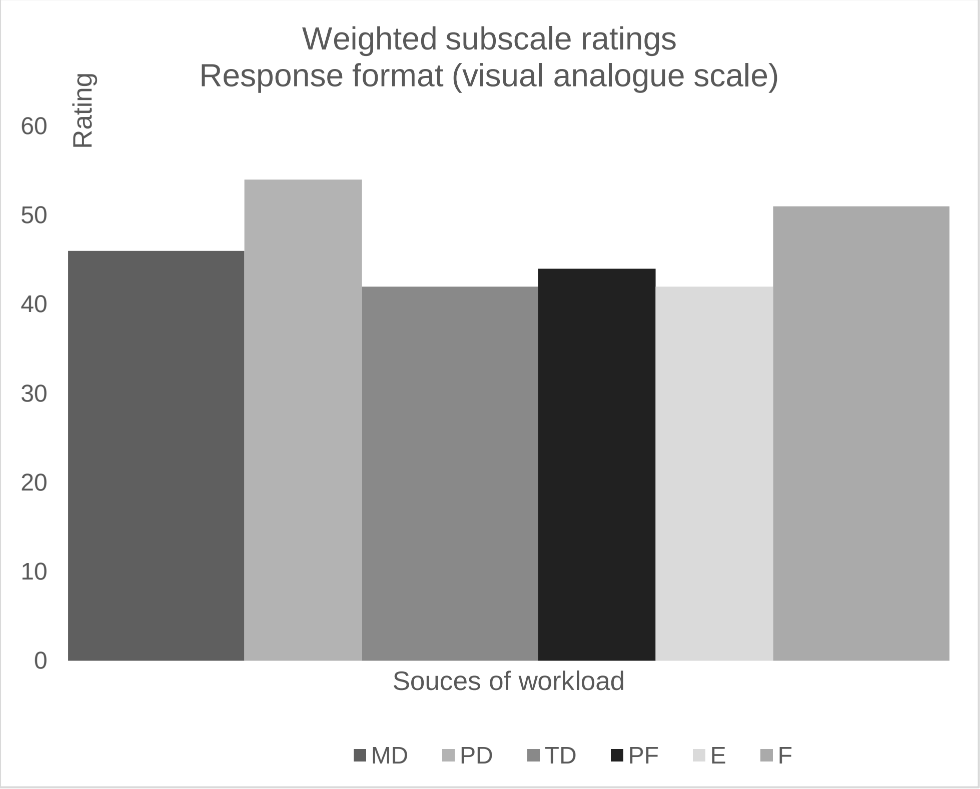


Figure 7.7. Graphic representation of weighted subscale ratings.

The descriptive statistics for the raw scores of the subscales and the weighted overall workload score is described in Table 7.26. It is observed that the overall weighted workload score was just below the midpoint of 50 for all design parameters except,

(1) font size 12 point, text/background polarity white on black, response format radio button; mean (SD) = 50.95 (14.70),

(2) font size 20 point, text/background polarity white on black, response format visual analogue scale; mean (SD) = 50.87 (21.30),

(3) font size 20 point, text/background polarity black on white, response format radio button; Mean (SD) = 49.36 (18.87),

(4) font size 12 point, text/background polarity white on black, response format visual analogue scale; mean (SD) = 40.08 (21.74).

According to the research by Grier (2015), the workload score of 50 is the midpoint. The observed workload scores that are below 50 are not remarkably below except for the score 40.08. Considering the frequency distribution of scores reported in the study by Grier (2015), the observed scores in the range of 50 is higher than 50% of the scores. Similarly, scores in the range 45 - 49 is higher than 40% of the scores. The score 40 is higher than 30% of the scores. Although the observed scores are somewhat high they are acceptable workload scores.

Table 7.26. Descriptive statistics of NASA-TLX subscales according to design parameters font size, text/background polarity and response format.

	mean (SD)							
	RB				VAS			
	12		20		12		20	
Subscales	BW	WB	BW	WB	BW	WB	BW	WB
Mental Demand	43.08 (22.08)	47.11 (26.16)	46.80 (27.32)	53.01 (30.44)	48.96 (25.64)	36.00 (23.09)	48.80 (19.51)	52.17 (26.93)
Physical Demand	49.08 (28.60)	59.91 (23.75)	62.40 (36.90)	55.51 (29.83)	58.56 (29.73)	46.71 (26.73)	50.80 (23.11)	54.68 (30.56)
Temporal Demand	42.60 (22.07)	44.62 (22.23)	34.00 (24.85)	43.83 (25.66)	38.64 (21.24)	42.28 (24.29)	38.40 (17.37)	51.13 (30.79)
Performance	49.44 (27.29)	51.73 (28.38)	33.20 (29.41)	52.17 (29.78)	38.76 (27.33)	48.18 (28.82)	44.80 (23.99)	49.75 (32.13)
Effort	41.16 (21.35)	44.09 (28.40)	54.40 (31.26)	43.62 (26.64)	41.52 (27.75)	40.98 (26.65)	41.20 (20.62)	43.83 (32.24)
Frustration	45.84 (28.22)	51.73 (28.75)	58.80 (35.99)	43.20 (30.43)	54.72 (31.10)	45.05 (28.80)	58.80 (23.70)	48.63 (31.32)
Weighted Overall workload	45.05 (15.36)	50.95 (14.70)	49.36 (18.87)	48.03 (20.88)	46.22 (17.44)	40.08 (21.74)	46.59 (19.17)	50.87 (21.30)

Note: RB: Radio button; VAS: Visual analogue scale; SD: Standard deviation; BW: Black on white; WB: White on black.

A mixed 2×2×(2) ANOVA was used to test the effects of font size, polarity and response format on respondents' workload experience. The details of the mixed ANOVA are presented in Table 7.27. No main significant effect or interaction effect was observed for the effect of font size, text/background polarity and response format on the weighted subscale scores.

Table 7.27. Mixed ANOVA summary table for NASA-TLX (Study 2-B).

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Polarity[a]	1	7.95	7.95	0.03	0.87	0.00
Font size	1	458.11	458.11	1.46	0.23	0.01
Polarity × Font size	1	26.73	26.73	0.09	0.77	0.00
Error (Font size and Polarity)	98	30685.79	313.12			
Response format	1	226.00	226.00	0.62	0.43	0.01
Response format × Polarity	1	467.18	467.18	1.28	0.26	0.01
Response format × Font size	1	413.76	413.76	1.13	0.29	0.01
Response format × Polarity × Font size	1	824.58	824.58	2.26	0.14	0.02
Error (Response format)	98	35829.01	365.60			

[a] text/background polarity

7.5.7.4 Summary of mixed ANOVA

In summary, faster completion time was observed when visual analogue scale response format was used for both font sizes. Perceived enjoyment was higher when visual analogue scale response format was used. With regard to workload, no significant results were observed for font size, text/background polarity and response format.

7.6 Discussion

In summary, this chapter presented the method and analysis to test the effect of three design parameters (font size, text/background polarity and response format) in online psychometrics on desktop computers. The hypothesis for each of these design parameters, perceived enjoyment and workload were stated. The method of the experiment to test each of the hypotheses was presented. The psychometric properties of the questionnaires were tested for each of the combinations of the design parameters. The hypotheses were tested with the mixed-measures ANOVA.

7.6.1 Psychometric analysis

7.6.1.1 Reliability analysis

The experiment reported that all questionnaires were reliable across all design parameters for both response formats (Likert scale [using radio buttons] and Likert scale [using Visual analogue scale]) (see Section 7.5.1).

7.6.1.2 Validity

Three subscales of PSSUQ exhibited strong correlation consistent with previous research (Lewis, 1995, 2002) across both response formats, thereby showing evidence of convergent validity. The remaining correlations were low, showing evidence of discriminant validity.

7.6.1.3 Factor structure

In addition to the first attempt of the translated versions of all the questionnaires being adopted, this is also the first attempt of validating the questionnaires with two response formats along with the manipulation of design parameters such as font size and text/background polarity. For the usability questionnaire PSSUQ, the factor structure established earlier by Lewis (1995, 2002) could not be replicated for. The factor structure of the SUS across certain design parameter combinations (see Section 7.5.4) was consistent with previous established studies (e.g. Sauro and Lewis, 2017). The factor structure of disorientation, perceived ease of use and perceived usefulness for all combinations of design parameters consistent with previous studies by van Schaik and Ling (2005, 2007) was evident. The factor structure of PE was consistent with previous research by Davis (1992) for all combinations of design parameters and response formats. A summary of the different extraction techniques and the rotation methods is summarised in Table 7.28.

Table 7.28. Factor extraction and rotation methods.

Questionnaire	Factor extraction method		Rotation method	
	RB	VAS	RB	VAS
PSSUQ	ULS	ULS	Oblimin	Oblimin
SUS	ULS	ULS	Oblimin	Oblimin
DIS-PEU-PU	ULS	ULS	Oblimin	Oblimin

Note: RB: Likert scale using radio button; VAS: visual analogue scale; ULS: unweighted least squares.

7.6.2 Mixed ANOVA

7.6.2.1 Completion time

The following hypotheses were developed in Chapter 3 and tested here.

Hypothesis 1: completion time decreases with increasing font size;

Hypothesis 3b: positive text/background polarity decreases completion time.

Hypothesis 4: completion time of Likert scales is shorter with radio button format than visual analogue scale format.

No significant effect of font size was observed and hence Hypothesis 1 is rejected. Hypothesis 3b is also rejected as no significant effect of text/background polarity was observed. Hypothesis 4 is rejected, as with visual analogue scale presentation, completion time was faster; this is the opposite result of the hypothesis, which states that radio button presentation is faster.

7.6.2.2 Perceived enjoyment

Hypothesis 8: perceived enjoyment increases with increasing font size;

Hypothesis 10b: positive text/background polarity enhances perceived enjoyment;

Hypothesis 11: perceived enjoyment is higher for Likert scale using radio button format than with visual analogue scale format.

No significant effect of font size and text/background polarity was evident, thus rejecting Hypothesis 8 and 10b. A significant main effect for response format on perceived enjoyment was observed. The mean scores for perceived enjoyment were higher for the visual analogue scale format thus rejecting Hypothesis 11.

7.6.2.3 Workload

Hypothesis 5: workload decreases with increasing font size;

Hypothesis 6: positive text/background polarity decreases workload;

Hypothesis 7: workload is lower with radio button format than with visual analogue scale format;

There was no evident significant main effect of font size, text/background polarity and response format on the workload scores, thus rejecting Hypothesis 5, 6 and 7.

7.7 Limitations

Given the sample size 102 participants, a post hoc power analysis indicated 100% chance (power = 1 – β = 1.0) of detecting a large and medium effect size (f = 0.40, f = 0.25) and 50% chance (power = 1 – β = 0.53) of detecting a small effect size (f = 0.10). Therefore, the sample size was deemed to be sufficient for the repeated measures ANOVA tests. The KMO values in the factor analysis of certain psychometric scales indicated that the data were not suitable for factor analysis for certain combinations of design parameters. This could be due to insufficient data collection or insufficient comprehension of the questions due to inappropriate Arabic translation. However, the rigorous process of translation and back-translation suggests that the latter is unlikely. Nevertheless, the translation and comprehension of the translated questionnaire could have limited the factor replication of the previously well-established psychometric questionnaires. With regards to sample size, Nunnally (1978) recommended having 10 times as many participants as variables; therefore the required sample size would need to be 10 × 10 variables = 100 for the SUS, 10 × 14 variables = 140 for the disorientation scale, perceived usefulness scale and perceived ease of use scale combined, and 10 × 19 variables = 190 for the PSSUQ. In addition, Tabachnick & Fidell (pg. 613, 2013) agree that ‘it is comforting to have at least 300 cases for factor analysis’. However, the number of participants was only 102 indicating sufficient sample size for factor analysis of SUS and insufficient for all other scales. Therefore, recommendations for future work include increased sample size for factor analysis (Tabachnick & Fidell, 2013) and refinement of the Arabic-language translation of the questionnaires.

Chapter 8

Conclusion and Discussion

Contents

8.1 Overview

The research questions addressed in Chapter 1 were the following:

(1) What technical system is required to support research on design parameters in online psychometrics?

(2) What are the effects of design parameters in online psychometric measurement?

(3) How can the knowledge acquired by answering Research Questions 1 and 2 be applied to guide system design?

The aim of this chapter is to (a) evaluate the technical system developed to answer research question (1), (b) summarise the results of Study 1 and Study 2 conducted to investigate and report research question (2) and (c) provide details of design guidance for online psychometrics drawn from the results of experiments detailed in this thesis to answer research question (3).

This chapter begins with a summary of the research process, followed by an evaluation of the research approach presented in Chapter 1 and presenting the design guidance for online psychometrics derived from Study 1 and Study 2. Finally, the contribution to knowledge of the research presented in this thesis is summarised.

8.2. A summary of the research

8.2.1. Literature review

As a basis for the original research presented in this thesis, a literature review was conducted of psychometrics, human-computer interaction, online psychometrics as well as tools to support research in online psychometrics. The aim of the literature review was to address Research Question 1 by identifying existing literature in online psychometrics and the design parameters in online psychometrics and to critically evaluate these by identifying research gaps. From the literature review it was evident that online psychometrics can be influenced by design factors, for example questionnaire layout (Norman et al., 2001; van Schaik & Ling, 2003, 2007; van Schaik et al., 2015) and design of response format (van Schaik & Ling, 2007). However, other design parameters that influence task completion time such as font size, response target size and response format were not addressed. Moreover, online psychometric questionnaire design on mobile devices was not addressed in the existing research.

Further, tools to support web-based psychometric research were investigated. Online survey tools are plentiful in the consumer market such as SurveyMonkey, Qualtrics, QuestionPro, among many others. These tools offer various templates and styling features (design parameters). However, a lack of separation between content, style and design of questionnaires was observed in the commercial survey tools.

Thus, due to the lack of features in the existing survey administration tools (e.g., Question pro) to support manipulation of design parameters for online psychometrics, it was decided that a new technical system must be developed for this research. This is addressed in Chapter 4 as the development of the Online Psychometric Questionnaire Design Tool (OnPQDT). The OnPQDT is a web-based tool built using PHP and MySQL programming languages. This tool provides a web-based psychometric research environment (WPRE) that enables the manipulation of design parameters, the administration of online psychometric instruments and data collection. A strength of the OnPQDT as compared to other consumer survey tools, is the flexible questionnaire design. In particular, (a) a questionnaire item can be part of multiple questionnaires, so questionnaires share items and (b) an extensible set of design parameters (e.g. font size, questionnaire layout) is supported that can be manipulated and applied to different questionnaires once created (see Section 4.5.2.4). In addition, the option to lock the zoom feature for touch-screen devices and the option to progress automatically without clicking the next button are features specifically implemented in the OnPQDT and not available as much as the author is aware in other available survey tools. After the development of the technical system was completed, several tests by the author and tests with many volunteers that are not reported in this thesis were conducted to ensure the system was stable. For the purpose of research reported in this thesis, participants responded to five psychometric questionnaires.

Next, they answered a perceived-enjoyment questionnaire and the NASA-TLX questionnaire to rate their quality of experience of answering the five psychometric questionnaires under different combinations of the manipulated design parameters. The functionalities of the OnPQDT that were used for both Study 1 and Study 2 are detailed in Table 8.1.

Table 8.1. Functionalities of OnPQDT implemented for Study 1 and Study 2.

Special features of OnPQDT implemented in this research	Study 1[a]	Study 2-A[a]	Study 2-B[b]
Single-item layout (one question-item per page)	Y	Y	Y
Zoom lock feature	Y	Y	NA
Automatic progress to the next page	Y	Y	Y

[a] touch-screen mobile device.

[b] desktop computers.

Y: Yes; NA: Not applicable.

The details of the psychometric questionnaires and the design parameters that were manipulated in the study are provided in Tables 8.2 and 8.3.

Table 8.2. Psychometric questionnaires used in this research.

Questionnaires	Study 1[a]	Study 2-A[a]	Study 2-B[b]
Disorientation (Ahuja & Webster, 1991)	Y	Y	Y
Perceived ease of use (Davis, 1992)	Y	Y	Y
Perceived usefulness (Davis, 1992)	Y	Y	Y
PSSUQ (Lewis, 1995, 2002)	Y	Y	Y
SUS (Brooke, 1996)	Y	Y	Y
Perceived enjoyment (Davis, 1992)	Y	Y	Y
NASA-TLX (Hart & Staveland, 1988)	N	Y	Y

[a] mobile device.

