A Methodology for Confidence-based Adaptive Numeracy Skill Assessment in Healthcare

Mandana Omidbakhsh

A Thesis
In the Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Computer Science and Software Engineering) at
Concordia University
Montreal, Quebec, Canada

August 2016

ABSTRACT

A Methodology for Confidence-based Adaptive Numeracy Skill Assessment in Healthcare

Mandana Omidbakhsh, Ph.D.

Concordia University, 2016

Numeracy skill level of patients has great influence on their preferences and priorities for the treatment options concerning their healthcare. The elicitation of patient preferences in healthcare, along with the increasing degree of patients' participation in their own treatment decision- making immensely signify the importance of the topic. Numeracy in healthcare domain is a measure of the ability of patients to understand and digest numerically presented information so as to make appropriate health decisions and understand risk factors.

Not properly numeracy-assessed patients are prone to make inaccurate and inappropriate decisions for their medical treatments. There are many challenges that the researchers face in designing and developing patient-sensitive numeracy assessment methods. The adaptability of the numeracy assessment is considered to be one of the most important issues to address. The existing methods of numeracy testing do not take confidence as a parameter in consideration for adaptive assessment. Numeracy assessment without confidence is prone to guess work. A better result in measurement is achieved when confidence in the knowledge is also appraised. More importantly, patients may act up on knowledge when they have confidence in it. Thus, we aimed to develop a novel model for Patient Numeracy Assessment based on this parameter.

We proposed a goal-driven Confidence-based model for Patient Numeracy Assessment (C-PNA), which (1) is adaptable to each individual patient, (2) covers the full sets of numeracy skills, and (3) considers confidence. Our adaptive model is based on a conceptual math model. Accordingly, to develop our model, we applied the Confidence Based Learning method for the measurement of confidence and (4) created the item bank, (5) defined the selection algorithm and (6) specified the associated scoring protocol applicable for the

assessment of numeracy. To validate the feasibility of our model, we conducted several empirical studies and demonstrated that the results are statistically significant.

We also (7) introduced a novel quality model for the evaluation of patient numeracy assessment methods. The quality model, which is inspired by ISO/IEC 25022, covers both (8) objective and (9) subjective characteristics regarding patient interface with numeracy assessment methods. We further applied this quality model to compare our numeracy assessment methods with the other existing methods. We were able to establish the place of our Confidence-based Patient Numeracy Assessment (C-PNA) method among the other numeracy assessment methods based on the empirical studies we performed. Empirical data provided the evidence for high satisfaction and trust, and significant effectiveness and usage efficiency of our patient numeracy assessment method.

The results of the empirical studies reveal that our model for the assessment of patient numeracy skill could be consequently pertained for Patient Preference Elicitation (PPE) systems. Preliminary research in support of PPE is reported in the thesis. In particular, it focuses on strategies to improve outcome of the treatments and decisions highly depending on patient’s numeracy skill. It will be instrumental in tailoring decision-supporting interventions to particular patient needs.

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my thesis advisor, Dr. Olga Ormandjieva for her expertise, understanding and continuous support. Without her thoughtful encouragement and careful supervision, this thesis would never have taken shape. She is an excellent example of a successful researcher and female professor.

I am greatly indebted to the late Dr. Thiruvengadam Radhakrishnan, my co-advisor, for his motivation, guidance and helpful advices. I am blessed to have spent time in his tutelage.

I extend my deepest thanks to Dr. Sudhir Mudur by giving generously of his time and for his contribution to the direction and the richness of this research.

I express the profoundest gratitude to my beloved parents for being role models in academic interest and for their special encouragement, and invaluable support.

I would like to thank my husband who inspired and supported me throughout this research.

A special feeling of gratitude to my sister who always supported me emotionally.

I dedicate this work to my son who became my ultimate motivation for this work.

I am thankful to all the participants in my study for their kind cooperation and time.

Lastly, I would like to thank anyone who has contributed to the success of this research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

Chapter 1 Introduction

In this chapter, we introduce the motivation for this research work, define the research statement and objectives, and clarify the significance of the work. We specifically discuss the major contributions and also give an overview of the organization of the thesis.

1.1 Research Motivation and Challenges

In recent years, patients have become more and more involved in their healthcare and are encouraged to participate in making decisions about their treatments. Thus, there is a great demand in research for increasing patients' involvement in decision making that result in improving patient knowledge and reducing decisional conflict and passivity in decision making (Murray, et al., 2007). The process of eliciting patient preferences which is the process of describing the care options, gathering and framing evidence in a format comprehensible to patients and subsequently measuring patients' preferences has received a lot of attention (Taylor, 2000). Numeracy and framing are two major concerns that have immense impact on patient preferences and if not dealt properly, may result in inaccurate and unreliable preferences (Lloyd, 2003). Patients need to understand quantitative information without the impact of the format and the framing of the information presented to them.

Numeracy in healthcare domain is a measure of the ability of patients to understand and digest numerically presented information so as to make appropriate health decisions and understand risk factors. The existing methods of numeracy testing do not take confidence as a parameter in consideration for adaptive assessment. Numeracy assessment without confidence is prone to guess work. A better result in measurement is achieved when confidence in the knowledge is also appraised. More importantly, patients may act up on knowledge when they have confidence in it. This is the main challenge to be addressed in this work.

1.2 Problem Statement and Research Objectives

The research described in this thesis constitutes an attempt to address the related challenges by proposing a novel confidence-based adaptive testing model which links Human Computer Interface (HCI) and Patient Education. Ultimately, it provides a pathway in resolving the issues in patient preference elicitation in healthcare domain.
The research objectives of this work are summarized below:

Objective 1: To investigate and design a new adaptive testing model that assesses the numeracy skills of patients: (See Chapter 3)

> **Sub-Objective 1:** To investigate and incorporate the parameter of confidence in the adaptive assessment. We design a selection algorithm, and specify a scoring method for this purpose. The work is published in (Omidbaksh & Radhakrishnan, May, 2014).
>
> **Sub-Objective 2**: To review the state of the art in the area of adaptive assessment in healthcare and propose a structure for an item bank for numeracy skill with full coverage.
>
> **Sub-Objective 3:** To derive a new goal-driven patient numeracy assessment model. The model is published in (Omidbaksh & Ormandjieva, Dec., 2015).

Objective 2: To research and design a novel quality model for the evaluation of patient numeracy assessment methods. The work is published in (Omidbaksh & Ormandjieva, July, 2016). (See Chapter 4)

Objective 3: To empirically validate our model that gains patient satisfaction, trust and discretionary usage with accurate results. The results are summarized in (Omidbaksh & Ormandjieva, July, 2016). (See Chapters 5 & 6)
This is achieved through building an online application and conducting Controlled Experiment 1, Controlled Experiment 2 and Controlled Experiment 3.

Objective 4: The fourth objective of this thesis is to investigate domain expertise as a foundation for designing a patient preference elicitation process, propose the architecture of patient preference elicitation system along with the personalization of healthcare information for patients. The research findings to this objective are published in (Omidbakhsh, et al., March, 2010) (Omidbaksh & Ormandjieva, Aug., 2016) (Omidbaksh, et al., Oct., 2010). (See Chapter 2 & Chapter 7)

1.3 Methodology

To accomplish the research objectives listed in Section 1.2, the following research methodology steps are examined in this thesis as shown in Figure 1.1:

Step 1. In the first step, a study is conducted with the objective of defining a process for patient preference elicitation. This study, as detailed in Chapter 7, shows that there are two significant concerns in the process: Numeracy and Framing.

Step 2. In the second step, we reviewed the existing methods for numeracy assessment in healthcare and computer adaptive assessment in general. (Chapter 2)

Step 3. The outcome of the second step integrated with a psychological educational model and adaptive assessment process helped us to develop a new model for patient numeracy assessment: C-PNA (Confidence-based Patient Numeracy Assessment), in the third step. (Chapter 3)

Step 4. A quality model assessment is designed for comparing different numeracy methods in the fourth step. (Chapter 4)

Step 5. In the fifth step, based on proposed quality model, C-PNA is validated with other numeracy assessment methods in three empirical studies. These studies are presented in Chapter 6.

Step 6. Finally, in the sixth step we analyzed the data collected in the empirical studies and made conclusions about the work that has been done. (Chapter 6)

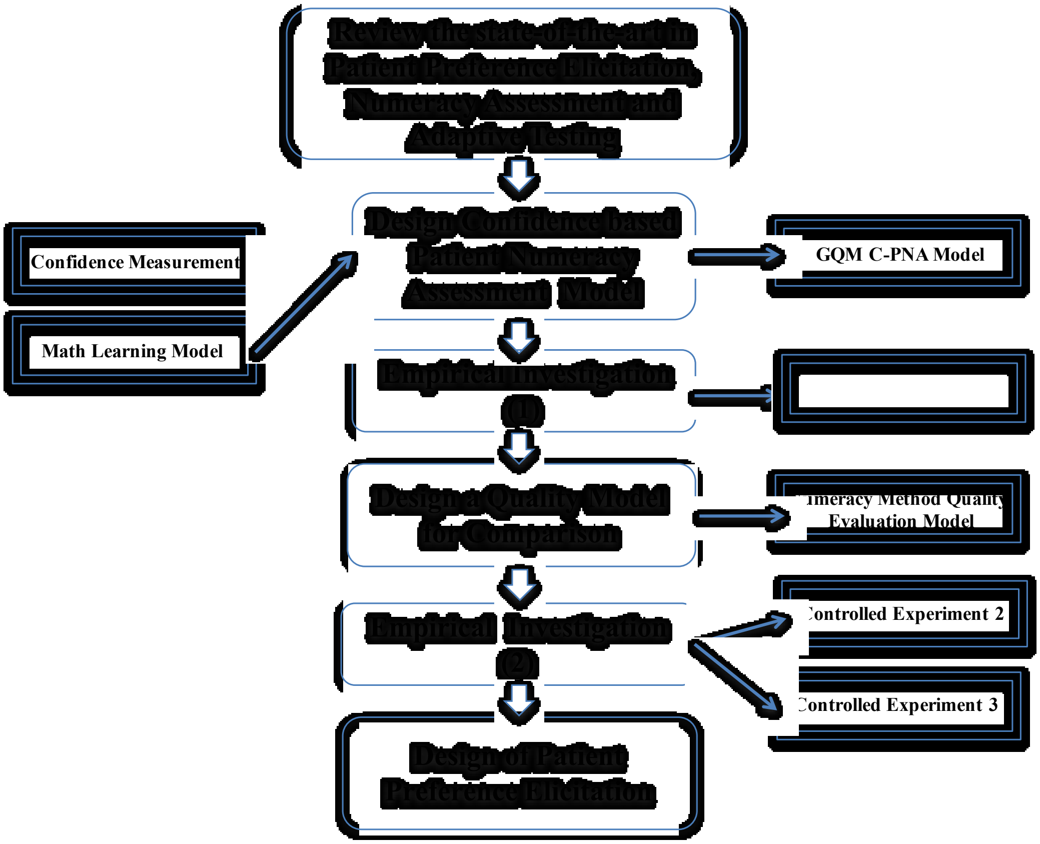

Review the state-of-the-art in Patient Preference Elicitation, Numeracy Assessment and Adaptive Testing
Confidence Measurement
Math Learning Model
Design Confidence based Patient Numeracy Assessment Model
GQM C-PNA Model
Empirical Investigation (1)
Design a Quality Model for Comparison
umeracy Method Quality Evaluation Model
Empirical Investigation (2)
ontrolled Experiment 2
ontrolled Experiment 3
Design of Patient Preference Elicitation

Figure 1.1 Outline of the Methodology

The novelty of this research lies in the interdisciplinary nature of adaptive testing methodology, confidence measurement, and the interaction between patients and the application along with subjective characteristics such as trust, satisfaction and discretionary usage that are taken into consideration.

1.4 Organization of the Thesis

The organization of this thesis is as follows:

In **Chapter 2**, we discuss the background both in Numeracy and Computerized Adaptive Testing and summarize the shortcomings of existing methods for the assessment of numeracy in healthcare. We also introduce current techniques relevant to empirical studies carried out applicable for our study.

In **Chapter 3,** we present our C-PNA model for patients as an adaptive assessment method based on the principles of computerized adaptive testing, and validate our model through empirical investigation which is discussed in Chapter 5.

In **Chapter 4,** we design a quality model for the evaluation of numercy assessment methods and then we applied the model to different controlled experiments which are discussed in Chapter 6.

In **Chapter 5**, we describe how we build our online platform for conducting empirical studies.

In **Chapter 6**, we present the empirical evaluation of our C-PNA model in real environment.

In **Chapter 7**, we propose a Patient Preference Elicitation model, which includes the two modules of numeracy assessment and framing and the architecture for the discussed process.

Finally, in **Chapter 8**, we summarize our work and its major contributions, and we present avenues for future research in this domain.

In the next chapter, we focus on numeracy assessment and provide a literature review of the existing numeracy assessment methods.

Chapter 2 Background and Related Work

Undoubtedly, one of the prominent elements of patient education is skill building and numeracy skill assessment is a very essential primary step in this regard. In this chapter we aim to explain the definition of numeracy in healthcare domain, to review the existing patient numeracy assessment approaches, and present the motivation for our novel quality model. Only through comparison, we could establish the place of our confidence-based numeracy assessment method C-PNA among the other numeracy assessment methods.

2.1 Numeracy

Numeracy or specifically health numeracy, is described as "the capacity to access, process and understand basic health information and services needed to make appropriate health decisions" (US Department of Health and Human Services, 2000). It is also defined as "accessing, processing, interpreting, communicating, and acting on numerical, quantitative, graphical, bio- statistical, and probabilistic health information necessary for effective health decision" (Goldbeck, et al., 2005).

Individual-level competencies can be categorized into four groups, namely: i) basic, ii) computational, iii) analytical, and iv) statistical literacy (Goldbeck, et al., 2005).

The basic group is considered as number recognition, while the computational group involves comparisons, arithmetic, and the use of simple formulas.

Analytical group encompasses inference, estimation, percentage, and frequencies.

Statistical literacy is concerned with an understanding of concepts such as chance and uncertainty (Lipkus, et al., 2001), sampling variability, margins of error, and randomization in clinical trials (Goldbeck, et al., 2005), and the ability to use such concepts to evaluate scientific information (Ancker & Kaufman, 2007).

Examples of each category are as follows:

- Basic: Identifying numbers of pills to take from a prescription bottle or matching number on bottle with pills.

- Computational: Determining net carbohydrates, calories, or nutrients based on information from a label or computing Body Mass Index (BMI).
- Analytical: Determining whether cholesterol levels are within a normal range, comparing insurance benefits across companies.
- Statistical: Understanding risk, life expectancy, and methods of randomized controlled trials (RCTs) in determining safety and efficacy.

Numeracy assessment is evidently important in the domain of patient education in understanding treatment options and in the personalization of the elicitation process (Omidbaksh, et al., Oct., 2010). The quantitive information necessary in order for patients to compare different treatment options and consequently to make decisions compatible with their own preferences has to be personalized according to the numeracy level of each patient. However, there are considerable variations in the numeracy skills which patients need to comprehend the risks and benefits of their treatment options. Therefore, the level of numeracy of each patient needs to be assessed. Figure 2.1 shows the definition of numeracy and its relation with other related enviromental factors such as emotions, language, etc. and summarizes the related literature based on the survey performed in this research.