[b] desktop computer.

Y: Yes; N: No.

Table 8.3. Design parameters used in this research.

Design parameters	Study 1[a]	Study 2-A[a]	Study 2-B[b]
Font size	36 point; 44 point	44 point; 64 point	12 point; 20 point
Text-background colour/polarity	Black on white; Blue on white	Black on white; White on black	Black on white; White on black
Response target size	22.6 mm; 15.5 mm; 12.7 mm	22.6 mm	12.7 mm
Response format	Likert scale using: radio button	Likert scale using: radio button; visual analogue scale	Likert scale using: radio button; visual analogue scale

[a] mobile device.

[b] desktop computer.

8.2.2 Main findings of design parameters in online psychometrics

Psychometric usability questionnaires were introduced into the field of human-computer interaction in the late 1980s and early 1990s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Corbett, 1993; Lewis, 1995). There are a plethora of web tools that help create and administer these questionnaires on-line. With regards to the design of these questionnaires, several design templates with question options are contained within these web-based tools as design choices. However, there is a lack of research on the design of web-based psychometric questionnaires (van Schaik and Ling, 2007). As an attempt to fill this gap, the author conducted a study testing four design parameters: font size, text/background colour, text/background polarity, size of response option and format of the response option.

The academic-research aim was to determine the effect of the design parameters on outcomes. The practical aim was to determine, how they should be applied in online psychometrics, for example to decrease completion time while maintaining the psychometric properties of these questionnaires, when administered online. The main findings are as follows.

8.2.2.1 Font size

Study 1 and Study 2-A were conducted on mobile devices while Study 2-B was conducted on desktop computers. Study 1 tested font sizes 36 point and 44 point. Previous research studies conducted on desktop computers report that reading time is faster as font size increases (e.g., Bernard, Chaparro, Mills, & Halcomb, 2003).

In the research by Rello et al., (2016) readability and comprehension increased with increasing font size. This may be true also for small-screen devices. When larger font size leads to faster reading, completion time is faster. Empirical evidence for faster completion time when font size increases was evident from the results of Study 1 that showed font size 44 point produced a faster completion time than font size 36 point.

Thus, in Study 2-A font sizes 44 point and 64 point were examined to test if further increasing font size would further decrease completion time. However, the results showed that responses with font size 44 point were faster. Increasing font size results in more lines presented on the screen due to small-screen size and thus reading involves increased eye movements and loss of overview (Dyson, 2001 as cited in Rello et al., 2016). In the research reported by Rello et al. (2016) for desktop computers, readability continued to increase with increasing font size from 14 point and levelled off after 22 point. There was no significant effect observed for larger font size beyond 22 point and the best readability was obtained with font size 18 point. In this research, increasing the font size from 44 point to 64 point on mobile devices brought about having less text on one line, thus resulting in increased number of lines. With an increase in the number of lines, reading becomes slower thereby increasing completion time. Thus the results of Study 2A showed no significant increase in completion time with increased font size from 44 point to 64 point.

For Study 2-B, conducted on desktop computers, font sizes 12 point and 20 point were chosen in accordance to the font sizes 44 point and 64 point used in Study 2-A.

The font size 12 point was higher than the minimum font size of 10 point recommended by (Nielsen, 2002) and font size 20 point was higher than the recommended font size of 18 point by Rello et al. (2016). However, in Study 2B there was no significant effect on the completion time observed on desktop computers. The lack of an effect may have occurred because little reading was required, as psychometric items were not displayed in more than two lines of text.

8.2.2.2 Text/background colour/polarity

Often, colour combinations suitable for the desktop view are also applied on small-screen devices. The text can be well-read when colours with high contrast are used (Nielsen, 2000). In comparison with various studies that identified combinations such as blue-on-white (Ling & van Schaik, 2002; Ramadan, 2011) and black-on-white (Piepenbrock et al., 2014; Ramadan, 2011) as beneficial for accuracy, reading speed, the speed of visual search as well as user preference, in this research there was no direct significant effect on completion time, except in one instance.

In Study 1, no significant effect of text/background colour was observed. However, an advantage for negative display polarity was observed in Study 2-A when the font size was 64 point and response format was visual analogue scale. Although optimal legibility of text requires black on white (positive polarity), white on black (negative polarity) is also good (Neilsen, 2000 as cited in Hall & Hanna, 2004). In Study 2-B for text/background polarity no significant effect on completion time was observed.

The explanation for lack of significant effect could be that both design parameters choices for polarity (positive and negative) in Study 2-B and both colour contrasts (blue on white and black on white) in Study 1, were appropriate for the purpose of presenting and responding to online psychometrics. Therefore the anomalous result for 64 point/visual analogue scale would require replication in future research to provide further evidence that this is a genuine result that needs a different explanation.

8.2.2.3 Response target size

The response target size was manipulated in Study 1 with three response target sizes 12.7 mm (small), 15.5 mm (medium), 22.6 mm (large) to determine its effect on questionnaire completion time. However, response target size had no significant effect either on completion time or on perceived enjoyment. Fitts' model for motor movement (1954) defined the time required to move to a target area as function of the ratio between the distance to the target and the width of the target. Fitts' Law was not repealed with the advent of smartphone or tablets (Tognazzini, 2003). In this research, the movement time required for accurately touching the response target was not measured, but was part of the total completion time that was measured. Therefore, any effect on movement time may have been undetectable on mean completion time, given the variance of the completion time.

Moreover, the three widths of the response options were higher than the minimum width of 7 - 10 mm as recommended in various studies (e.g. Android Developers Guide, 2018; Henze, Rukzio, & Boll, 2011; Park & Han, 2010), thus reducing the number of erroneous taps outside the target size; therefore, the chosen target sizes were sufficiently wide and, possibly as a consequence, no significant effect was observed.

8.2.2.4 Response format

Research by van Schaik and Ling (2007) and van Schaik et al. (2015) reported that presentation of single items and a direct interaction mechanism (radio button) produced faster completion of online psychometric questionnaires than drop down boxes and visual analogue scales respectively, and were therefore recommended. In this research, two response formats Likert scale using radio button and Likert scale using visual analogue scale were manipulated in Study 2 and both were presented in single-item format. For survey input design, research by Stapleton (2013) recommended vertical radio buttons as the input type for all questions on mobile devices because this input type leads to less biased data and is displayed consistently on mobile devices. The study by MacKenzie, Sellen, and Buxton (1991) concluded that a dragging task was slower than a pointing task. Also, Toepoel (2017), suggested a better way to use visual analogue scale by employing the 'point-and-click' principle. Therefore, in Study 2, the response formats were designed and manipulated as vertically oriented radio buttons and visual analogue scale with point-and-click principle. For the response format radio button, all options of the Likert scale were visible. However, for response format visual analogue scale, only the end descriptors were visible with bars indicating the response options. Therefore, for the radio button response format, the respondents were able to see the appropriate option and make a choice with a touch.

However, although the visual analogue scale was implemented with the point-and-click principle, it may have been more demanding due to the limited screen size of mobile devices and unavailability of visible options like the radio button on both mobile and desktop devices.

In Study 2-A, the results for the effect of response format on completion time are mixed. Completion time with response format radio button was faster than with visual analogue scale when font size was 44 point. However, completion time with response format visual analogue scale was faster only for one specific combination of design parameters, font size 64 point and the text/background polarity white on black.

A practical explanation for the mixed results in Study 2-A can be related to the orientation of the response format, orientation of the mobile device, number of lines and the direction of the finger and eye movement as follows. On Mobile devices, scrolling is typically from up to down or down to up. Since the response format radio button was vertically oriented, aligned to the portrait orientation of the mobile device, the normal tendency to move the finger up and down was faster than the movement from left to right, required for the visual analogue scale. Also, with font size 44 point, there was no increase in the number of lines. However, when font size is 64 point, a new factor comes in namely increased number of lines to read and therefore, the tendency to co-ordinate the finger movement from right to left or left to right along with the eye movement for reading the increased number of lines from right to left (for Arabic text in this research) provides a potential explanation for the faster completion time.

A significant effect of response format on completion time was observed in Study 2-B conducted on desktop computers. Completion time was faster with visual analogue scale as the response format. This was presumably because, (a) the visual analogue scale was implemented using the point-and-click principle and (b) an input device such as the mouse was used to point-and-click on the desktop. Moreover, the screen size of the desktop, is larger with the attention of respondents being more focussed (Reeves et al., 1999). This is because, visual search in the vast amount of available space becomes difficult and therefore larger screens may compel more attention.

8.2.2.5 Effect of design parameters on perceived enjoyment

Font size, response target size and text/background colour had no significant effect on perceived enjoyment in Study 1. Similarly, font size and text/background polarity had no significant effect on perceived enjoyment in Study 2. However, response format had a significant effect on perceived enjoyment in both Study 2-A and Study 2-B. The perceived enjoyment scores for the response format visual analogue scale were observed to be higher than the response format radio button.

The reason for higher enjoyment may be that, users found the visual analogue scale implemented with the point-and-click principle more attractive than the traditional radio buttons. In addition, a red mark was placed when the user clicked on the desired location, thereby presumably increasing the enjoyment of the respondents during participation.

8.2.2.6 Effect of design parameters on workload

Workload was measured only in Study 2. Although differences in completion time was significant, workload was least sensitive to all design parameter manipulations. Answering psychometric questionnaires involves giving a spontaneous response without deliberation and this applies to all the experimental conditions formed by the combinations of design parameters. The lack of variation in workload between the conditions may be due to this spontaneous response process. The workload scores for most design parameter combinations in both Study 2-A and Study 2-B were around the midpoint score of 50 indicating an acceptable workload score. However, with vast differences in design parameter manipulations such as extremely small font size or extremely large font size, fatigue should increase and workload should ultimately show sensitivity to the differences. Also, response format visual analogue scale was implemented with the point-and-click principle that required very less effort. However, if the visual analogue scale is implemented with the drag-and-drop principle, that requires more effort, workload may show sensitivity on mobile devices and desktop devices.

8.2.2.7 General discussion on the main findings

In a broader sense, the findings could be explained as follows. With regards to font size, although there is no research in online psychometrics as much as the author is aware till date that has measured completion time based on font size, literature reports that on desktop computers, bigger font size produces better readability than smaller font sizes (e.g. Bernard et al., 2003). In addition, literature also reports research on reading speed affected by font size.

This is evident in the research reported by Abubaker and Lu (2012) on desktop computers that the mean reading time for bigger font sizes was less than for smaller font sizes. With regards to research on small-screen devices, the empirical evidence obtained from this study is that faster completion time with font size 44 point was evident in the setting of online psychometrics. In addition, differences in font size can also imply the amount of information displayed on each line. Thus, an optimal font size can be a trade-off between readability and the amount of displayable information; therefore, bigger is not always better (Abubaker & Lu, 2012). In the setting of online psychometrics, the results of this study report that 44 point font size had faster completion time than 36 point and 64 point font sizes on small screen devices. Furthermore, these results could be further extended to various other settings such as educational assessments. With variations in font sizes, at one extreme, the examinee may need to spend more time locating information or may have to do different cognitive processing (for e.g. scroll page due to large font size) and at the other, if the font size is smaller it will possibly be hard to read. Thus future research based on the results from the setting of online psychometrics could be directed to test if this in turn may affect educational test completion time and/or further affect test scores.

According to literature, the Likert scale was devised in order to measure ‘attitude’ in a scientifically accepted and validated manner in 1932 as cited in Edmondson (2005) and McLeod (2008). Moreover, Likert scale (measures human attitude) is an example of a scale in psychometrics that is used widely in the social-science and educational research (Croasmun & Ostrom, 2011).

In addition to being a well-accepted technique for attitude measurement mainly due to its simplicity and reliability, one other advantage of the Likert-type scales is that it is familiar to most respondents (ten Klooster, Visser, & de Jong, 2008). Because the respondents have been exposed to the Likert scales and are more familiar, the higher enjoyment of the Likert scale may be due to the mere-exposure effect (Zajonc, 1968; 2001). According to this effect, the repetition of an arbitrary stimulus leads to mild affection for the stimulus.

With regards to the response format, in this study, the Likert scale was implemented using the radio button and the visual analogue scale formats. In terms of perceived enjoyment, according to research by Puleston (2011 as cited in Roster, Lucianetti, & Albaum, 2015) one of the main advantages of the visual analogue scale is the respondents' enjoyment when responding to questionnaires. In support of this claim, in the current research, the mean score of perceived enjoyment in terms of response format was higher for visual analogue scale than radio button for both mobile and desktop devices. It is important to note that in the current research a variant of the traditional visual analogue scale format was implemented; instead of click-drag-drop interaction, point-and-click interaction was implemented. This choice of implementation was made due to the ease of use as reported in the research by Toepoel (2017).

In terms of completion time in the current study, the radio button format on mobile devices had faster completion time while the visual analogue scale format had faster completion time on desktop computers.

This further indicates that in general, the effect of input device plays on completion time is moderated by response format. While using the mouse to interact with the visual analogue scale on desktop computers is an advantage, using the finger to respond on small screen devices is cumbersome due to the limited screen size. However, an alternative for small screen devices could be the use of stylus with a pointed tip. While the use of visual analogue scale may not directly apply to an educational setting, the implementation of Likert scale using radio button could be of a great advantage in the educational setting.

8.3. Evaluation of research approach

The evaluation of the research reported in thesis is based on two approaches: technical innovation and empirical research evidence. Technical innovation can be seen as an instrument with necessary and positive change to an existing means. In this section, it is addressed as the development of software to incorporate research requirements identified to be lacking in currently available similar software.

8.3.1 Technical innovation

The development of the OnPQDT is a significant step as it provides a new research environment for online psychometrics (see Section 4.6). It represents innovation as it separates content, style and design of questionnaires, thus enabling independent manipulation of design parameters for online psychometrics; this functionality is required for and forms the basis of a programme of comprehensive research to study the effects of design parameters in online psychometrics.

Design includes the design parameters of visual elements that play an important role in displaying the questionnaire content such as questionnaire items and response type (see Section 4.2.1). Style on the other hand is the presentation style of online psychometrics such as sequencing of items and the navigation (see Section 4.2.2). Within OnPQDT, design and style have been implemented separately (see Figures 4.1, 4.2). The interface of OnPQDT in general, is simple and has been developed solely for research purposes. Moreover, this research environment has enabled the author to test the effects of design parameters in online psychometrics as reported in this thesis. Therefore, OnPQDT has functioned the way it was intended to for manipulating design parameters, designing and administering psychometric questionnaires, along with collecting responses.

8.3.2 Empirical research studies

The specifics of the research design dealing with control, validity and the subjects in the study are discussed. A strength of the research approach followed in this thesis lies in the systematic testing of the design parameters. For each of the selected design parameters (e.g., font size) one or more hypotheses were developed and tested in experiments using the OnPQDT that was developed. Psychometric properties such as reliability and validity of the questionnaires *Disorientation*, *perceived ease of use, perceived usefulness, perceived enjoyment* and usability (*PSSUQ and SUS*), translated into Arabic, were established. Mixed-measure of ANOVA was used to test the effect of the design parameters on completion time, perceived enjoyment and workload. Completion time in online psychometrics is important considering the fact that usually, psychometric questionnaires are lengthy.

Therefore, reducing completion time is an important consideration in online psychometrics. Similarly, since psychometric questionnaires are lengthy, it is important that respondents enjoy responding to the questionnaires spontaneously but not carelessly, with less workload. Thus, perceived enjoyment and workload also play an important role in online psychometrics.

The controlled variables within the scope of this research were *font size, response target size, text-background colour, text/background polarity and response format.* The choice of manipulations of these independent variables were based on theoretical considerations and practical considerations (e.g., Buchner & Baumgartner, 2007; Ling & van Schaik, 2002; Parhi et al., 2006; Ramadan, 2011; Van Schaik & Ling, 2007). This research has helped to identify a framework general enough to have wide applicability across the use of small-screen devices for responding to questionnaires and specifically develop a design guide for online psychometric questionnaires. The chosen manipulations were applied consistently within each experiment. By controlling the manipulation of the independent variables such as font size, text/background colour/polarity, response target size and response format, the effect of the design parameter manipulation on completion time, perceived enjoyment and workload was measured.

The experiments in this research followed a mixed-measures design, using both repeated measures and independent measures, where each participant was tested in only one condition of the between-subject independent variables. Randomization was achieved at the level of each participant for the between-subjects variables such as font size and text-background colour/polarity. The repeated measures variable in

Study 1 was the response target size and in Study 2 was the response format. The control of these independent variables was exercised at the level of the individual participant. The order of the repeated measures administration helped to balance the effect of repeatedly measuring the participants on the same variable: response target size in Study 1 and response format in Study 2. A check on the data collected from Study 1 and Study 2, indicated that there was no pattern observed in the loss of data due to participants leaving the study. During the experiment in every session, the author noted that loss of data due to technical glitches was very rare and occurred only in a few instances.

The research is observed to be high in terms of internal validity, especially when a significant effect in completion time was observed due to the manipulation of the independent variables, such as font size (Study 1 and Study 2-A) through random allocation and response format (Study 2-B) through counterbalancing of orders. Hypothesis 1, completion time decreases with increasing font size, received supportive evidence in Study 1. However, it was rejected in Study 2-A, as completion time increased with the increasing font size.

Hypothesis 4, completion time of Likert scales is shorter with radio button format than visual analogue scale format, was rejected in Study 2-B, as faster completion time was observed for the visual analogue format. Hypothesis 11, perceived enjoyment is higher with Likert scale using radio button format than with visual analogue scale was rejected, as the perceived enjoyment score of the respondents was higher for the visual analogue scale in both Study 2-A and Study 2-B.

A threat to the internal validity of the study was observed through the variable mobile device with which the participants responded in Study 1 and Study 2-A. This variable remained uncontrolled. The participants used their own mobile device to participate in the study and therefore, the device was identified as the confounding variable. Every model of the mobile device has different specifications such as processor speed and this may have an effect on the task completion time. However, in Study 2-B, the experiment was conducted on desktop computers in the University's computer lab where all the devices had identical software and hardware configurations. It was not within the scope of the author's budget to control the confounding variable in Study 1 and Study 2-A by providing devices for participation in the experiment. For this reason, the results of the experiments conducted on small-screen devices are not compared against the results of the desktop computers where the device was not a confounding variable.

This research is low on external validity since Study 1 and Study 2 was conducted within only one university in Kuwait and there was no participation outside of this university. Thus the results cannot be generalized beyond the population of this university.

This research has high ecological validity for Study 1 and Study 2-A with regards to the mobile device. Unlike desktop computers, social media popup messages and email notifications appear on mobile devices. This was not controlled during the experiment. However, instructions were provided by the author before the start of the experiment that although social media and email notifications may appear, respondents must continue in the Study. As much as the author is aware, there were no instances of participants leaving the online psychometric questionnaire webpage and returning due to the social media and email notifications.

The construct validity of this research is established in a much broader sense through the quality of the psychometric properties of the questionnaires. Factor analysis was used as the tool to determine the validity of the psychometric questionnaires (Stapleton, 1997). The factor structure of most questionnaires was not satisfactory and did not reproduce existing solutions from previous research, with the exception of the questionnaires to measures disorientation, perceived ease of use and perceived usefulness in Study 2-B. In addition, reliability values were established for the five psychometric questionnaires according to different combinations of design parameters. In general, the questionnaires exhibited acceptable reliability values except in rare instances for certain combinations of design parameters for both Study 1 and Study 2. Threats to construct validity due to the translation process is observed and explained in the limitations of the study (see Section 8.5.2.3).

8.4. Design guidance

The results from the mixed-measures experiments reported in this thesis, now provide evidence for design guidance in online psychometrics specifically with regards to completion time, perceived enjoyment and workload. In order to address Research Question 3, a summary is provided in Tables 8.4, 8.5 and 8.6 and the following sections present a discussion of the applicability of the findings from the studies to the evaluation and (re)design of online psychometric questionnaires.

Table 8.4. Completion time for online psychometrics on Mobile devices (Study 1; Study 2-A).

Design	Study			
Parameter	Study 1	Study 2-A	Result	Design guidance
Font size (point)	36;44	44;64	Consistent faster completion with 44 point. Medium to small effect sizes were observed.	44 point should be chosen over 36 point and 64 point.
Response target size (mm)	12.7;15.5; 22.6	15.5	No significant effect observed	Target sizes 12.7mm, 15.5mm, 22.6mm are equally suitable.
Text/ background colour	black/blue on white	NA	No effect	Black on white and Blue on white text/background colours are equally suitable.
Text/ background polarity		black on white; white on black	Faster completion time observed for negative polarity in combination with font size 64 point and visual analogue scale response format.	Tentatively, negative polarity should be applied with 64 point and visual analogue scale response format.
Response format (Likert scale)	Radio button	Radio button; Visual analogue scale (point-and-click format)	In Study 1, faster completion time was observed for response format radio button with font size 44 point. In Study 2-A, faster completion time was observed for response format visual analogue scale with font size 64 point and negative polarity.	Tentatively, response format radio button should be chosen with font size 44 point and response format visual analogue scale to be chosen with font size 64 point and negative polarity.

NA : Not applied.

Table 8.5. Completion time for online psychometrics on Desktop devices (Study 2-B).

Design Parameter		Result	Design guidance
Font size (point)	12; 20	No significant effect observed.	Tentatively, 12 point and 20 point are equally suitable
Response target size (mm)	12.7	Response target size was not a repeated measure.	No guidance can be provided
Text/background polarity	Black on white; white on black	No significant effect observed.	Tentatively, no guidance can be provided.
Response format (Likert scale)	Radio button; Visual analogue scale (point-and-click format)	Faster completion time was observed with response format visual analogue scale implemented using the point-and-click format.	Likert scale using visual analogue scale with point-and-click format should be chosen over radio button.

Table 8.6. Perceived enjoyment and workload for online psychometrics on Mobile devices and Desktop computers.

Design Parameter	Study 2-A (Mobile)	Study 2-B (Desktop)	Results	Design guidance
Font size (point)	44;64	12; 20	PE, WL: No significant effect observed	PE, WL: Tentatively, font sizes are equally preferable
Response target size (mm)	15.5	12.7	PE, WL: No significant effect observed	PE, WL: Tentatively, response target sizes are equally preferable
Text/ background polarity	Black on white; white on black		PE, WL: No significant effect observed	PE, WL: Tentatively, positive and negative polarity are equally preferable
Response format (Likert scale)	Radio button; Visual analogue scale (point-and-click interaction)		PE: Higher perceived enjoyment scores for response format visual analogue scale than radio button was consistently observed in both Study 2-A and Study 2-B WL: No significant effect observed	PE: Visual analogue scale with point-and-click interaction should be chosen for administration of Likert scales over the radio button format. WL: Tentatively, both response formats are equally preferable

PE: Perceived enjoyment; WL: Workload.

8.4.1 Completion time

8.4.1.1 Font size

The values for design parameter font size in the studies were as follows:

1. Study 1 (mobile): 36 point; 44 point;

2. Study 2-A (mobile): 44 point; 64 point;

3. Study 2-B (desktop): 12 point; 20 point.

Results of Study 1 and Study 2-A indicated that completion time on the mobile device was significantly affected by font size. Font size 44 point led to faster completion of questionnaires consistently in both Study 1 and Study 2-A. The effect sizes observed in both Study 1 and Study 2-A occurred with medium to small effect sizes.

For desktop computers, various research studies provide evidence that bigger font size is better (e.g., Rello et al., 2016). This is true for online reading that requires the comprehension of large texts. However, in online psychometrics, each questionnaire item is typically not longer than two lines and hence reading speed presumably does not affect the response time a great deal. Study 2-B investigated the effect of font size (12 point and 20 point) for online psychometrics on desktop computers. However, design advice cannot be offered for desktop computers since font sizes smaller than 12 point and bigger than 20 point were not investigated.

Design guidance: currently, based on this study, in terms of reducing completion, 12 point and 20 point font sizes are both equally suitable for desktop computers and font size 44 point should be chosen over 36 point and 64 point for mobile devices specifically in online psychometrics.

8.4.1.2 Text/background colour and polarity

The values for design parameter text/background colour in the studies were as follows:

1. Study 1 (mobile): black on white; blue on white;

2. Study 2-A (mobile): black on white; white on black;

3. Study 2-B (desktop): black on white; white on black.

There was no effect of text/background colour combination or text/background polarity observed for completion time in Study 1 and Study 2-B. However in Study 2-A, when text/background polarity was white on black, faster completion time was observed for one combination of design parameters font size 64 point and response format visual analogue scale.

Tentative *design guidance*: in terms of reducing completion time, negative polarity when applied with font size 64 point for the response format visual analogue scale should be chosen over positive polarity on mobile devices. However, because the advantage of negative polarity occurred only for a specific combination of font size and response format, further research is recommended before more definitive guidance can be given.

8.4.1.3 Response target size

No significant effect was observed for response target size in Study 1. However, it is to be noted that the setting of the study was among university students, who are very familiar using small-screen devices. Moreover, the width of the response target size used in Study 1 (12.7 mm/15.5 mm/22.6 mm) was higher than 10 mm recommended in various research (e.g., Android Developers Guide, 2018; Henze et al., 2011; Park, Han, Park, & Cho, 2008) to promote accurate touching. However, it is important to test the effect of response target size among various age-groups of people, including people with dyslexia and other challenges.

Design guidance: in terms of reducing completion time, response target sizes of 12.7 mm/15.5 mm/22.6 mm are equally preferable on small-screen devices. However, for desktop computers, 12.7 mm is preferable; therefore, design considerations other than completion may drive the choice of target size.

8.4.1.4 Response format

A point-and-click format for the visual analogue scale was implemented in Study 2. In Study 2-B on desktop computers, faster completion time was observed for the visual analogue scale compared to the radio button. In addition, the use of the input device such as the mouse contributes to the faster completion time with the point-and-click principle on desktop computers (MacKenzie et al., 1991). Therefore Hypothesis 4, completion time of Likert scales is shorter with radio button format than visual analogue scale format, was rejected for presentation on desktop computers.

For mobile devices, the results were more complex. An interaction effect for response format was significant in Study 2-A that was conducted on mobile devices. Faster completion time was observed for response format radio button when font size was 44 point. Also, faster completion time was observed for response format visual analogue scale when font size was 64 point and text/background polarity was white on black. Therefore, based on the results from Study 1, Hypothesis 4, completion time of Likert scales is shorter with radio button than visual analogue scale format was partially supported only when font size 44 point was used. However, based on the results from Study 2-A, Hypothesis 4, completion time of Likert scales is with shorter with radio button that visual analogue scale format was partially rejected, as response format visual analogue scale was faster for one combination of design parameters, font size 64 point and text/background polarity white on black.

It is observed that visual analogue scale was slightly faster for only one combination of design parameters in Study 2-A, and was definitely faster in Study 2-B. Thus the design guidance of the ‘point-and-click’ principle (Toepoel, 2017) applied for the response format visual analogue scale provided the advantage for faster completion time partially on mobile devices and consistently on desktop computers, compared to the traditional click-and-drag format particularly in online psychometrics.

Design guidance: in terms of reducing completion time, on desktop computers, Likert scale using visual analogue with point-and-click principle should be chosen as the response format over radio button. However, on mobile devices, the tentative design guidance is that radio button should be chosen over visual analogue with font size 44 point and visual analogue scale should be chosen when font size is 64 point and text/background polarity is white on black. Yet, because of the mixed results observed in Study 2-A for the design parameter combinations of response format with a specific font size and polarity, further research is recommended before more definitive guidance can be given.

8.4.2 Perceived enjoyment

Pertaining to perceived enjoyment, only one design parameter response format was observed to have significant effect. The significant effect of response format was consistently observed in both Study 2-A and Study 2-B. Likert scale using visual analogue format produced higher perceived enjoyment scores than the radio button format consistently in Study 2. Thus, Hypothesis 11, perceived enjoyment is higher with Likert scale using radio button format than with visual analogue scale, was rejected. Moreover, the point-and-click principle (Toepoel, 2017) implemented for the design of the visual analogue scale contributed to the advantage of perceived enjoyment both on mobile device and desktop computers.

Design guidance: in terms of increasing perceived enjoyment, the point-and-click format visual analogue format should be chosen for the administration of Likert scales over the radio button format.

8.4.3 Workload

No specific design guidance with regards to font size, response target size, text-background colour/polarity and response format can be given as no empirical research evidence was significant in the results of Study 2. Therefore, in terms of reducing workload the manipulations of response format, polarity and font size are equally suitable for the administration of psychometrics on mobile devices and desktop computers.

8.5 Limitations

8.5.1 Technical innovation

The author was the sole user of the OnPQDT. A first limitation observed was that, although multiple administrators/researchers can be added, there is no individual memory/workspace allocated.

Second, Arabic script is written from right to left. When the main script is Arabic, the layout and structure of pages and documents are also set from right to left. However, currently in the OnPQDT, although the main text in Arabic was displayed from right to left, the structure and layout of the page was presented from left to right. Participants were briefed on this limitation by the author before the start of the experiment in each study and therefore they were able to complete the experiment in spite of the limitation.

8.5.2 Empirical research evidence

8.5.2.1 KMO values

The KMO values for certain combinations of design parameters and questionnaires indicated the unsuitability of data for factor analysis. According to Kaiser (1974), KMO values below 0.5 are unacceptable. Such low unacceptable KMO values were observed for certain cases in both Study 1 and Study 2. The inability to replicate the factor structure of the established questionnaire (e.g. PSSUQ) could be one of the implications of low KMO values.

Replication of factor structure gives practitioners in human-computer interaction and usability increased confidence in using the questionnaire (e.g., Lewis 2002, 2018). In this study, among all the five psychometric questionnaires, the factor structure of PSSUQ questionnaire was not replicated in both Study 1 and Study 2 (see Sections 5.4.3, 6.6.3 and 7.5.3). The factor structure of SUS was replicated in both Study 1 and Study 2 except for certain combinations of design parameters (see Sections 5.4.4, 6.6.4 and 7.5.4).

8.5.2.2 Screen size

Questionnaires *disorientation, perceived ease of use* and *perceived usefulness* resulted in a three factor structure consistently in Study 2-B conducted on desktop computers for all combinations of design parameters. However, in Study 1 and Study 2-A, conducted on mobile devices, the three-factor structure was evident only for certain combinations of design parameters. This difference may be due to the limitation of the screen size for mobile devices.

Small screens may pose problems for both human perception and visual attention (Chen et al., 2003; Kim & Kim, 2012; Maniar, Bennett, Hand, & Allan, 2008). These research studies were based on video-based/multimedia learning and image displays on small-screen devices. For example, small screens often show information with limitations (e.g., distortions in brightness, colour, font, and spacing between characters, lines, and words) compared to the large screen; this makes it more difficult to perceive the information that is presented. Larger screens contribute to greater attention because, people tend to pay more attention when they receive a message on a large screen (Reeves et al., 1999).

8.5.2.3 Arabic translation

The translation of the questionnaires may have also been a factor for the non-replication of factor structures. It is important that meanings are not lost during the translation process. One of the effects of translation could be that the factor structure in psychometric analysis is affected when items load on the same factor due to lack of precision in the translated language. Moreover, if the same meaning is not conveyed across participants consistently, there may be threats to construct validity. This is because in terms of meaning the same items, when translated, the words may not map onto the same underlying construct as the original items before translation. Another consequence of translation with regards to completion time is that, there might arise a situation when too much time is consumed to merely understand the translated item. Otherwise, a lack of understanding may result in faster completion time. However, these implications may be less likely, due to the rigorous translation process that was followed for this research.

Nevertheless, the translation process focused on the literal equivalence of the translated items (in particular the item stems) with the items before translation, but did not consider the underlying constructs that the items represent.

8.5.2.4 Data collection and analysis

A limitation observed was that the results of the study are largely based on one type of device the iPhone. Therefore, the results may only generalise to psychometric measurement on iPhones. In addition, a procedural issue in data collection was the lack of participant profile data in terms of last access to the evaluation site (PeopleSoft-SIS in Study 1 and Moodle based LMS in Study 2). The participants were not asked to use the system as part of the experiment before answering the psychometric tests. Therefore, a limitation was the lack in precise recall of the experience leading to a less accurate description of their evaluation. Further, it was not the objective of the study to statistically compare the results from mobile devices and desktop computers. Therefore, a repeated measures study of desktop vs mobile was not designed.

Furthermore, with regards to psychometric data analysis of the questionnaires, a confirmatory factor analysis was not conducted. This is because, the questionnaires used in the study did not exist in Arabic and hence the first step was to validate the factor structure of the translated questionnaires using the exploratory factor analysis method and compare the resulting factor structure with the factor structure of the English versions of the questionnaire.