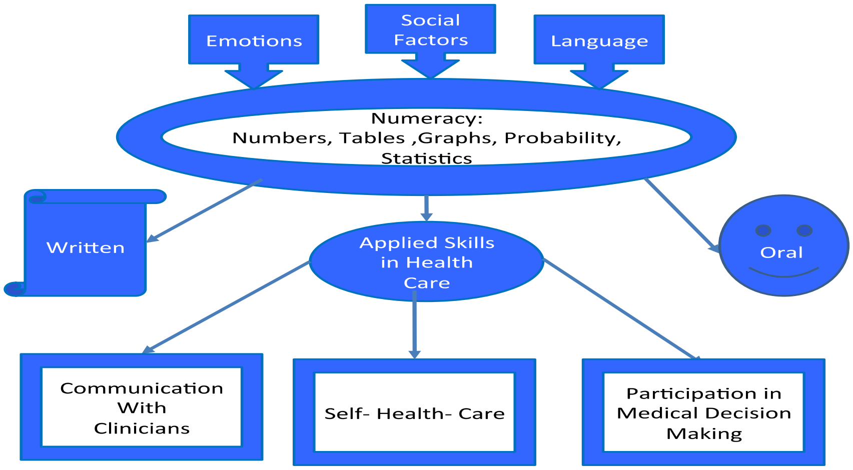

Emotions
Social Factors
Language
Numeracy:
Numbers, Tables ,Graphs, Probability, Statistics
Written
Applied Skills in Health Care
Oral
Communication With Clinicians
Self- Health- Care
Participation in Medical Decision Making

Figure 2.1 Numeracy Skill and Environmental Factors

The organization of the rest of this chapter is as follows: In Section 2.2, we summarize the literature on approaches to numeracy assessment for patients. We, then, review the computer adaptive testing, the feasibility of using it for healthcare domain, and discuss its shortcomings in Section 2.3 and Section 2.4 subsequently. In Section 2.5, we explain in what ways our approach is similar or different from the existing methods. We review the background concerning the empirical investigation approaches and discuss our approach for empirical studies, in Section 2.6 and finally, we conclude the chapter in Section 2.7.

2.2 Related Work

There have been several methods for numeracy assessment in the literature. Some are considered as more established standard pioneers and some are developed in the more recent years. In (Woloshin, et al., 2000), (Schwartz, et al., 1997), Schwartz et al. assessed patients' numeracy with three questions and scored it as the total number of correct responses. In (Lipkus, et al., 2001), Lipkus et al. evaluated a set of eleven questions that compose more questions that directly evalute the patients' ability of risk understanding.

Rapid Estimate of Adult Literacy in Medicine (REALM) (Davis, et al., 1991) measures the individual's ability to read common medical words and lay terms for parts of body and illness.

Wide Range Achievement Test (WRAT-3), (Weintraub, 2000) assesses basic skills in reading, arithmetic, and spelling. The test takes approximately 30 minutes to administer. Test of Functional Health Literacy in Adults (TOFHLA) is designed in two parts: 17-item numeracy questions and 50-item reading comprehension questions with three passages. It uses actual health-related materials such as prescription bottle labels and appointment slips. S-TOFHLA, a shortened version of TOFHLA, consists of 4 numeracy questions and 36 reading comprehension with two passages. It needs half a time for administration compared with TOFHLA.

Medical Data Interpretation Test (MDIT) (Woloshin, 2005) assesses the individual's ability to interpret and understand medical statistics and understand concepts regarding risk. The test includes 18 questions based on the individual’s daily encounter with health information.

The Newest Vital Sign (Weiss, et al., 2005) is another functional test it consists only six questions based on the nutrition label states.

Subjective Numeracy Scale (Fagerlin, et al., 2007) is designed and developed as a subjective numeracy measure claiming that is more useful than objective numeracy measures.

STAT-Confidence Scale (Woloshin, et al., 2005) is another subjective test which includes only three questions about the confidence in the medical statistics of patients.

Asthma Numeracy Skills (Apter, 2006) assesses the understanding of numerical concepts in asthma self-management instructions with a 4-item Asthma Numeracy Questionnaire.

Diabetes Numeracy Test (DNT), with 43 items (Huizinga, 2008) is an assessment tests for investigating the numeracy skills in patients with diabetes.

Warfarin Management Test in patients, assesses the patients’ ability for taking warfarin (an anticoagulant) to handle basic numerical concepts (Estrada, 2004).

Numeracy Understanding in Medicine Instrument (NUMi) (Schapira, et al., 2012) is based on using item response theory scaling methods. The test has 20 items with an item bank calibrated with 1000 patients. Table 2.1 represents the different categories of numeracy assessment approaches. Table 2.2 summarizes different numeracy methods along the category, type and number of questions.

Table 2.1 Categories of Numeracy Assessment Approaches

Category/Type	1	2
1	General (C1.1)	Disease Specific (C1.2)
2	Objective (C2.1)	Subjective (C2.2)
3	Composite (C3.1)	Numeracy Focused (C3.2)
4	Basic Skills (C4.1)	Higher Skills (C4.2)

The existing numeracy methods have some limitations (Huizigna, et al., 2009): first, none includes the full set of skills and knowledge associated with numeracy. Second, potential confounders such as test anxiety, and distress are not taken into account and at last high-end means of communication and technology are not considered in the assessment.

To obtain reliable measurement, specific health education interventions should be individually tailored for patients (Lipkus, et al., 2008) (Davis, et al., 2002) and the numeracy level of patients should be assessed.

Table 2.2 Numeracy Assessment Methods

Name	Abbreviation	Category/Type	#Questions
Schwartz	-	(C1.1)(C2.1)(C3.2)(C4.2)	3
Lipkus	-	(C1.1)(C2.1)	11
Slosson Oral Reading Test–Revised	SORT-R	(C1.1)(C2.1)	50 Score Sheet Word list
Rapid Estimate of Adult Literacy in Medicine	REALM	(C1.1)(C2.1)	3 lists
Wide Range Achievement Test	WRAT-3	(C1.1)(C2.1)	reading15 letters and pronounce 42 words 55 spelling +55arithmetic
National Adult Literacy Survey	NALS	(C1.1)(C2.1)(C4.1)	-
Test of Functional Health Literacy in Adults	TOFHLA	(C3.1)(C2.1)	50 reading+17numeracy
Short TOFHLA	S-TOFHLA	(C3.1)(C2.1)	36 reading + 4 numeracy
Medical Data Interpretation Test	MDIT	(C1.1)(C2.1)	18
The Newest Vital Sign	-	(C3.1)(C2.1)	6
Subjective Numeracy Scale	SNS	(C2, 2)	8
STAT-Confidence scale	-	(C2, 2)	3
Asthma Numeracy Skills	-	(C1.2)(C2.1)	4
Diabetes Numeracy Test	-	(C1.2)(C2.1)	43
Warfarin Management	-	(C1.2)(C2.1)	3+

2.3 Computerized Adaptive Testing

Computerized Adaptive Testing (CAT), also called tailored testing, is a method for administering tests that dynamically adapts to each individual patient's skill or knowledge level. Unlike fixed-number exams in which the same number of questions is presented to all patients in the same order, adaptive exams choose questions based on the patient's performance level; not all patients receive the same set or the same number of questions. This reduces the time patients spend and the frustration they face in answering questions.

When a patient answers a question in an adaptive test correctly, a question with a higher difficulty rating is presented to them. If their answer is not correct, an easier question is asked. After each and every question is answered, the patient's skill level is re-evaluated and the next appropriate question to be asked is determined. The patient’s score is determined, not based on the number of right or wrong answers given, but rather on the average difficultly level of the questions answered correctly.

CAT has been successful in education evaluation; however, the feasibility of its usage is arguable. Patients’ health outcomes, needs, status or even preferences are elicited from different questionnaires. Generally, short questionnaires are more favorable for patients. Though there is a compromise in precision and reliability in favor of practicality for short questionnaires.

Item Response Theory (IRT) is a statistical framework that is predominantly used for CAT. In the next Section, we review Response Theory models.