In addition, confirmatory factor analysis should be run using a data set different from the data set used for the exploratory data analysis, such that the validity of the exploratory factor analysis structure found as a result of the exploratory factor analysis will be shown by using confirmatory approach with a different data set (Orcan, 2018). However, the data collected for the Study was scarcely sufficient for the exploratory factor analysis thus limiting the approach to implement a confirmatory factor analysis.

8.6 Future work

8.6.1 Technical innovation

First, future work should include the facility for more than one researcher/administrator to be registered with individual memory/workspace. Second, help and documentation should be developed and provided within the tool to assist new researchers. Third, the design guidance that has been developed (see Section 8.4) should be provided on the interface of the tool where design parameters are manipulated as recommendations to researchers. It is to be noted that OnPQDT will remain a tool for research unlike many other commercial tools.

The use of computers for online testing has increased considerably with the recognition of the vast potential for computer-based tests (Gikandi, Morrow, & Davis, 2011; Leeson, 2006). In a technology-enriched environment, good interface designs for online testing both on small-screen and desktop computers should be researched and reported (Fulcher, 2003; Nicol, 2007).

Therefore, an extended possible application of the concept of OnPQDT lies in online educational assessment, specifically design parameters for online assessments. For example, the choice of font size, when very large, can affect cognitive processing and may require the examinee to spend more time locating the information. Alternatively, if the text is very small, it will become possibly difficult to read, thus requiring more completion time than what is allocated. In addition, item presentation can play an important role; for example hurried responses may occur when all items are presented whole-form (Hofer & Green, 1985; van Schaik & Ling, 2007; van Schaik et al., 2015). Therefore, design parameters such as font size, item presentation may affect test completion times (according to the results presented in this thesis) and test scores (Bridgeman, Lennon and Jackenthal, 2003). An empirical evidence is required to provide design guidance for online educational testing. As much as the author is aware, there is little research that addresses design parameters for educational online assessments directly and systematically.

A relevant study is the research conducted by Bridgeman, Lennon and Jackenthal (2003). A main issue of interest presented by the authors was the variation in legibility of information presented on the screen due to the size, resolution and various font settings. For the purpose of the study the researchers investigated and reported only the effects of screen size, screen resolution, and display rate on computer-based test scores. The findings showed that test scores were affected by screen-size and screen resolution.

Research on presentation of text for people with dyslexia also exist specifically for font type and this needs to be applied to online assessments to test the effect of font type on test scores (e.g., Rello & Baeza-Yates, 2017). In addition, other design parameters such as text/background colour, line-length and questionnaire presentation layout can be applied to online assessments to test the effect of design parameters on test performance. The web-based OnPQDT can provide the research environment for this.

Last but not the least, the interface of the OnPQDT must be enhanced to support the right to left orientation of the Arabic language to support research among the Arab speaking population. The OnPQDT must also be able to administer questionnaires in Arabic by following the appropriate text orientation of the language. Enabling features to be culture specific can directly impact user performance, thus merging culture and usability (Barber & Badre, 1998; Reinecke & Bernstein, 2011).

8.6.2 Empirical research evidence

Naturally, given the time and resources available in this project, the set of design parameters manipulated and reported in this thesis in terms of empirical research evidence is not comprehensive. Many other parameters have been identified and included within the tool for manipulation and are available to conduct numerous studies. Among the many design parameters, two examples for design parameters applicable in online psychometrics are font type and line spacing.

Along with font size, font type plays an important role for visual interface in human-computer interaction in terms of reading speed and visual search (Banerjee, Majumdar, Pal, & Majumdar, 2011; Ling & van Schaik, 2006). In addition to font type, a significant effect on task performance, with wider line spacing leading to better accuracy and to faster reaction times was reported in the research by Ling & van Schaik (2007). A reaction may imply the action of responding to questionnaires in online psychometrics. Therefore, faster reaction time may imply quicker responses to questions, leading to faster completion time. Faster completion is specifically important in online psychometrics; therefore design parameters such as font type and line spacing should be considered in future research.

Additionally, another design parameter manipulation is the different text/background colour combination for clear distinction between questions and answers. Existing research on the effect of colour on visual search tasks has demonstrated that higher contrasts between text and background colour led to faster searching (e.g., Bhattacharyya, Chowdhury, Chatterjee, Pal, & Majumdar, 2014; Ling & van Schaik, 2002). Research by Ko (2017) reported that colour combinations may play an even more important role than luminance contrast in the overall legibility. Also, it was reported that search time corresponded to highest legibility. Although these research studies were not in online psychometrics, the importance still applies because different colour combination for questions and answers can distinguish the answers from questions and thereby potentially result in reduced search time resulting in faster questionnaire completion.

Design parameter manipulations can make text legible and thereby lead to faster reading (e.g. Grobelny & Michalski, 2015; Rello, Pielot, & Marcos, 2016), and impact completion time.

8.7 Summary of contribution to knowledge

The contribution to knowledge reported in this thesis is summarised here.

8.7.1 The development of a novel tool for research in online psychometrics

Research Question 1 (Section 8.1) was addressed by the development of OnPQDT solely for research in online psychometrics. OnPQDT is a novel tool that enables the manipulation of design parameters, the creation of online questionnaires, questionnaire administration and data collection. With a plethora of survey tools available in the consumer market, OnPQDT is designed to be used solely by researchers. The separation between content, style and design of questionnaires in OnPQDT (see Chapter 4) enables the manipulation of design parameters for online psychometrics, thereby supporting research to provide empirical research evidence-based design guidance. In addition, the extensible database system allows the set of design parameters to be extended. Moreover, the system provides the platform to use OnPQDT for research in similar fields such as online educational testing.

8.7.2 The effect of design parameters on human-computer interaction outcomes

Hypotheses developed for the chosen design parameters manipulated in Study 1 and Study 2 were systematically and successfully tested (see Chapters 5, 6 and 7). The results obtained helped to address Research question 2 (Section 8.1) regarding the effect of design parameters in online psychometrics. The systematic testing involved the mixed-measures analysis of variance tests of the hypotheses. This was in addition to the systematic testing and reporting of the psychometric properties of the translated questionnaires according to the different design parameter combinations. These questionnaires were to measure disorientation, perceived ease of use, perceived usefulness and usability questionnaires (such as PSSUQ and SUS). The systematic testing and reporting of the effects of design parameters brought about the development of new knowledge in online psychometrics with regard to these parameters in relation to completion time, perceived enjoyment and workload. The design parameter font size was manipulated with different values and significant effects were observed in completion time. Furthermore, the manipulation of the response format produced significant effects on both completion time and perceived enjoyment. In a particular instance, an interaction effect was also observed between font size and response format. Together these results contribute new knowledge about the effect of design parameters in online psychometrics.

8.7.3 Design guidance

Based on the new knowledge obtained from the effect of design parameters in online psychometrics, design guidance was developed. Thus Research Question 3 is addressed and contributes to the development of design guidance for online psychometrics (Section 8.4). Design guidance with regards to font size, response format and polarity specifically in online psychometrics has been developed in this research. In particular specific design guidance was produced regarding the design parameters font size, polarity, response target size and response format in terms of reducing completion time. Furthermore, specific design guidance was produced regarding the design parameter response format in terms of increasing perceived enjoyment.

8.7.4 Psychometric properties

Systematic evaluation of the psychometric properties of the questionnaires according to the combination of different design parameters indicated an acceptable level of reliability for most combinations of the design parameters except in specific cases. Factor analysis consistent with a structure reported for the standard questionnaires: SUS*, Disorientation, Perceived ease of use* and *Perceived usefulness* was evident for most design parameter combinations except in specific cases. However, the factor structure for PSSUQ was not replicated in either Study 1 or Study 2.

8.8 Final words

This chapter summarised the research phases presented in this thesis and their findings, as well as the limitations of the studies and suggestions for future work. The practical implications of findings of the studies were discussed. Next, these findings were used to guide design decisions. Finally, the contribution to knowledge of the research project reported in this thesis was summarised.

The starting point of this project was to apply human-computer interaction and interaction-experience knowledge to the field of online psychometrics on both small-screen and desktop computers. It can be concluded that the main research questions of the thesis were addressed and answered by the development of the OnPQDT and the two empirical studies. The results of the studies presented in this thesis can be used to further develop the testing of further design parameters in online psychometrics and provide guidance for design.

References

Abubaker, A. A., & Lu, J. (2012). The optimum font size and type for students aged 9-12 reading Arabic characters on screen: A case study. *Journal of Physics: Conference Series*, *364*(1). https://doi.org/10.1088/1742-6596/364/1/012115

Ahuja, J., & Webster, J. (2001). Perceived disorientation: an examination of a new measure to assess web design effectiveness. *Interacting with Computers*, *45*(2001), 31–33. Retrieved from http://www.sciencedirect.com/science/article/pii/S0953543801000480

AlGhannam, B. A., Albustan, S. A., Al-Hassan, A. A., & Albustan, L. A. (2017). Towards a Standard Arabic System Usability Scale: Psychometric Evaluation using Communication Disorder App. *International Journal of Human-Computer Interaction*, *00*(00), 1–6. https://doi.org/10.1080/10447318.2017.1388099

Andrews, D., Nonnecke, B., & Preece, J. (2003). Conducting Research on the Internet : Online Survey Design , Development and Implementation Guidelines. *International Journal of Human-Computer Interaction*, *16*(2), 185–210. https://doi.org/10.1207/s15327590ijhc1602_04

Android Developers Guide. (2018). App Widget Design Guidelines, 1–9. Retrieved from https://developer.android.com/guide/practices/ui_guidelines/widget_design.htm

Aranyi, G., Schaik, P. van, & Barker, P. (2012). Using think-aloud and psychometrics to explore users' experience with a news Web site. *Interacting with Computers*, *24*(2), 69–77. https://doi.org/10.1016/j.intcom.2012.01.001

Banerjee, J., & Bhattacharyya, M. (2011). Selection of the optimum font type and size interface for on screen continuous reading by young adults: an ergonomic approach. *Journal of Human Ergology*, *40*(1–2), 47–62.

Banerjee, J., Majumdar, D., Pal, M. S., & Majumdar, D. (2011). Readability, Subjective Preference and Mental Workload Studies on Young Indian Adults for Selection of Optimum Font Type and Size during Onscreen Reading. *Al Ameen Journal of Medical Sciences*, *4*(2), 131–143. Retrieved from http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=59806916&site=ehost-live

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447310802205776

Barber, W., & Badre, A. (1998). Culturability: The Merging of Culture and Usability. In *Proceedings of the 4th Conference on Human Factors and the Web* (pp. 1–14).

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A method for CATI and a new look at non attitudes. *Public Opinion Quarterly*, *55*(3), 331–346. https://doi.org/10.1086/269265

Bassili, J. N., & Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly*, *60*(3), 390. https://doi.org/10.1086/297760

Battiste, V., & Bortolussi, M. (1988). Transport Pilot Workload: A Comparison of Two Subjective Techniques. In *Proceedings of the Human Factors Society* (pp. 150–154). https://doi.org/10.1177/154193128803200232

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human Computer Studies*, *59*(6), 823–835. https://doi.org/10.1016/S1071-5819(03)00121-6

Bevan, N., Carter, J., & Harker, S. (2015). ISO 9241-11 revised: What have we learnt about usability since 1998? In *Lecture Notes in Computer Science* *(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-20901-2_13

Bhattacharyya, D., Chowdhury, B., Chatterjee, T., Pal, M., & Majumdar, D. (2014). Selection of character/background colour combinations for onscreen searching tasks: An eye movement, subjective and performance approach. *Displays*, *35*(3), 101–109. https://doi.org/10.1016/j.displa.2014.03.002

Birnbaum, M. (2004). Human research and data collection via the Internet. *Annu. Rev. Psychol.*, *55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Bonnardel, N., Piolat, A., & Le Bigot, L. (2011). The impact of colour on Website appeal and users' cognitive processes. *Displays*, *32*(2), 69–80. https://doi.org/10.1016/j.displa.2010.12.002

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495. https://doi.org/10.1080/10447318.2015.1064648

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale : A test of alternative measurement models On the dimensionality of the System Usability Scale : a test of alternative measurement models. *Cognitive Processing*, *9*(July), 193–197. https://doi.org/10.1007/s10339-009-0268-9

Bradley, C. (2013). *Handbook of psychology and diabetes: a guide to psychological measurement in diabetes research and practice*. *Routledge*. https://doi.org/https://doi.org/10.4324/9781315077369

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing*, *6*(3), 188–205. https://doi.org/10.1007/s007790200019

Bridgeman, B., Lennon, M. Lou, & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/S15324818AME1603_2

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x

Brock, R., Barry, R., Lawrence, E., Dey, J., & Rolffs, J. (2012). Internet Administration of Paper-and-Pencil Questionnaires Used in Couple Research Assessing Psychometric Equivalence. *Assessment*, *19*(2), 226–242. Retrieved from http://asm.sagepub.com/content/19/2/226.short

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*. https://doi.org/10.1002/hbm.20701

Brooke, J. (2013). SUS : A Retrospective. *Journal of Usability Studies*, *8*(2), 29–40. https://doi.org/10.1074/jbc.R115.675280

Brown, & Venkatesh. (2005). Model of Adoption of Technology in Households: A Baseline Model Test and Extension Incorporating Household Life Cycle. *MIS Quarterly*, *29*(3), 399. https://doi.org/10.2307/25148690

Buchanan, T, & Smith, J. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125–144. Retrieved from http://onlinelibrary.wiley.com/doi/10.1348/000712699161189/abstract

Buchanan, Tom. (2002). Online Assessment: Desirable or Dangerous? *Professional Psychology: Research and Practice*. https://doi.org/10.1037/0735-7028.33.2.148

Buchner, A., & Baumgartner, N. (2007). Text - Background polarity affects performance irrespective of ambient illumination and colour contrast. *Ergonomics*, *50*(7), 1036–1063. https://doi.org/10.1080/00140130701306413

Buchner, Axel, Mayr, S., & Brandt, M. (2009). The advantage of positive text-background polarity is due to high display luminance. *Ergonomics*. https://doi.org/10.1080/00140130802641635

Buskirk, T. D., Saunders, T., & Michaud, J. (2015). Are Sliders Too Slick for Surveys? An Experiment Comparing Slider and Radio Button Scales for Smartphone, Tablet and Computer Based Surveys. *Methods*, *9*(92), 229–260. https://doi.org/10.12758/mda.2015.013

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. Retrieved from http://psycnet.apa.org/journals/bul/56/2/81/

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods*, *41*(1), 113–117. https://doi.org/10.3758/BRM.41.1.113

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. https://doi.org/10.1016/0003-6870(84)90205-9

Carlbring, P., Brunt, S., & Bohman, S. (2007). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human …*, *23*(3), 1421–1434. Retrieved from http://www.sciencedirect.com/science/article/pii/S0747563205000415

Carroll, J. M., & Carrithers, C. (1984). Training wheels in a user interface. *Communications of the ACM*, *27*(8), 800–806. https://doi.org/10.1145/358198.358218

Cebi, S. (2013). Determining importance degrees of website design parameters based on interactions and types of websites. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2012.10.036

Chen, L. Q., Xie, X., Fan, X., Ma, W. Y., Zhang, H. J., & Zhou, H. Q. (2003). A visual attention model for adapting images on small displays. *Multimedia Systems*, *9*(4), 353–364. https://doi.org/10.1007/s00530-003-0105-4

Chen, P. P.-S. (1976). The entity-relationship model---toward a unified view of data. *ACM Transactions on Database Systems*. https://doi.org/10.1145/320434.320440

Chen, P. P.-S. (2002). The Entity Relationship Model --- Toward a Unified View of Data. In *Software Pioneers: Contributions to Software Engineering*. https://doi.org/10.1007/978-3-642-59412-0_18

Chin, J. P., Diehl, V. a, & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *CHI '88- Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/57167.57203

Clark, L. A., & Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Coles, M., Cook, L., & Blake, T. (2007). Assessing obsessive compulsive symptoms and cognitions on the internet: evidence for the comparability of paper and Internet administration. *Behaviour Research and Therapy*, *45*(9), 2232–2240. Retrieved from http://www.sciencedirect.com/science/article/pii/S0005796707000022

Colman, A. M. (2009). A Dictionary of Psychology (3rd ed.). *Oxford University Press*. https://doi.org/10.1093/acref/9780199657681.001.0001

Copping, L. T., Campbell, A., & Muncer, S. (2014). Psychometrics and life history strategy: the structure and validity of the High K Strategy Scale. *Evolutionary Psychology : An International Journal of Evolutionary Approaches to Psychology and Behavior*, *12*(1), 200–222. https://doi.org/10.1177/147470491401200115

Copping, L. T., Campbell, A., Muncer, S., & Richardson, G. B. (2017). The psychometric evaluation of human life histories: A reply to Figueredo, Cabeza de Baca, Black, Garcia, Fernandes, Wolf, and Woodley (2015). *Evolutionary Psychology*, *15*(1), 1–14. https://doi.org/10.1177/1474704916663727

Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.78.1.98

Couper, M. P. (2008). *Designing Effective Web Surveys*. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511499371

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230–253. https://doi.org/10.1086/322199

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-Type Scales in the Social Sciences. *Environmental and Resource Economics*, *40*(1), 19–22.

Darroch, I., Goodman, J., Brewster, S., & Gray, P. (2005). The effect of age and font size on reading text on handheld computers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3585 LNCS, pp. 253–266). https://doi.org/10.1007/11555261_23

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. *Journal of Applied Social Psychology*. https://doi.org/10.1111/j.1559-1816.1992.tb00945.x

Davis, F. D. F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, FD. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(September 1989), 319–340. https://doi.org/10.2307/249008

Davis, Fred, Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, *22*(14), 1111–1132. https://doi.org/10.1111/j.1559-1816.1992.tb00945.x

Davis, S., & Wiedenbeck, S. (2001). The mediating effects of intrinsic motivation, ease of use and usefulness perceptions on performance in first-time and subsequent computer users. *Interacting with Computers*, *13*(5), 549–580. Retrieved from http://www.sciencedirect.com/science/article/pii/S0953543801000340

Denscombe, M. (2006). Web-Based Questionnaires and the Mode Effect An Evaluation Based on Completion Rates and Data Contents of Near-Identical Questionnaires Delivered in Different. *Social Science Computer Review*, *24*(2), 246–254. Retrieved from http://ssc.sagepub.com/content/24/2/246.short

Dyson, M. C. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human Computer Studies*, *54*(4), 585–612. https://doi.org/10.1006/ijhc.2001.0458

Edmondson, D. R. (2005). Likert scales: A History. In *the Conference on Historical Analysis and Research in Marketing, Vol. 12, 2005* (pp. 127–133).

Fazio, R. H., & Olson, M. A. (1990). The MODE model: Attitude-Behavior Processes as a Function of Motivation and Opportunity. *Advances in Experimental Social Psychology*, *23*. https://doi.org/Retrieved from https://www.guilford.com/books/Dual-Process-Theories-of-the-Social-Mind/Sherman-Gawronski-Trope/9781462514397/contents

Finstad, K. (2006). The System Usability Scale and Non-Native English Speakers. *English*, *1*(4), 185–188. https://doi.org/1

Fitts, P. M. (1954). the Amplitude of Movement. *Journal of Experimental Psychology*, *47*, 381–391.

Frughling, A., & Lee, S. (2005). Assessing the Reliability, Validity and Adaptability of PSSUQ. *AMCIS 2005: Proceedings of the 11th Americas Conference on Information Systems*.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, *20*(4), 384–408. https://doi.org/10.1191/0265532203lt265oa

Funke, F. (2016). A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. *Social Science Computer Review*, *34*(2), 244–254. https://doi.org/10.1177/0894439315575477

Funke, F., Reips, U. D., & Thomas, R. K. (2011). Sliders for the smart: Type of rating scale on the web interacts with Educational Level. *Social Science Computer Review*, *29*(2), 221–231. https://doi.org/10.1177/0894439310376896

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers and Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Gillan, D., & Cooke, N. (1995). Methods of cognitive analysis for HCI. *Conference Companion on Human Factors in …*, *7*(11), 349–350. Retrieved from http://dl.acm.org/citation.cfm?id=223716

Giosan, C. (2006). High-K Strategy Scale: A Measure of the High-K Independent Criterion of Fitness. *Evolutionary Psychology*, *4*(1), 147470490600400. https://doi.org/10.1177/147470490600400131

Goldberg, L. R., John, O. P., Kaiser, H., Lanning, K., & Peabody, D. (1990). An Alternative " Description of Personality ": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229.

Gradišar, M., Turk, T., & Humar, I. (2010). Factors Affecting Reading Speed Measurements of Coloured Web Pages. *Organizacija*, *43*(4), 173–185. https://doi.org/10.2478/v10051-010-0017-y

Greco, M., Stucchi, N., Zavagno, D., & Marino, B. (2008). On the portability of computer-generated presentations: the effect of text-background color combinations on text legibility. *Human Factors*, *50*(5), 821–833. https://doi.org/10.1518/001872008X354156

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality1. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316447703700403

Grier, R. A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. *Proceedings of the Human Factors and Ergonomics Society*, *2015-Janua*, 1727–1731. https://doi.org/10.1177/1541931215591373

Grobelny, J., & Michalski, R. (2015a). Computers in Human Behavior The role of background color , interletter spacing , and font size on preferences in the digital presentation of a product. *Computers in Human Behavior*, *43*, 85–100. https://doi.org/10.1016/j.chb.2014.10.036

Grobelny, J., & Michalski, R. (2015b). The role of background color, interletter spacing, and font size on preferences in the digital presentation of a product. *Computers in Human Behavior*, *43*, 85–100. https://doi.org/10.1016/j.chb.2014.10.036

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview survey. Wiley series in probability and statistics. Survey methodology section*.

Hall, R. H., & Hanna, P. (2004a). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology*, *23*(3), 183–195. https://doi.org/10.1080/01449290410001669932

Hall, R. H., & Hanna, P. (2004b). The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology*, *23*(February 2015), 183–195. https://doi.org/10.1080/01449290410001669932

Hansen, W. J. (1971). User engineering principles for interactive systems. In *Proceedings of the May 16-18, 1972, spring joint computer conference on - AFIPS '72 (Spring)*. https://doi.org/10.1145/1479064.1479159

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hayes, M. H., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*.

Heijden, H. van der. (2004). User Acceptance of Hedonic Information Systems. *MIS Quarterly.*, *28*(4), 695–704.

Henze, N., Rukzio, E., & Boll, S. (2011). 100,000,000 Taps: Analysis and Improvement of Touch Performance in the Large. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI ’11*, 133. https://doi.org/10.1145/2037373.2037395

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., … Verplank, W. (1992). ACM SIGCHI Curricula for Human-Computer Interaction. *Acm Sigchi*, *2008*(February 17), 162. https://doi.org/10.1017/CBO9781107415324.004

Higgins. (1987). Self-discrepancy: a theory relating self and affect. *Psycnet.Apa.Org*, *94*(3), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors*. https://doi.org/10.1177/001872089203400405

Hofer, P. J., & Green, B. F. (1985). The Challenge of Competence and Creativity in Computerized Psychological Testing. *Journal of Consulting and Clinical Psychology*, *53*(6), 826–838. https://doi.org/10.1037/0022-006X.53.6.826

Höhne, J. K., Schlosser, S., & Krebs, D. (2017). Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys Using Paradata. *Field Methods*. https://doi.org/10.1177/1525822X17710640

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, *48*(1), 71–74. https://doi.org/10.1145/1039539.1039541

Humar, I., Gradisar, M., Turk, T., & Erjavec, J. (2014). The impact of color combinations on the legibility of text presented on LCDs. *Applied Ergonomics*, *45*(6), 1510–1517. https://doi.org/10.1016/j.apergo.2014.04.013

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. *Computer Networks*, *31*(11), 1129–1137. https://doi.org/10.1016/S1389-1286(99)00013-4

Kaiser, H. F. (1974). An index of factor simplicity Journal. *Psychometrika*, *39*(1), 31–36.

Kaya, N., & Epps, H. H. (2004). RELATIONSHIP BETWEEN COLOR AND EMOTION: A STUDY OF COLLEGE STUDENTS. *College Student Journal*, *38*(3), 396–410.

Kim, D., & Kim, D. J. (2012). Effect of screen size on multimedia vocabulary learning. *British Journal of Educational Technology*, *43*(1), 62–70. https://doi.org/10.1111/j.1467-8535.2010.01145.x

Kim, G. J. (2015). Human – Computer Interaction Fundamentals and Practice. *Human Computer Interaction: Fundamentals and Practice*, 1–12. Retrieved from http://www.ittoday.info/Excerpts/HCI.pdf

Kirakowski, J;Claridge, N;Whitehand, R. (1998). Human centered measures of success in web site design. In *Proceedings of the Fourth Conference on Human Factors and the Web* (pp. 1–9). Retrieved from http://research.microsoft.com/en-us/um/people/marycz/hfweb98/kirakowski/

Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Kirakowski, J., & Murphy, R. (2009). A comparison of current approaches to usability measurement. *I-Hci 2009*, (October), 13–17.

Ko, Y. H. (2017). The effects of luminance contrast, colour combinations, font, and search time on brand icon legibility. *Applied Ergonomics*, *65*, 33–40. https://doi.org/10.1016/j.apergo.2017.05.015

Kortum, P., & Sorber, M. (2015). Measuring the Usability of Mobile Applications for Phones and Tablets. *International Journal of Human-Computer Interaction*, *31*, 518–529. https://doi.org/10.1080/10447318.2015.1064658

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. *Improving Surveys with Paradata: Analytic Uses of Process Information*. https://doi.org/10.1002/9781118596869

Larossi, G. (2006). *The power of survey design : a user's guide for managing surveys, interpreting results, and influencing respondents*. https://doi.org/10.1596/978-0-8213-6392-8

Leeson, H. V. (2006). The Mode Effect: A Literature Review of Human and Technological Issues in Computerized Testing. *International Journal of Testing*, *6*(1), 1–24. https://doi.org/10.1207/s15327574ijt0601_1

Legge, G. E., Parish, D. H., Luebker, A., & Wurm, L. H. (1990). Psychophysics of reading. XI. Comparing color contrast and luminance contrast. *Journal of Optical Society of America*, *7*(10).

Lewis, J. R. (1992). Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/154193129203601617

Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questioannaires, Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. Retrieved from http://www.tandfonline.com/doi/full/10.1080/10447319509526110

Lewis, J. R. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction*, *14*(3–4), 463–488. https://doi.org/10.1080/10447318.2002.9669130

Lewis, J. R. (2015). Introduction to the Special Issue on Usability and User Experience: Methodological Evolution. *International Journal of Human-Computer Interaction*, *31*(9), 555–556. https://doi.org/10.1080/10447318.2015.1065689

Lewis, J. R. (2017). Revisiting the Factor Structure of the System Usability Scale. *Journal of Usability Studies*, *12*(4), 183–192.

Lewis, J. R. (2018a). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R. (2018b). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*, *00*(00), 1–9. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric Evaluation of the EMO and the SUS in the Context of a Large-Sample Unmoderated Usability Study. *International Journal of Human-Computer Interaction*, *31*(8), 545–553. https://doi.org/10.1080/10447318.2015.1064665

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5619 LNCS, pp. 94–103). https://doi.org/10.1007/978-3-642-02806-9_12

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: when there’s no time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099. https://doi.org/10.1145/2470654.2481287

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505. https://doi.org/10.1080/10447318.2015.1064654

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*. https://doi.org/2731047

Lin, H., Wu, F. G., & Cheng, Y. Y. (2013). Legibility and visual fatigue affected by text direction, screen size and character size on color LCD e-reader. *Displays*, *34*(1), 49–58. https://doi.org/10.1016/j.displa.2012.11.006

Ling, J, & Schaik, P. Van. (2002). The effect of text and background colour on visual search of Web pages. *Displays*, *23*(5), 223–230. Retrieved from http://www.sciencedirect.com/science/article/pii/S0141938202000410

Ling, J, & Schaik, P. Van. (2006). The influence of font type and line length on visual search and information retrieval in web pages. *International Journal of Human-Computer Studies*, *64*(5), 395–404. Retrieved from http://www.sciencedirect.com/science/article/pii/S1071581905001679

Ling, Jonathan, & Schaik, P. van. (2002). The effect of text and background colour on visual search of Web pages. *Displays*, *23*(5), 223–230. https://doi.org/10.1016/S0141-9382(02)00041-0

Ling, Jonathan, & Schaik, P. van. (2007). The influence of line spacing and text alignment on visual search of web pages. *Displays*, *28*(2), 60–67. https://doi.org/10.1016/j.displa.2007.04.003

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*. https://doi.org/10.1037/1082-989X.4.2.192

MacKenzie, I. S., Sellen, A., & Buxton, W. A. S. (1991). A comparison of input devices in element pointing and dragging tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91* (pp. 161–166). https://doi.org/10.1145/108844.108868

MacKenzie, I. S., & Zhang, S. X. (2001). An empirical investigation of the novice experience with soft keyboards. *Behaviour and Information Technology*, *20*(6), 411–418. https://doi.org/10.1080/01449290110089561

Maniar, N., Bennett, E., Hand, S., & Allan, G. (2008). The effect of mobile phone screen size on video based learning. *Journal of Software*, *3*(4), 51–61. https://doi.org/10.4304/jsw.3.4.51-61

Marcotte, E. (2010). *Responsive Web Design*. *Internet document*. https://doi.org/10.11635/2319-9954/1/1/18

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, *55*(1), 2–19. https://doi.org/10.1109/TPC.2011.2182569

McDonald, S., & Stevenson, R. J. (1996). Disorientation in hypertext: The effects of three text structures on navigation performance. *Applied Ergonomics*, *27*(1), 61–68. https://doi.org/10.1016/0003-6870(95)00073-9

McDonald, S., & Stevenson, R. J. (1998). Effects of Text Structure and Prior Knowledge of the Learner on Navigation in Hypertext. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(1), 18–27. https://doi.org/10.1518/001872098779480541

McLeod, S. (2008). Likert Scale Likert Scale Examples How can you analyze data from a Likert Scale ?, 1–3.

Mehta, R. P., & Zhu, R. J. (2009). Blue Or Red? Exploring the Effect of Color on Cognitive Performance. *Advances in Consumer Research*, *36*, 1045–1046. Retrieved from http://www.acrwebsite.org/volumes/14185/volumes/v36/NA-36

Mizobuchi, S., Ren, X., & Yasumura, M. (2002). An empirical study of the minimum required size and the number of targets for pen on the small display. In *MobileHCI* (Vol. 2411, pp. 184–194). https://doi.org/10.1007/3-540-45756-9

Moneta, G. B., & Csikszentmihalyi, M. (1996). The Effect of Perceived Challenges and Skills on the Quality of Subjective Experience. *Journal of Personality*. https://doi.org/10.1111/j.1467-6494.1996.tb00512.x

Naglieri, J., Drasgow, F., & Schmit, M. (2004). Psychological Testing on the Internet: New Problems, Old Issues. *American …*, *59*(3), 150–162. Retrieved from http://psycnet.apa.org/journals/amp/59/3/150/

Neilsen, J. (2000). *Designing Web Usability: The practice of simplicity*. *Interactive Marketing*. https://doi.org/10.1057/palgrave.im.4340116

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64. https://doi.org/10.1080/03098770601167922

Nielsen, J. (1993). *Usability Engineering*. *Usability Engineering*. https://doi.org/10.1145/1508044.1508050

Nielsen, J. (2001). Designing Web Usability: The practice of simplicity. *Interactive Marketing*, *3*, 77–78. https://doi.org/10.1057/palgrave.im.4340116

Nielsen, J. (2002). Let Users Control Font Size. *Nielsen Norman Group*, 1–6.

Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM*, *26*(4), 254–258. https://doi.org/10.1145/2163.358092

Norman, D., & Nielsen, J. (2016). The Definition of User Experience (UX). *Nielsen Norman Group Publication*, 1. Retrieved from https://www.nngroup.com/articles/definition-user-experience/

Norman, K. L., Friedman, Z., Norman, K., & Stevenson, R. (2001). Navigational Issues in the Design of On-Line Self-Administered Questionnaires. *Behaviour and Information Technology*, *20*(1), 37–45.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387

Nunnally, J., & Bernstein, I. (1994). Psychometric Theory, 3rd edn, 1994. *McGraw-Hill, New York*.

Nunnally, J. C. (1978). *Psychometric THEORY* (Vol. 1).