2.3.1 Item Response Theory Models

Item Response Theory applies mathematical models for predicting a set of ability scores by linking actual performance on test items, item statistics, and examinee abilities (Reckase, 1981). IRT calculates the probability that an examinee may answer a specific item correctly. The probability of correctly answering increases, if the examinee’s ability is at a higher level than the difficulty level of that item. Item characteristic curve describes the relationship between the examinee's item performance and the abilities underlying item performance. Each item contains one, two or three parameters namely discrimination, difficulty, and guessing parameters. Based on these parameters the appropriate item is adapted for each examinee. In this way, the items are selected from the item bank corresponding to the estimated ability level of examinees.

In IRT-based CAT, item characteristic functions which are essential for the functioning of valid tests are determined during calibration. Good calibration is essential for effective adaptive testing. Calibration strategy, size of item bank, item response model and size of calibration sample are the parameters that should be considered. As item calibration needs huge sets of test sessions previously done by examinees, the major issue for calibration is limited access to examinees in occupational settings. An effective way to address this issue is online calibration, which is estimating the parameters of pre-test items which are presented to examinees during the course of their testing with operational items. In IRT-based CAT, the items are selected from an IRT calibrated item bank either by Bayesian method or by maximum likelihood information (Hambleton, et al., 1991).

IRT models are classified based on the number of item parameters and the type of items. IRT models that include tests with binary item responses such as multiple-choice items are called dichotomous and those that include tests with items which have more than two response categories are called polytomous (Thissen & R. J. Mislevy, 2000). The former IRT models can be the one-, two- and three-parameter logistic model (1PL, 2PL and 3PL) respectively: Rasch and Birnbaum models. The equation for the Rasch (Single-parameter logistic) model, Two-parameter logistic model and Three-parameter logistic model are as the followings: (where e is the base of natural logarithm: 2.71828)

Equation 2.1

$$p_i(\theta) = 1 + \frac{1}{1 + e^{(\theta - b_i)}}$$

Equation 2.2

$$p_i(\theta) = 1 + \frac{1}{1 + e^{-a_i\,(\theta - b_i)}}$$

Equation 2.3

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i\,(\theta - b_i)}}$$

i: item number; *i*=1, 2, 3, …, N, and N is the total number of items

θ: examinee's ability

a_i : discrimination parameter of item *i*

b_i : difficulty parameter of item *i*

c_i : guess parameter of item *i*

Some of the important challenges in IRT-based testing are related to (a) item bank calibration, (b) over-exposure control, (c) test security, and (d) examinee review allowance. Although the problem of protecting the content of the items from public knowledge to prevent cheating is obviously important in educational testing, it is not a concern in the health care field. Therefore, the challenges of item over-exposure and test security are not our focus. Generally speaking, there is no incentive for cheating in this field and it is in the best interest of all patients to give the right answers. Item bank calibration and examinee review allowance are the two major challenges for CAT applied in healthcare.

In brief, although IRT-based CAT models are theoretically efficient, due to issues regarding item banks, administration, and security, they tend to be difficult for development and maintenance (Weiss & Kingsbury, 1984). IRT-based CAT seems very attempting for numeracy assessment in healthcare, though there are some concerns in its usage that we discuss in the next Section.

2.4 Shortcomings of Existing Approaches to Patient Numeracy Assessment

We believe that IRT-based CAT is not appropriate for patient numeracy assessment. Firstly, item pools should be developed extensively. IRT-based CAT systems need to select presentation of items based on the item response characteristic curves from large data sets.

Furthermore, there is a continuous need for item bank expansion and refinement to be supported.

Thirdly, item calibration is a complex procedure, which is the prerequisite of IRT-based CAT. Item calibration is concerned with data analysis procedures that provide estimates of IRT for each test question. All the items in the item bank need to be tested on a large number of patients and as item calibration needs huge sets of test sessions previously done by patients; the major issue for calibration is limited access to patients especially in medical settings, in which gathering patients' data is involved with complex and exhaustive process of informed consent.

"Informed consent is an on-going process that starts with the researcher's first contact with the individual and continues until the study is complete or the participant withdraws. Any discussion of informed consent with the participant, the written informed consent form and any other written information given to participants should provide adequate information for the participant to make an informed decision about his/her participation [in research]." (Health Canada Official, 2015)

Generally, 200 to 1000 patients are required to adjust parameters of IRT test for item calibration (Wainer & Mislevy, 2000) and in the healthcare domain having access to this number of patients is not facile.

Furthermore, the uni-dimensionality assumption of IRT model is not satisfied in our case. Most of educational and psychological tests are often multidimensional. IRT models focus only on one dimension to estimate the ability level of patients. We claim that we need a multidimensional model for the assessment of numeracy.

Lastly, as the reality is multidimensional, we consider at least two major dimensions namely, knowledge and confidence for our patient numeracy assessment. We need more than one dimension for our work. So, we examine multidimensional CAT and we investigate it further as below.

For our problem, we consider each test item has to ‘reflect’ one or more underlying dimensions of the ability and we assume that the test is dichotomous which includes binary (correct/incorrect) responses. Following Goldstein and Wood, (Goldstein & Wood, 1989) notation, if f_j represents the factor score for patient *j* and π_{ij} represents the probability that patient *j* responds correctly to item *i*. Then a simple item response model is:

Equation 2.4

$$\pi_{ij} = ai + b_i f_j$$

A reasonable estimate of f_j is calculated by the ‘raw score’ i.e. percentage (or total) of items answered correctly. If we use b_i as a weight, a more efficient estimate is given by a weighted percentage as in IRT model:

Equation 2.5

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \operatorname{logit}(\pi_{ij}) = a_i + b_i f_j$$

However, we intend to apply a unique two-dimensional assessment process of CBL in which a single answer for each question generates two factors simultaneously; knowledge and confidence. Thus, by including the confidence factor we characterize the patient’s ability with two underlying factors. Our assumption is that the two factors of knowledge and confidence are independent. Confidence is measured based on the measurement discussed in Section 3.3. Consequently, the logistic model can be generalized as follows:

Equation 2.6

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\operatorname{logit}(\pi_{ij}) = a_i + b_{1i} f_{1j} + b_{2i} f_{2j}$$

Thus, we need to use a multidimensional model. Multidimensional IRT models model response data hypothesized to arise from multiple abilities, unlike uni-dimensional models which require a single ability dimension. However, because of the greatly increased complexity, the majority of IRT applications apply only a uni-dimensional model for modeling the data known to be multidimensional. In doing so, there may be incorrect inferences about characteristics of the items (e.g., discriminations) as well as about patient's proficiency (Kang, 2006).

Therefore, the existing CAT models are not well suited for our domain and we need a new model for the assessment of numeracy. In the next Section, we compare our numeracy assessment approach with the existing ones.

2.5 Confidence-based Patient Numeracy Assessment Model

We proposed a confidence based adaptive testing model (Omidbaksh & Radhakrishnan, May, 2014) that assesses the patients' numeracy skill by integrating the parameter of confidence in the adaptive assessment. In (Omidbaksh & Ormandjieva, Dec., 2015), we introduced our goal-driven modeling for Confidence-based Patient Numeracy Assessment named C-PNA.

We developed a novel model to measure different objective and subjective quality characteristics of numeracy assessment methods (Omidbaksh & Ormandjieva, July, 2016). The objective characteristics of our hierarchal quality model are composed of four characteristics: Accuracy, Effectiveness, Productivity and Usage Efficiency and the subjective characteristics are composed of Satisfaction, Discretionary Usage and Trust at one layer and Comfort, Pleasure and Understandability at another layer.

We compared our confidence-based numeracy assessment method C-PNA with well-established numeracy assessment methods mentioned in Table 2.2 based on the categories listed in Table 2.1. C-PNA has the full coverage of all categories and types for numeracy assessment approaches as shown in Table 2.3. We could simplify this coverage as follows: (C1,1) (C1,2), (C2,1), (C2,2), (C3,1), (C3,2), (C4,1), (C4,2).