Orcan, F. (2018). Exploratory and Confirmatory Factor Analysis: Which One to Use First? *Journal of Measurement and Evaluation in Education and Psychology*, *9*(4), 414–421. https://doi.org/10.21031/epod.394323

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*. https://doi.org/10.1037/h0055737

Parhi, P., Karlson, A. K., & Bederson, B. B. (2006). Target Size Study for One-handed Thumb Use on Small Touchscreen Devices. *Proceedings of the 8th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'06)*, 203–210. https://doi.org/10.1145/1152215.1152260

Park, Y. S., & Han, S. H. (2010). One-handed thumb interaction of mobile devices from the input accuracy perspective. *International Journal of Industrial Ergonomics*, *40*(6), 746–756. https://doi.org/10.1016/j.ergon.2010.08.001

Park, Y. S., Han, S. H., Park, J., & Cho, Y. (2008). Touch key design for target selection on a mobile phone. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services - MobileHCI ’08* (p. 423). https://doi.org/10.1145/1409240.1409304

Parush, A., & Yuviler-Gavish, N. (2004). Web navigation structures in cellular phones: The depth/breadth trade-off issue. *International Journal of Human Computer Studies*, *60*(5–6), 753–770. https://doi.org/10.1016/j.ijhcs.2003.10.010

Pearson, R., & Schaik, P. van. (2003). The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. *International Journal of Human Computer Studies*, *59*(3), 327–353. https://doi.org/10.1016/S1071-5819(03)00045-4

Perry, D. E., & Wolf, A. L. (1992). Foundations for the study of software architecture. *ACM SIGSOFT Software Engineering Notes*, *17*(4), 40–52. https://doi.org/10.1145/141874.141884

Perry, K., & Hourcade, J. (2008). Evaluating one handed thumb tapping on mobile touchscreen devices. In *Proceedings of Graphics Interface 2008* (pp. 57–64). Retrieved from http://dl.acm.org/citation.cfm?id=1375725

Pettit, F. (1999). Exploring the use of the World Wide Web as a psychology data collection tool. *Computers in Human Behavior*, *15*(1), 67–71. Retrieved from http://www.sciencedirect.com/science/article/pii/S0747563298000338

Peytchev, A., & Hill, C. A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review*. https://doi.org/10.1177/0894439309353037

Piepenbrock, C., Mayr, S., & Buchner, A. (2014). Positive display polarity is particularly advantageous for small character sizes: Implications for display design. *Human Factors*, *56*(5), 942–951. https://doi.org/10.1177/0018720813515509

Piepenbrock, C., Mayr, S., Mund, I., & Buchner, A. (2013). Positive display polarity is advantageous for both younger and older adults. *Ergonomics*. https://doi.org/10.1080/00140139.2013.790485

Ramadan, M., Mohamed, A., & El-Hariry, H. (2010). Effects of cathode ray tube display formats on quality-assurance auditor's performance. *Human Factors and Ergonomics In Manufacturing*. https://doi.org/10.1002/hfm.20166

Ramadan, M. Z. (2011). Evaluating college students' performance of Arabic typeface style, font size, page layout and foreground/background color combinations of e-book materials. *Journal of King Saud University - Engineering Sciences*, *23*(2), 89–100. https://doi.org/10.1016/j.jksues.2011.03.005

Reenskaug, T. (1979). Models - Views - Controllers. *Technical Note, Xerox PARC*.

Reenskaug, T. (2003). The Model-View-Controller (MVC) Its Past and Present. *University of Oslo Draft*.

Reeves, B., Lang, A., Kim, E. Y., Tatar, D., Reeves, B., Lang, A., & Kim, E. Y. (1999). The Effects of Screen Size and Message Content on Attention and Arousal The Effects of Screen Size and Message Content on Attention and Arousal. *Media Psychology*, *1*(1), 49–67. https://doi.org/10.1207/s1532785xmep0101

Reinecke, K., & Bernstein, A. (2011). Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transactions on Computer-Human Interaction*, *18*(2), 1–29. https://doi.org/10.1145/1970378.1970382

Rello, L., & Baeza-Yates, R. (2017). How to present more readable text for people with dyslexia. *Universal Access in the Information Society*. https://doi.org/10.1007/s10209-015-0438-8

Rello, L., Pielot, M., & Marcos, M.-C. (2016a). Make It Big!: The Effect of Font Size and Line Spacing on Online Readability. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (May), 3637–3648. https://doi.org/10.1145/2858036.2858204

Rello, L., Pielot, M., & Marcos, M. (2016b). Make It Big!: The Effect of Font Size and Line Spacing on Online Readability. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3637–3648. https://doi.org/10.1145/2858036.2858204

Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., & Eisenberg, M. (2005). Design principles for tools to support creative thinking. *Creativity Support Tools: Report from a US National Science Foundation Sponsored Workshop*, (February 2014).

Reutskaja, E., & Hogarth, R. M. (2009). Satisfaction in Choice as a Function of the Number of Alternatives: When "Goods Satiate. *Psychology & Marketing*, *26*(3), 197–203. https://doi.org/10.1086/651235

Ritter, P., Lorig, K., Laurent, D., & Matthews, K. (2004). Internet versus mailed questionnaires: a randomized comparison. *Journal of Medical Internet Research*, *6*(3), e29. https://doi.org/10.2196/jmir.6.3.e29

Riva G., Davide F., Ij. W. A. (Eds. . (2003). Persuasive Effects of Presence in immersive virtual environments.

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring slider vs. Categorical response formats in web-based surveys. *Journal of Research Practice*, *11*(1).

Rust, J., & Golombok, S. (1989). *Modern Psychometrics*. *The Science of Psychological Assessment*.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, 2215. https://doi.org/10.1145/1978942.1979266

Schaik, P. van, & Ling, J. (2003). Using on-line surveys to measure three key constructs of the quality of human–computer interaction in web sites: psychometric properties and implications. *International Journal of Human-Computer Studies*, *59*(5), 545–567. https://doi.org/10.1016/S1071-5819(03)00078-8

Schaik, P. van, & Ling, J. (2005a). Five psychometric scales for online measurement of the quality of human-computer interaction in web sites. *International Journal of Human-Computer Interaction*, (3), 309–322. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327590ijhc1803_4

Schaik, P. van, & Ling, J. (2005b). the Psychometric Evaluation of Educational Intranets. *Journal of Educational Computing Research*, *33*(1), 81–100. https://doi.org/10.2190/06M5-HUK3-02AL-R993

Schaik, P. van, & Ling, J. (2007). Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction Interaction*, *14*(1), 4-es. https://doi.org/10.1145/1229855.1229859

Schaik, P. van, & Ling, J. (2008). Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers*, *20*(3), 419–432. https://doi.org/10.1016/j.intcom.2008.03.001

Schaik, P. van, Luan Wong, S., & Teo, T. (2015). Questionnaire layout and national culture in online psychometrics. *International Journal of Human Computer Studies*, *73*, 52–65. https://doi.org/10.1016/j.ijhcs.2014.08.005

Schmidt, W. (1997). World-Wide Web survey research made easy with WWW Survey Assistant. *Behavior Research Methods*, *29*(2), 303–304. Retrieved from http://www.springerlink.com/index/V2404483478MN782.pdf

Segall, N., Doolen, T. L., & Porter, J. D. (2005). A usability comparison of PDA-based quizzes and paper-and-pencil quizzes. *Computers and Education*. https://doi.org/10.1016/j.compedu.2004.05.004

Shneiderman, B. (1983). Direct Manipulation: A Step Beyond Programming Languages. *Computer*, *16*(8), 57–69. https://doi.org/10.1109/MC.1983.1654471

Shneiderman, B. (2002). Understanding human reactivites and relationships: an excerpt from Leonardo’s laptop. *Interactions*, *9*(5), 40–53. https://doi.org/10.1145/566981.566982

Shneiderman, B., & Plaisant, C. (2010). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. *Pearson* (Vol. 5th). https://doi.org/10.1016/0166-3615(93)90066-A

Simonsen, E., & Mortensen, E. L. (1990). Difficulties in translation of personality scales. *Journal of Personality Disorders*. https://doi.org/10.1521/PEDI.1990.4.3.290

Smith, S. L., & Mosier, J. N. (1986). Guidelines for Designing User Interface Software. *Guidelines for Designing User Interface Software*, *ESD-TR-86*-(ESD-TR-86-278), 0. https://doi.org/10.1001/archdermatol.2011.235

Stapleton, C. D. (1997). Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach. *Southwest Educational Research Association*, 1–8. Retrieved from http://ericae.net/ft/tamu/Efa.htm%5Cnpapers3://publication/uuid/0B1D72E5-D700-48C4-A378-C166C1ABDA4A

Stapleton, C. E. (2013). The smartphone way to collect survey data. *Survey Practice*, *6*(2), 1–7.

Sweeney, M., & Dillon, A. (1987). Methodologies Employed in the Psychological Evaluation of H.C.I. *Human–Computer Interaction–INTERACT '87*, 367–373. https://doi.org/10.1016/B978-0-444-70304-0.50065-0

Tarafdar, M., & Zhang, J. (2005). Analyzing the Influence of Web Site Design Parameters on Web Site Usability. *Information Resources Management Journal*. https://doi.org/10.4018/irmj.2005100104

Ten Klooster, P. M., Visser, M., & de Jong, M. D. T. (2008). Comparing two image research instruments: The Q-sort method versus the Likert attitude questionnaire. *Food Quality and Preference*, *19*(5), 511–518. https://doi.org/10.1016/j.foodqual.2008.02.007

Timpany, C. (2009). The affect of type and colour on readability in print and screen based environments. In *Proceedings of World Conference on Educational Media and Technology* (pp. 797–806).

Toepoel, V. (2017). Online survey design. *The Sage Handbook of Online Research Methods*, 184–202. https://doi.org/10.4135/9781473957992.n11

Tognazzini, B. (2003). First principles of interaction design. *Interaction Design Solutions for the Real World, AskTog*, (2007), 1–5. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Principles+of+Interaction+Design#0%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:First+principles+of+interaction+design#0

Tuch, A. N., Schaik, P. van, & Hornbæk, K. (2016). Leisure and Work, Good and Bad: The Role of Activity Domain and Valence in Modeling User Experience. *ACM Transactions on Computer-Human Interaction*, *23*(6), 1–32. https://doi.org/10.1145/2994147

Umami, M. K., Arezes, P. M., & Sampaio, Á. M. (2016). Understanding finger postures when touching targets on the touchscreen of mobile devices. *DYNA (Colombia)*, *83*(197), 31–37. https://doi.org/10.15446/dyna.v83n197.57587

Van De Looij-Jansen, P. M., & De Wilde, E. J. (2008). Comparison of web-based versus paper-and-pencil self-administered questionnaire: effects on health indicators in Dutch adolescents. *Health Services Research*, *43*(5 Pt 1), 1708–1721. https://doi.org/10.1111/j.1475-6773.2008.00860.x

Venkatesh, Visawanath. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research*. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, Visawanath, Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.1017/CBO9781107415324.004

Venkatesh, Visawanath, & Speier, C. (1999). Computer technology training in the workplace: a longitudinal investigatin of the effect of moods. *Organiztional Behavior and Human Decision Processes*, *79*(1), 1–28.

Venkatesh, Visawanath, & Speier, C. (2000). Creating an effective training environment for enhancing telework. *International Journal of Human-Computer Studies*, *52*, 991–1005. https://doi.org/10.1006/ijhc.1999.0367

Venkatesh, Viswanath, Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.1111/j.1540-4560.1981.tb02627.x

Vogt, W. P. (2005). *Dictionary of Statistics & Methodology*. *Dictionary of Statistics & Methodology* (3rd ed). Sage. https://doi.org/http://dx.doi.org/10.4135/9781412983907.n697

Woods, D. D. (1984). Visual momentum: a concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies*. https://doi.org/10.1016/S0020-7373(84)80043-7

Yan, T., & Olson, K. (2013). Analysing Paradaata to Investigate Measurement Error. *Improving Surveys With Paradata*, 74–95.

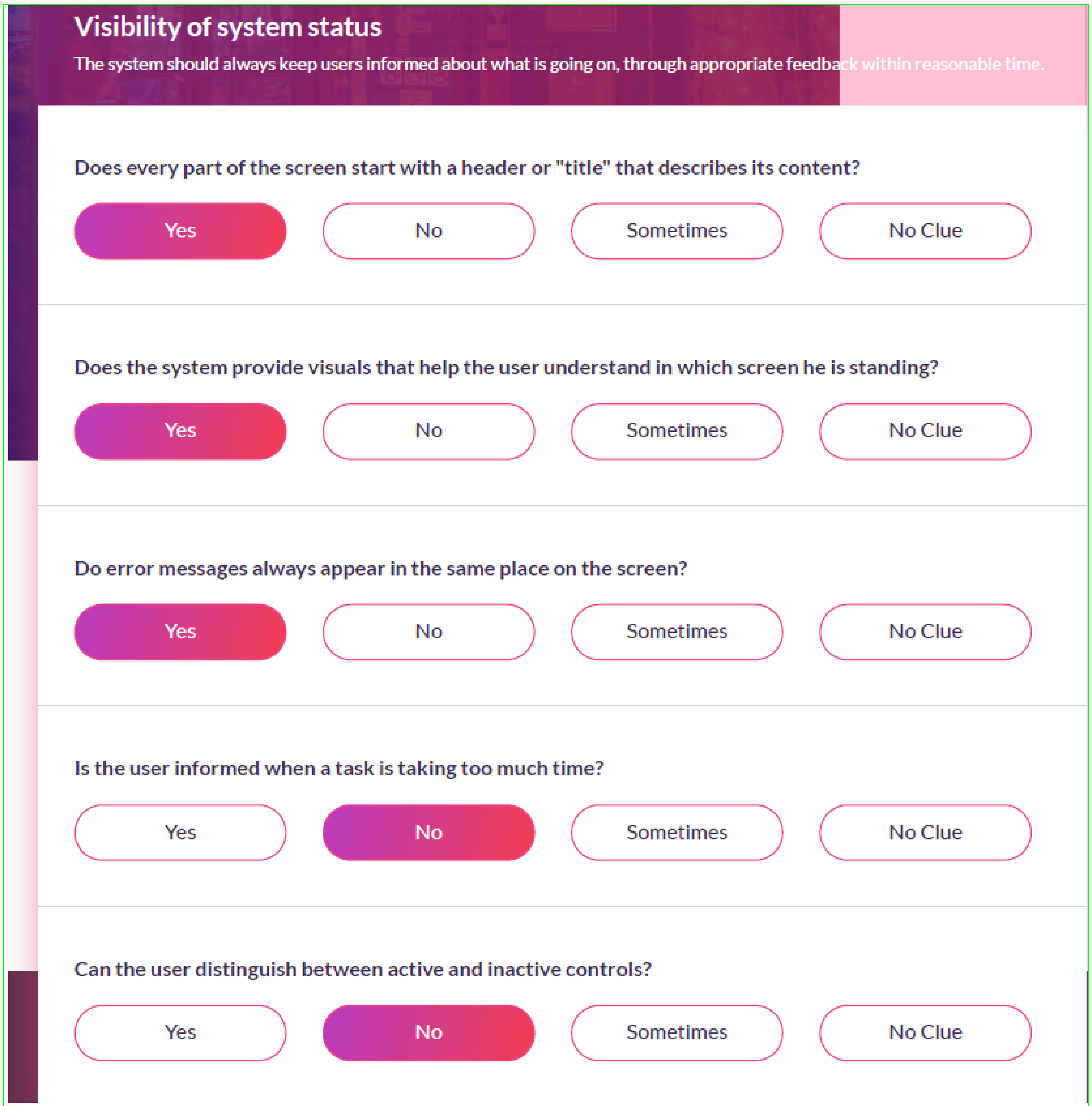
Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.1331

Zajonc, R. (University of M. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2).