Table 2.3 Coverage of Numeracy Categories C-PNA

C-PNA	**1**	**2**
1	General (C1.1) ✓	Disease Specific (C1.2)□ ✓
2	Objective (C2.1)□ ✓	Subjective (C2.2)□ ✓
3	Composite (C3.1)□ ✓	Numeracy Focused (C3.2)□ ✓
4	Basic Skills (C4.1)□ ✓	Higher Skills (C4.2)□ ✓

We also illustrated the strength of C-PNA compared to any IRT-based model as in Table 2.4. IRT-based models need extensive item pool creation, extension and item calibration which made them very time/effort consuming. However, C-PNA is multidimensional and needs less time/effort consuming.

Table 2.4 Comparison of C-PNA with IRT models

Requirements	C-PNA	IRT
Item Pool Creation	✓	✓
Item Calibration	✓	✓
Highly Effort/Time Consuming	✓	✓
Multidimensional	✓	✓
Item Pool Expansion	✓	✓

2.6 Empirical Studies in Software

An experimental design is a complete plan for researchers to apply various experimental conditions to the subjects in an experiment, to determine how the conditions affect behavior or the results of some activities (Fenton & Bieman, 2014). We needed to plan how the application of these conditions would help us to support or refute our hypotheses. Controlled experiments, case studies, and survey research are empirical investigations applied in Software engineering.

In the next section, we explain the methods that are most likely to be applied in software engineering and adapted from a number of different fields (Easterbrook, et al., 2008). We review the assessment techniques available, and provide guidelines for applying the method to empirically assess whether or not the research objectives were achieved.

2.6.1 Controlled Experiment

A controlled experiment is an investigation of a testable hypothesis that one or more independent variables are manipulated to measure their effect on one or more dependent variables (Fenton & Bieman, 2014). Controlled experiments help us determine precisely how the variables are related, and, specifically, whether a cause-effect relationship exists between them. They must be planned in advance. Each combination of values of the independent variables is called a treatment (Easterbrook, et al., 2008).

The simplest experiments involve only two treatments, representing two levels of a single independent variable (e.g. using a tool vs. not using a tool). When the experimental designs are more complex, more than two levels, or more than one independent variable are involved. In most software engineering experiments, we need human subjects to perform some tasks and then we measure the effect of the treatments on the subjects. The Empirical studies that include observations in which potential confounding variables cannot be controlled and/or subjects cannot be assigned to treatment or control groups are called observational studies, natural experiments, or quasi experiments (Fenton & Bieman, 2014).

2.6.2 Case Study

A case study is a quasi-experiment in which the key factors affecting the outcome of an activity are specified, and then the inputs, constraints, resources, and outputs of the activity are documented (Fenton & Bieman, 2014). The term *case study* is often applied to mean a working example. But a case study is considered different as an empirical method. Yin (Yin, 2009) introduces the case study as “an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident.” Case studies can be retrospective or planned.

2.6.3 Survey

A survey is a retrospective study of a situation that attempts to document the relationships and outcomes. After the occurrence of an event, the survey is carried out. Therefore, researchers have no control over the activity that is under study in the performance of the survey. As survey is retrospective, a situation can be recorded and compared with the similar ones. However, variables cannot be manipulated unlike controlled experiments and case studies.

2.6.4 Our Empirical Investigation Approach

The guidelines proposed by (Fenton & Bieman, 2014) are followed in order to choose the method appropriate for our research. Controlled experiments engage small numbers of people or events and demand high supervision. They are best described as '**research in the small**.' Case studies generally focus on at a typical project and do not attempt to capture information about all possible cases. They are known as '**research in the typical**.' Surveys referred to as '**research in the large'** are applied to record the general views over large groups of projects**.**

In this work, we chose mainly controlled experiments for the following reasons:

1. Our investigation is planned, and not retrospective.
2. The treatments that we propose have not been applied previously.
3. The replication cost of conducting the studies is low.
4. We have high level of replication in our study. As the same test was carried out many times, with different apps, different types of UI, and different types of patients.
5. There were limited number of participants in the study whom were carefully controlled.
6. We had a high level of control over the variables that could have impact on the outcome.

2.7 Summary

In this chapter, we presented the state of the art of the existing numeracy assessment approaches in healthcare. We showed how Computerized Adaptive Testing does not suffice for patient numeracy assessment. We compared our C-PNA model with the existing related work based on the coverage of numeracy skill categories. Moreover, we confirmed that C-PNA requires less effort and time for item bank creation, management, and item bank calibration in comparison to IRT-based models. C-PNA produces results as accurate and productive as them; along with obtaining higher satisfaction and trust in patients (will be discussed in Chapter 6).

In the next chapter, we introduce our adaptive testing model for patients' numeracy skill assessment.

Chapter 3 An Adative Testing Model for Assessment of Numeracy in Patients

In the past decade, there has been an increasing demand for development of new techniques for assessment of numeracy skill in healthcare domain. Patients' ability in understanding risk, uncertainty and probabilities and making trade-offs between benefits and harms play an essential role in their active participation in the decision making that surrounds their health care. To achieve reliable and accurate preferences, numeracy skill of patients should be assessed so that information regarding the risks and benefits of their treatment options can be presented in the readily comprehensible and perceivable format to them. Although there are several numeracy assessment measurements in healthcare, they all seem to lack deliberation of potential confounders such as confidence and anxiety.

We state that patients are a special group of users with diverse skill levels whose computer interaction for this purpose is relatively more critical than that of other groups of users. The existing testing models are not applicable for immediate use with this group. They require the collection of lots of information, which is burdensome to patients. The aim is to reduce patient burden without compromising the precision of the test. We need a new adaptive testing model that dynamically assesses the numeracy level of patients in a way that not only obtains reliable results but also suits patients better. If patients take responsibility to participate in making decisions about their treatments, they have to possess the numeracy skill necessary and they have to be confident enough to use that skill. However, there is no model, which takes confidence as a parameter in consideration for adaptive assessment. We incorporate the parameter of confidence in the adaptive testing. We propose a model that reflects both knowledge and confidence in the assessment and claim that the result of our assessment is more reliable. Consequently, this approach leads to behavioural outcomes and empowers patients.

In this chapter, we aim to address our Objective 1 (see Section1.2), to design an adaptive testing model for the assessment of patient numeracy skill.

3.1 Confidence-based Patient Numeracy Assessment (C-PNA) Model

Obtaining numeracy skills involves the development of math knowledge to be able to reason and apply basic numerical concepts. Patients, if numerically literate, need to be able to manage and respond to the mathematical demands of their health decisions. They, like all human beings, are majorly engaged with the affective variables with this development. Researchers in the field of educational psychology have been studying the relationship between affective variables and academic achievement for many years (Cokley, 2000) (Lopez, et al., 1997).

Strawderman (Strawderman, 2009) focuses on the math anxiety and suggests that there are initially three major dimensions, which are involved with the development of math knowledge. Consequently, she proposes a model for math development with the three dimensions. The dimensions are social/motivational dimension, intellectual/educational dimension and psychological/emotional dimension. The dimensions may have some overlaps and their boundaries are not well defined. Associated with each dimension is a continuum on which it is assumed that any patient at any particular time may be situated.

The social/motivational dimension is comprised of the forces that influence an individual through the agencies of family, friends, and society as a whole. The continuum associated with this dimension is behavior and this continuum has pursuit and avoidance at its two extremes.

Pursuit and avoidance are logical consequences of the value placed on mathematics, which are performed by the individual but influenced by the attitudes of significant others and by society in general. The intellectual/educational dimension includes the knowledge and skills an individual possesses (or expected to obtain) and their perception of success or failure in them. It is formed of those influences that are cognitive in nature. The continuum associated with this dimension is achievement, where individual perception is important. At the extremes of the achievement continuum lie success and failure that are the subjective evaluation regarding one's acquisition or use of mathematics skill and concepts.

The psychological/emotional dimension includes the individual’s emotional history, reactions to stimuli and arousal states, formed by the faculties that are affective in nature. Therefore, the continuum associated with this dimension is feelings and thus anxiety and confidence are at the either end of the feelings continuum.

The extremes of the three continua form positive and negative cycles. In the positive cycle, an individual who is successful in the use of mathematics will be more confident in situation involving math and more likely to pursue use of mathematics. Betz (Betz, 1978) indicates that the more confident an individual is toward using mathematics, the more likely they are to be successful in such tasks and the more confident the individual is toward learning and using mathematics, the more likely they are to pursue the study of mathematics. The negative cycle operates in the same way. Tobias and Weissbrod (Tobias & Weissbrod, 1980) believe failure in mathematics is an antecedent to math anxiety. When individuals avoid mathematics, it would be possibly the result of perceived or actual failure. Figure 3.1 illustrates the three dimensions and the three continua of math knowledge.

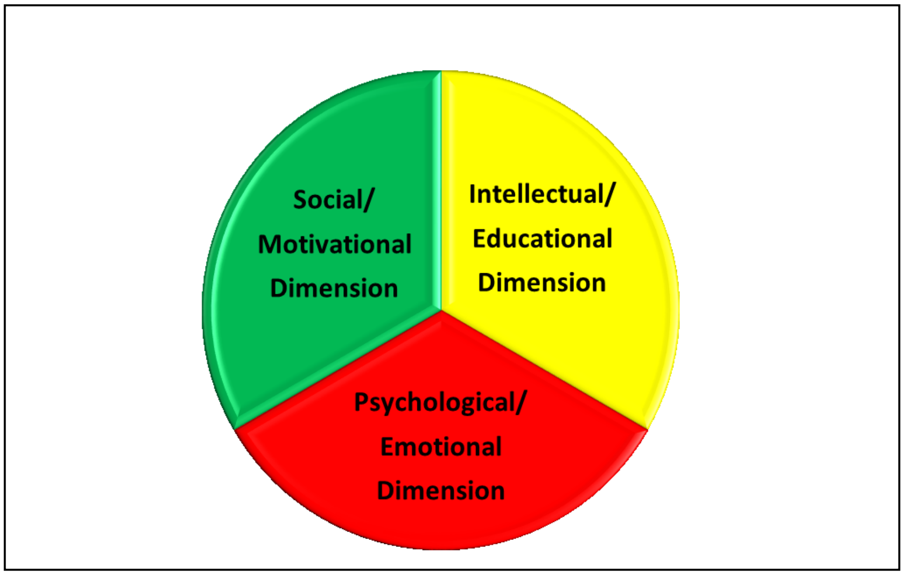


Figure 3.1 Multiple Dimensions of Math Knowledge (Strawderman, 2009)

In addition to the three dimensions mentioned, cognition, specifically the role of understanding may also contribute to the development of math knowledge. Ashcraft

(Ashcraft, 2001) indicates that math anxiety may inhibit certain cognitive functions that are necessary for learning mathematics. Hence, math anxiety would be possibly the reason for not understanding mathematics. Subsequently, learning is considered for the model, with respect to its role in how people move between the positive cycle and the negative cycle. The learning continuum has understanding and rote learning at its extremes. Learning by rote and learning with understanding are different processes and have very different outcomes. Skemp (Skemp, 1971) states that there are major differences between individuals who have learned with understanding and those who have learned by rote. Carpenter et al. (Carpenter, et al., 1981) indicate that individuals who have learned by rote cannot easily apply learned skills in solving problems.

Figure 3.2 shows the conceptual model for math anxiety including both affective and cognitive influence.

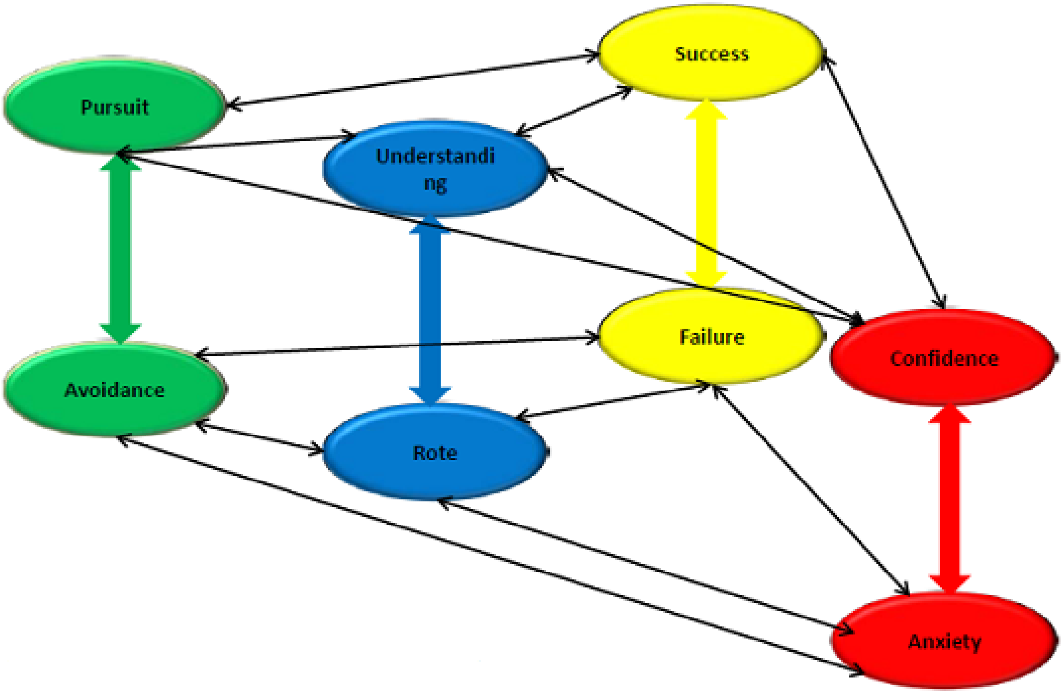


Figure 3.2 Conceptual Model for Math Anxiety (Strawderman, 2009)

In order to assess patients' numeracy, their psychology and their behavior towards learning should be majorly taken into consideration. The relationship between affective variables of behavior, feelings, learning and achievements is strongly noticeable for the patients in this matter.

Patients are reluctant on the attitudes of family, friends, clinicians and health professionals. Thus, the social/motivational dimension is completely explicable for them and their pursuit and avoidance on the subject are definable.

The knowledge and skills patients are expected to obtain and the perception of success or failure in them form the intellectual/educational dimension and, success and failure are considered two ranges in this domain.

Furthermore, patient's emotional history, reactions to stimuli and arousal states, which are not negligible in their attitude, justify the psychological/emotional dimension. This dimension brings anxiety and confidence into the picture. Moreover, the role of patients' understanding, cognition, is considerable in the development of math knowledge, as the learning of an individual is either by understanding or by rote. Therefore, the above-mentioned discussion led us to redefine theses dimensions corresponding to our domain. The four dimensions: learning, feelings, behavior and achievement exist for patients. We define the learning dimension with the two ranges understanding and rote, as parameter *X1* and assign the Difficulty as its value. We define the dimension of feelings with the two ranges anxiety and confidence, as parameter *X2* and assign the value of Confidence Level to this parameter,

Likewise, we define the dimension of behavior with two ranges pursuit and avoidance as parameter *X3* and assign the value of *Pursuit Level* to it. Finally, we rename the dimension of achievement with the ranges success and failure, as parameter *X4* and assign the value of *Success* to it.