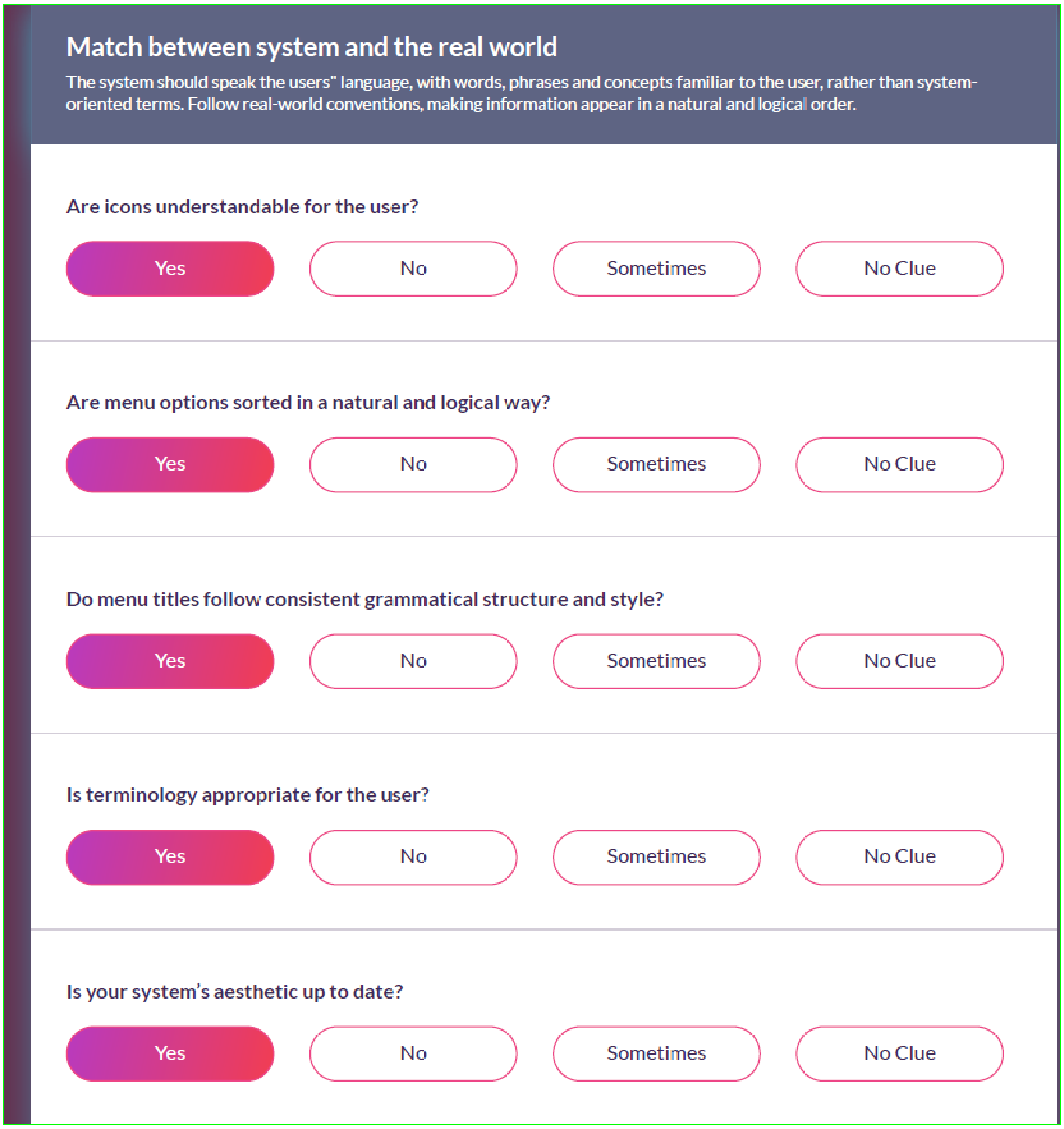
Zajonc, R. (University of M. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224.

Appendices

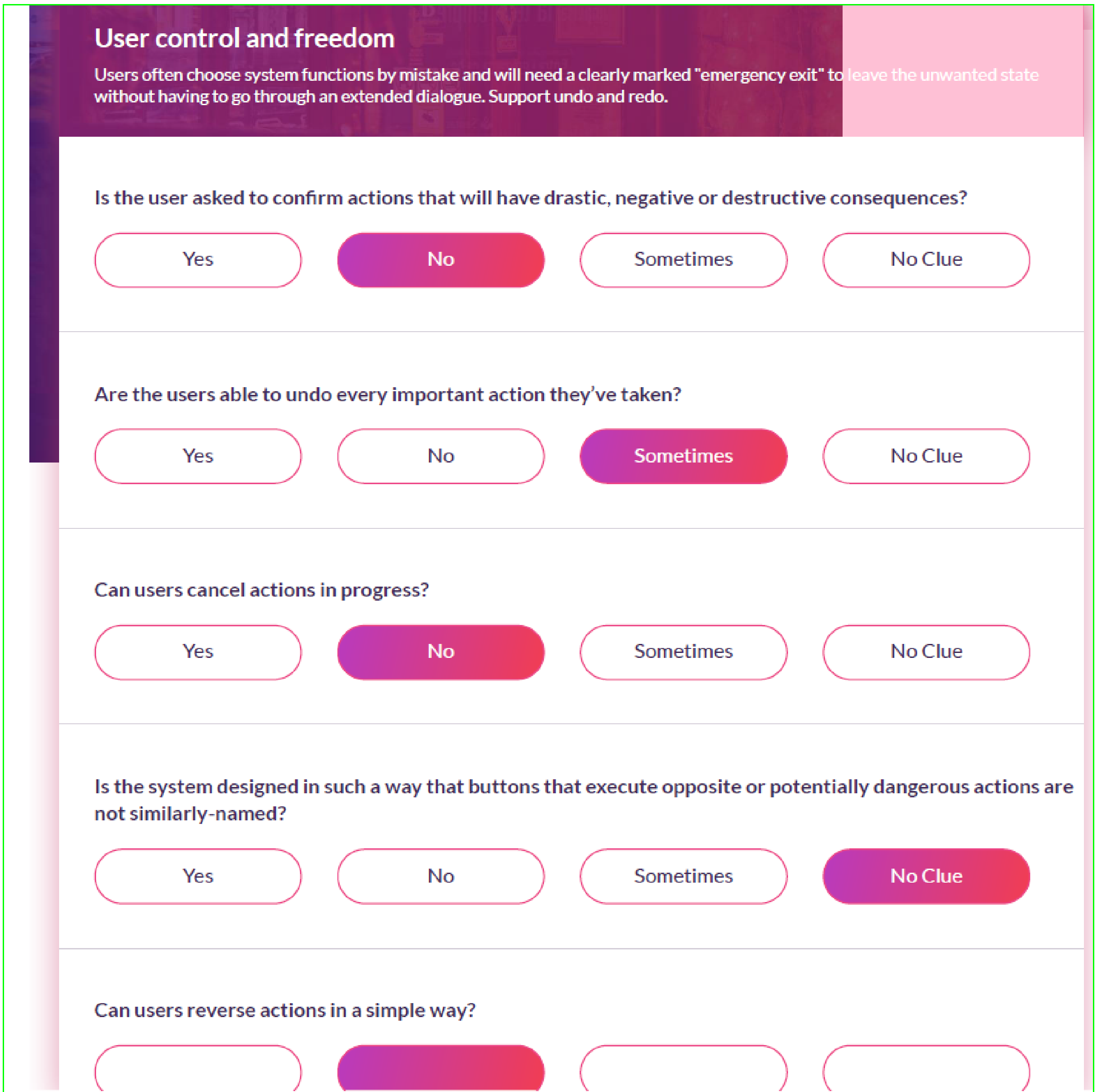
Subscales of the heuristics evaluation of the usability quiz.



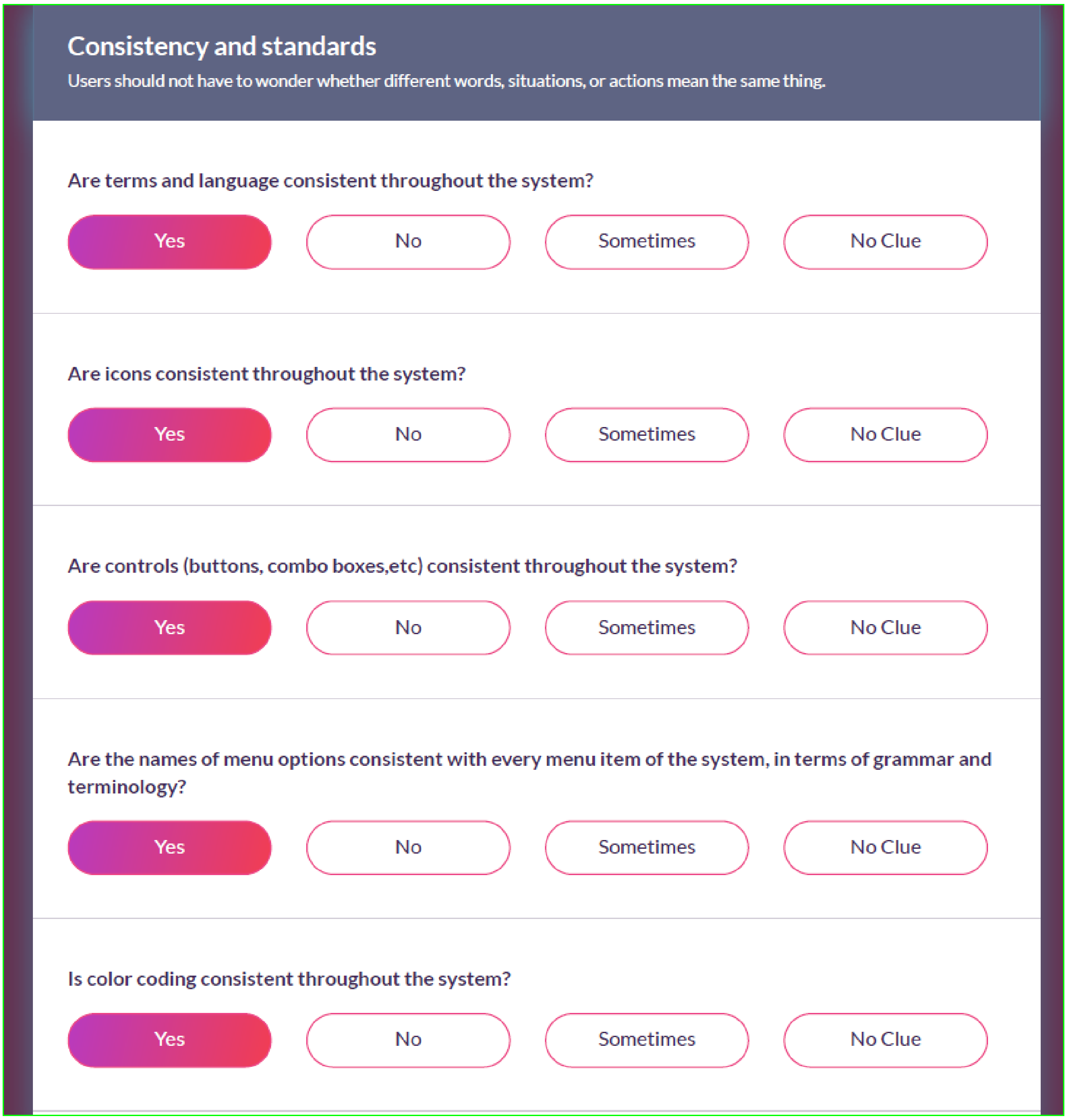
https://www.uruit.com/ux-quiz/1

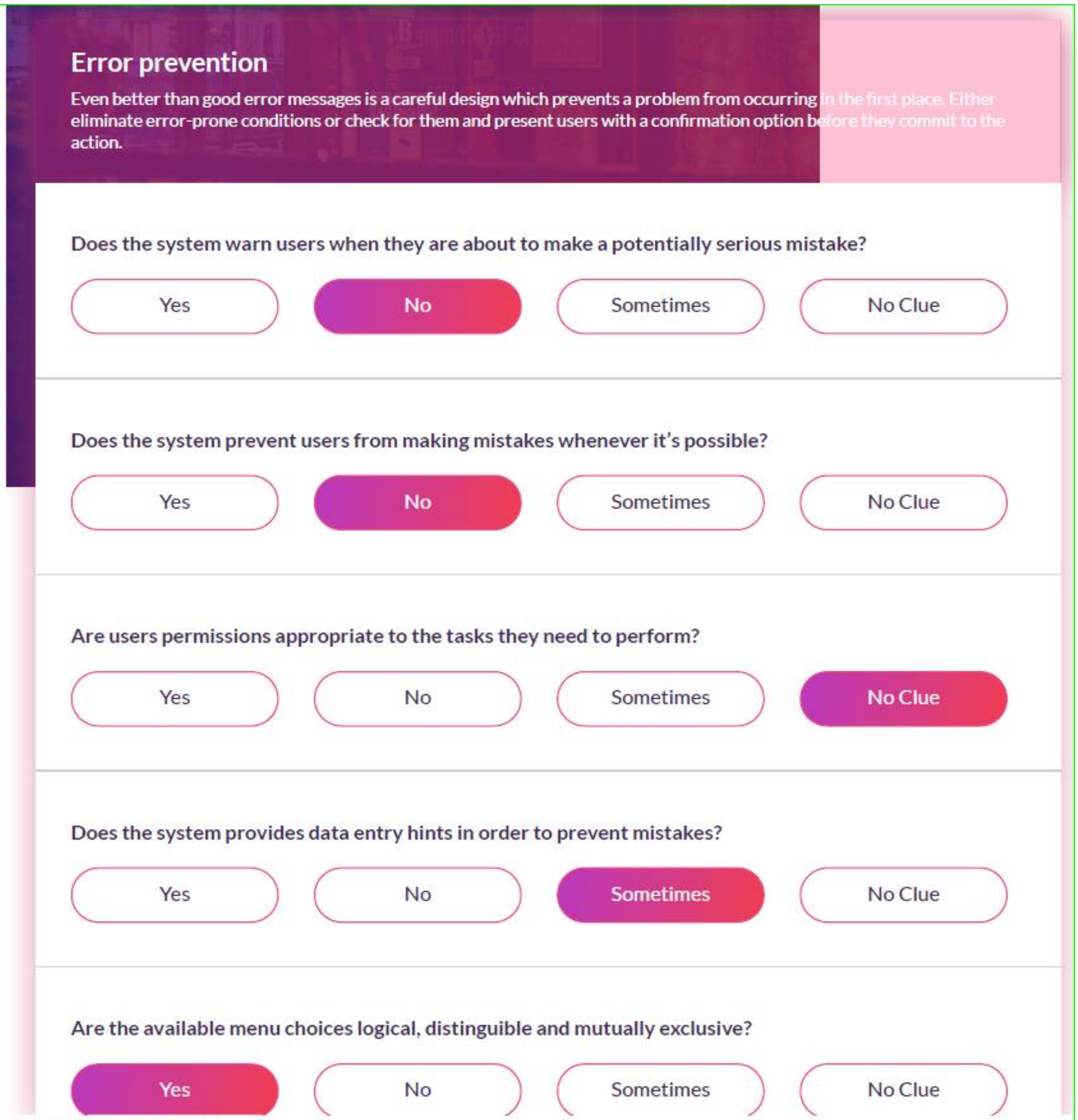


Match between system and the real world
The system should speak the users" language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.
Are icons understandable for the user?
Yes
No
Sometimes
No Clue
Are menu options sorted in a natural and logical way?
Yes
No
Sometimes
No Clue
Do menu titles follow consistent grammatical structure and style?
Yes
No
Sometimes
No Clue
Is terminology appropriate for the user?
Yes
No
Sometimes
No Clue
Is your system's aesthetic up to date?
Yes
No
Sometimes
No Clue