Generally, there is a relationship between these parameters and they have positive and negative effects on each other. Studies reveal that there is a direct link between health numeracy and emotions, and particularly math anxiety (Ashcraft, 2001) (Donelle, et al., 2007) (Eccles & Jacobs, 1986) (Hodge, 1999) (Rothman, et al., 2008). Thus, we work on the relationship among parameters learning, feelings, behavior and achievement (*X1, X2, X3* and *X4*) as presented in Figure 3.3. We formulate this relationship simply as: *learning(X1), feelings(X2), behavior(X3) => achievement(X4)*

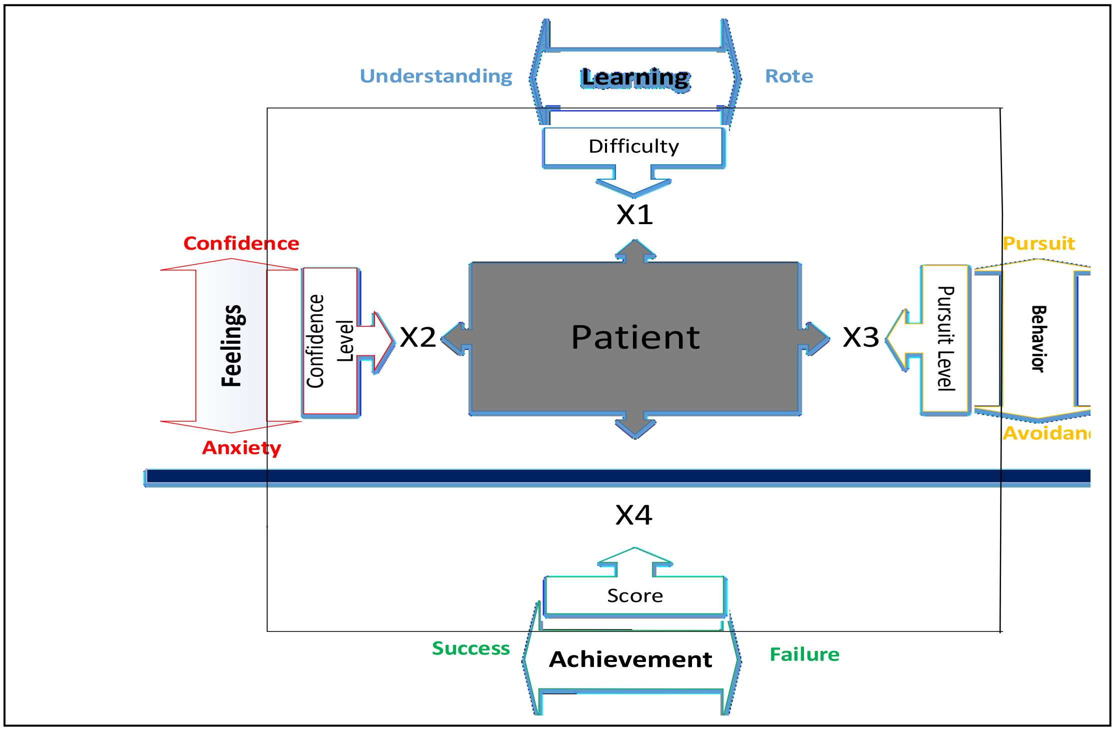


Figure 3.3 Our Model for C-PNA

Figure 3.3 shows our novel C-PNA model proposed in this research. As a means to assess the numeracy skills of patients, we intend to calculate the values for each parameter. In this manner, it would be possible to assess the achievement of an individual by considering the values of learning, feelings and behavior parameters.

Normally, the type of a question or an item reflects if the answer to that question would be given by understanding or by rote. In other words, the response shows if the individual has learnt the knowledge in question by rote or by understanding. For this purpose, we specify the items in our item bank by rote/understanding type and we name this specification the difficulty of the item. Therefore, for each item (*i*) in the item bank, we assign a *Difficulty* (*di*).

Furthermore, the feelings parameter, which ranges from anxiety to confidence, could be also estimated by a value. We call this value; Confidence Level and we attempt to apply a method to find estimation for this value.

Moreover, the behavior parameter can be defined by a binary value in each and every step of the assessment. The value of this parameter, *Pursuit Level*, is estimated by the patient's willingness to quit or continue the assessment.

In this way, by summing up these values, we could end up with a value for the parameter of achievement and we call this value *Score*. This *Score* not only represents the correctness of an item but also includes three different dimensions regarding the answer to that item.

3.2 Adaptive Testing for C-PNA

As we intend to create adaptive tests for the assessment of patients' numeracy, we need tests that dynamically adapt to each patient's skill or knowledge level. In adaptive tests, questions are selected in such a manner to maximize the precision of the test and generally, fewer questions are needed for accurate scores. The accuracy of the test has much more importance for patients. Their health decisions rely highly on their understanding of the numerical information and consequently, on their level of numeracy skill in which the information is represented to them. Unlike fixed number tests in which the same number of questions is presented to all patients in the same order, adaptive tests choose questions based on the performance level.

In other words, when a patient answers a question in an adaptive test correctly, a question with a higher difficulty rating is presented to them. If their answer is not correct, an easier question is asked. After each and every question is answered, the skill level is re-evaluated and the next appropriate question to be asked is determined. The score is determined, not based on the number of right or wrong answers given, but rather on the average difficultly level of the questions answered correctly.

Evidently, there are three major tasks for construction of computer adaptive tests: item bank creation, item bank calibration and definition of selection algorithms, along with minor tasks such as specifying scoring procedures, assignment of initial ability and definition of termination criteria.

Item bank creation is the task of providing large data sets needed for establishing items. Item banks should be calibrated afterwards and the possibility for their expansion and refinement should be provided as well. Item calibration refers to data analysis procedures that provide parameters for each test question. Item selection procedures are considered as an important part of adaptive testing algorithms. They determine the choice of the items, which are administered during testing. The item bank of the system is built based on our model. As Figure 3.4 shows that the output of the system based on C-PNA is the numeracy level of the patient.

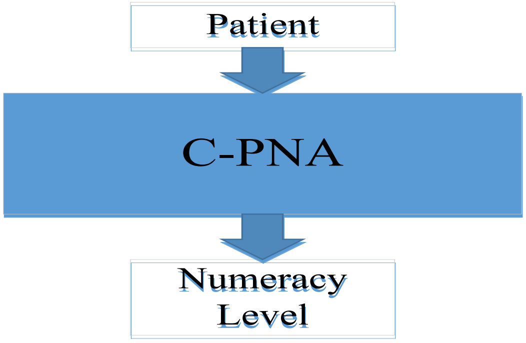


Figure 3.4 C-PNA System

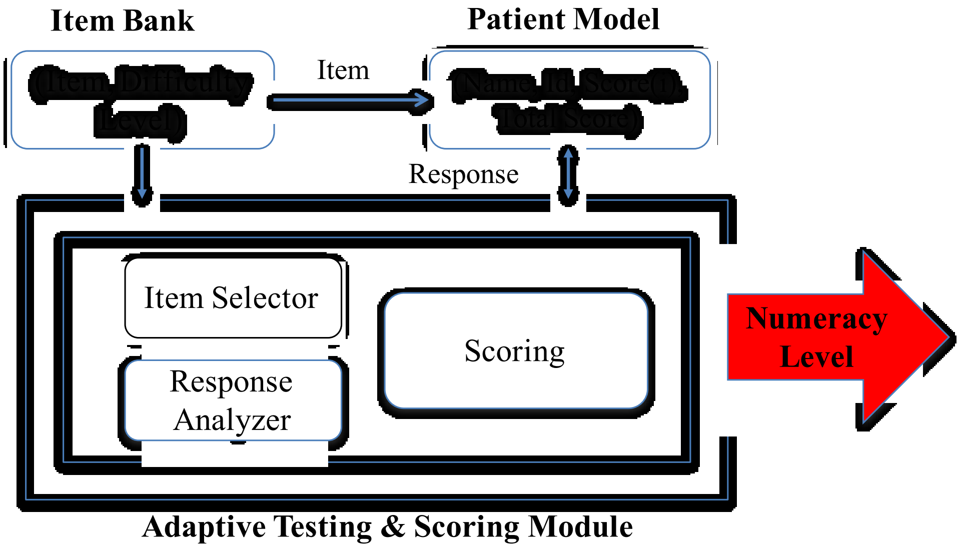


Figure 3.5 Architecture of C-PNA System

Figure 3.5 shows the overall architecture of C-PNA system. The components of our system are:

- An item bank which is comprised of the items along with their difficulty.
- A patient model which includes the basic information about the individual and the information regarding the items asked and the scores acquired on each item and their total score.
- An adaptive testing and scoring module which does three tasks:

1. Selection of the items from the item bank which is done by an item selector based on a selection algorithm.

2. Analyzing the responses received from patients, which is done by a response analyzer.

3. Calculating the score and specifying the level of numeracy based on a scoring method.

The iterative process of adaptive testing occurs as follows (Thissen & R. J. Mislevy, 2000): (1) based on the current ability estimate level, all the items that have not yet been administered are evaluated to determine the next “best” item. (2) The "best" next item is administered for the user to respond. (3) A new ability estimate is computed, given the responses to all the administered items. (4) Steps 1 to 3 are repeated until a stopping condition is reached. This condition could be, for instance, when a fixed test length has been met or pre-specified level of measurement precision such as standard error or error variance is met. This process is depicted in Figure 3.6.