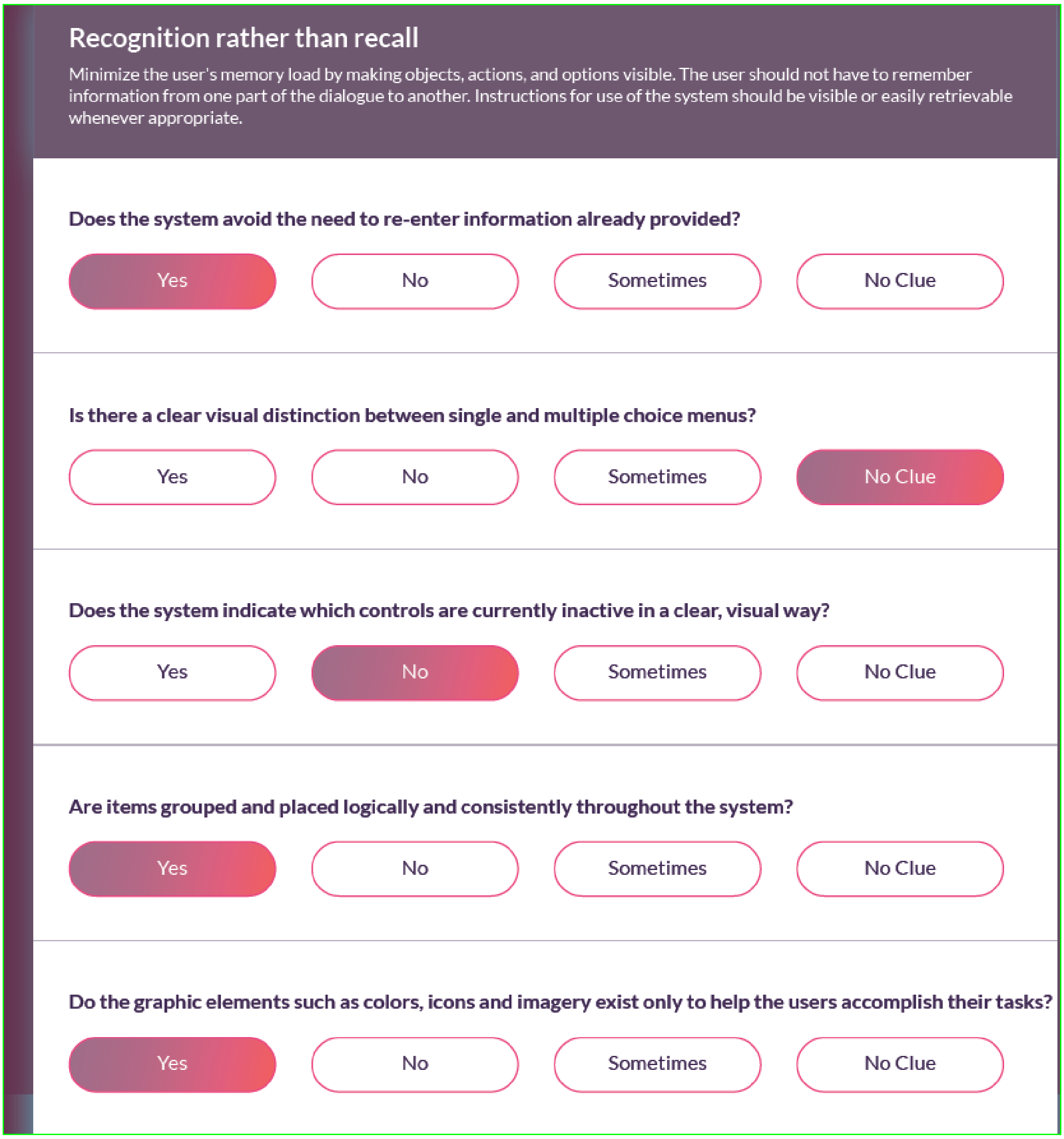


https://www.uruit.com/ux-quiz/2


Consistency and standards
Users should not have to wonder whether different words, situations, or actions mean the same thing.
Are terms and language consistent throughout the system?
Yes
No
Sometimes
No Clue
Are icons consistent throughout the system?
Yes
No
Sometimes
No Clue
Are controls (buttons, combo boxes,etc) consistent throughout the system?
Yes
No
Sometimes
No Clue
Are the names of menu options consistent with every menu item of the system, in terms of grammar and terminology?
Yes
No
Sometimes
No Clue
Is color coding consistent throughout the system?
Yes
No
Sometimes
No Clue



https://www.uruit.com/ux-quiz/3



Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Does the system avoid the need to re-enter information already provided?

Yes | No | Sometimes | No Clue

Is there a clear visual distinction between single and multiple choice menus?

Yes | No | Sometimes | No Clue

Does the system indicate which controls are currently inactive in a clear, visual way?

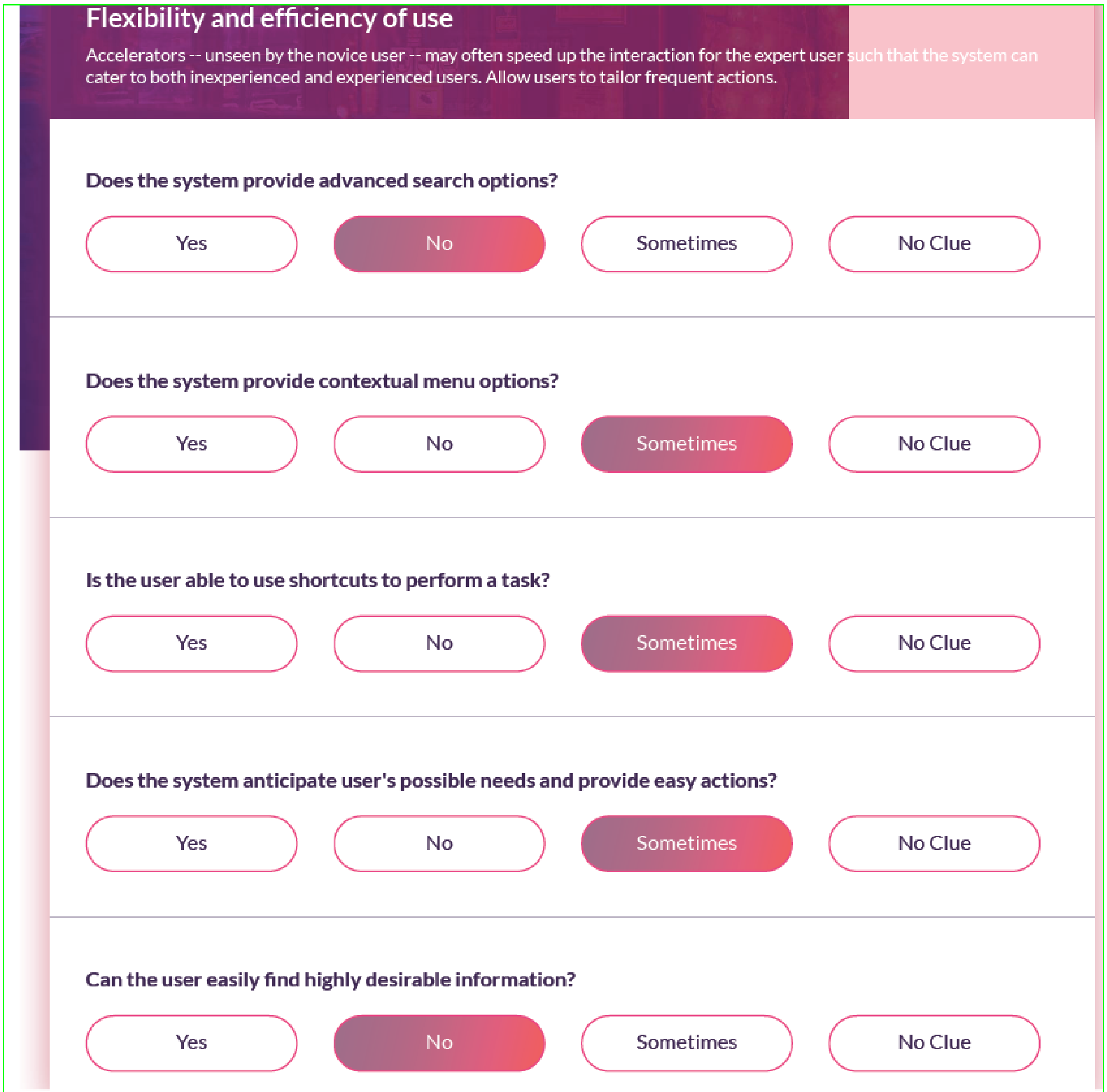
Yes | No | Sometimes | No Clue

Are items grouped and placed logically and consistently throughout the system?

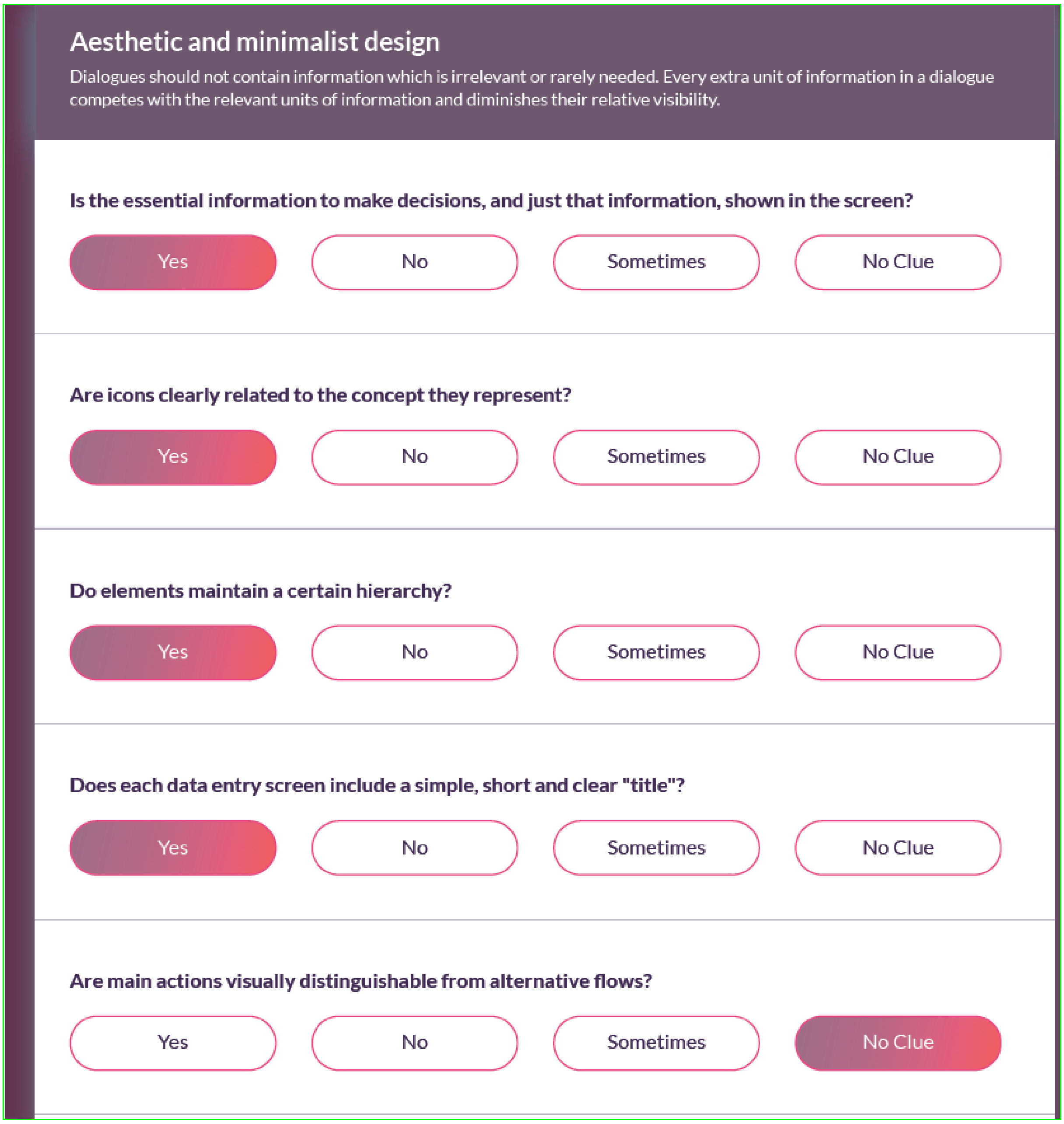
Yes | No | Sometimes | No Clue

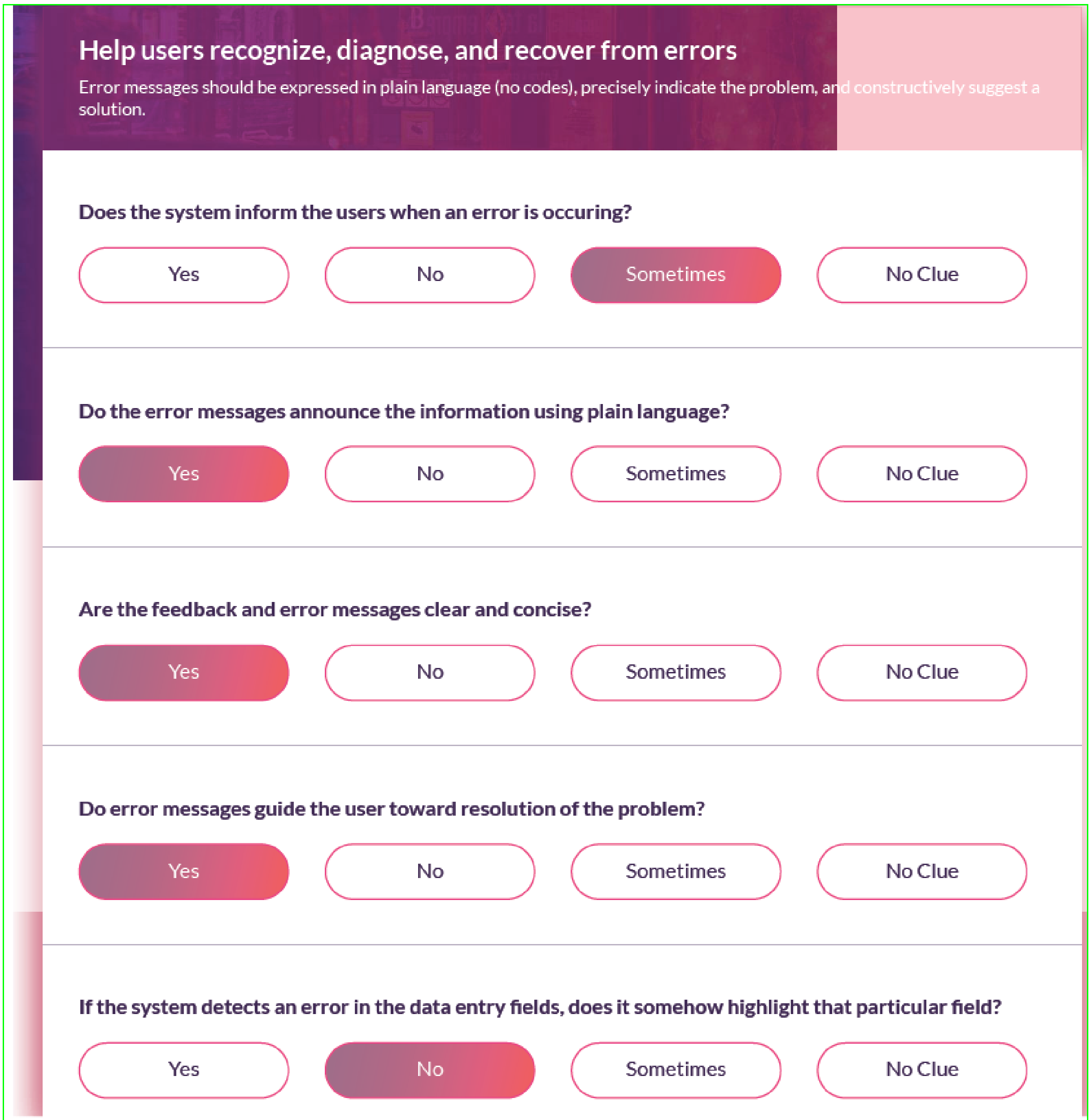
Do the graphic elements such as colors, icons and imagery exist only to help the users accomplish their tasks?

Yes | No | Sometimes | No Clue



https://www.uruit.com/ux-quiz/4


Aesthetic and minimalist design
Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
Is the essential information to make decisions, and just that information, shown in the screen?
Yes
No
Sometimes
No Clue
Are icons clearly related to the concept they represent?
Yes
No
Sometimes
No Clue
Do elements maintain a certain hierarchy?
Yes
No
Sometimes
No Clue
Does each data entry screen include a simple, short and clear "title"?
Yes
No
Sometimes
No Clue
Are main actions visually distinguishable from alternative flows?
Yes
No
Sometimes
No Clue


Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Does the system inform the users when an error is occuring?

Yes | No | Sometimes | No Clue

Do the error messages announce the information using plain language?

Yes | No | Sometimes | No Clue

Are the feedback and error messages clear and concise?

Yes | No | Sometimes | No Clue

Do error messages guide the user toward resolution of the problem?

Yes | No | Sometimes | No Clue

If the system detects an error in the data entry fields, does it somehow highlight that particular field?

Yes | No | Sometimes | No Clue

https://www.uruit.com/ux-quiz/5

Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Is the help information accurate, comprehensible and complete?

Yes | No | Sometimes | No Clue

Is the help information right where needed?

Yes | No | Sometimes | No Clue

Does the documentation allow the user to understand, interpret and proceed correctly?

Yes | No | Sometimes | No Clue

Can the user easily access help information without interrupting the work?

Yes | No | Sometimes | No Clue

After reading the help information, are the users able of continuing with their work right where they left it?

Yes | No | Sometimes | No Clue

The results of the heuristics evaluation of the usability quiz.