The fundamental elements of any computerized adaptive testing software are as follows (Weiss & Kingsbury, 1984):

- Item bank: An item bank is a necessary element of the system. This item bank has to be calibrated with a model.

- Starting rule: Since no item has been administered at the start of the test, no specific estimate of users‟ ability is calculated. Thus, other initial estimates of user ability may become very useful.

- Adaptive item selection rule: If there is an estimate of patients’ ability, we can select an item that is most appropriate for that estimate, i.e. selecting the item with the most information at that point.

- Ability estimation method: The software updates its estimate of the patients’ ability level, after administration of each item.

- Termination condition: The software has to continually administer items and update the estimate of the ability until either the item bank is exhausted or a termination condition is reached**.**

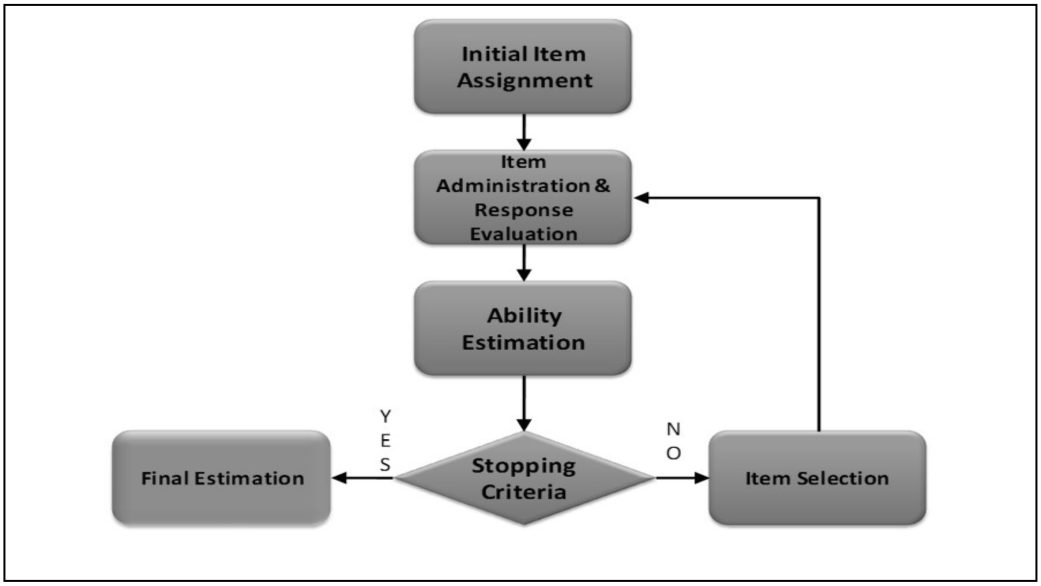


Figure 3.6 Adaptive Testing Flowchart

In the next sections, we discuss how we have estimated a value for Confidence Level and moreover, we focus on our proposed selection algorithm for the application in C-PNA system.

3.3 Confidence Measurement

Confidence is described as a state of being certain as defined in *Merriam-Webster Dictionary*. More specifically, it is described as a state of being certain either that a hypothesis/prediction is correct or that a chosen course of action is the best or the most effective (Hunt, 2003). Researchers namely, Hunt, Shuford, Brown and Bruno (Hunt & Furustig, 1989) (Bruno & Abedi, 1989) (Shuford & Brown, 1973) have established that there is a connection between correctness and confidence and that performance can be better measured by considering confidence along with knowledge. Measuring knowledge alone is prone to guesswork. In other words, knowledge and confidence are correlated. The more confident one is in the knowledge the more likely they can perform better upon it. Knowledge alone is necessary but not sufficient to create behavior (Bruno, 1993). Thus, a better result in measurement is achieved when confidence in the knowledge is also appraised.

In the assessment of numeracy skill of patients, we have to be sure that they have not only the knowledge but also the confidence in that knowledge. If patients take responsibility to participate in making decisions about their treatments, they have to possess the numeracy skill necessary and they have to be confident enough to use that skill.

Only patients who are confidently correct will take productive actions (Bruno, 1993). To rely on the results of the assessment, both knowledge and confidence should be measured. Generally, the result of traditional tests is represented by a score which is the number of correct answers. We propose an approach that reflects both the knowledge and the confidence in the assessment.

In a multiple-choice test, patients are asked to answer an item by choosing among different options. If the answer they choose is the correct answer, they get the score of one. If they don't choose an answer the score is zero and if they guess incorrectly, they get a score of zero or less than zero. Their total score of the test is calculated as the sum of the item scores. This type of test does not include the parameter of confidence in the assessment. Confidence can be measured in a way that patients belief in the correctness of an item is reflected by weighs assigned directly or indirectly to item answers (Echternacht, 1972). Evidently, we need to ask the patient to indicate what they believe the correct answer to be and how certain they are of the correctness of the answer.

One method to include confidence in the assessment is asking two consecutive questions. In the first question, we ask the patient to state what they believe to be the correct answer to a question, and then, in the second question, we ask them to state their confidence in the answer just selected. In such manner, the Confidence Level of the patient in their answers is appraised.

However, Adams (Adams, 2009) states that patients tend to overstate their confidence when knowledge and confidence are measured in two separate questions. Their answer to the confidence question is not a spontaneous reply of their true emotion but a logical reply.

Another method, which was proposed by Bruno (Bruno, et al., 2005), is a unique two-dimensional assessment process that was initially called Information Reference Testing. In this method, a single answer for each question generates simultaneously correctness and confidence metrics. If the patient does not know the answer to a question with multiple choices, they are undoubtedly led to guessing. They will get the credit for the knowledge that they do not possess though for the guessing which happened to be correct. Therefore, he believed in focusing on the knowledge quality rather than the score in assessment. The true knowledge of patient is what they actually know and not what they think they know. Bruno (Bruno, 1990) suggests a method for knowledge assessment that assesses the true knowledge that incorporates non-one-dimensional testing techniques to obtain the subject's knowledge and associated confidence in that knowledge.

Each question has multi-choice answers indicative of Confidence Level including a choice of right answer that indicates complete confidence, choices of wrong answers, at least one choice of combined right and wrong answers that indicates partial confidence, and a choice of no answer that indicates no confidence.

Based on the patient’s answer to the question, a weighted score according to a predefined scoring protocol with the comparison of the patient’s answer with the correct answer is calculated. The patient’s knowledge and confidence is categorized by objective into one of the four knowledge quadrants in the Learning Behavior Model (Bruno, 1993). These quadrants are as follows:

(a) Misinformed:

The patient believes confidently that their knowledge is correct, but it is actually incorrect.

(b) Uninformed:

The patient does not possess the knowledge and may not act and is in a state of paralysis.

(c) Partially informed (doubt):

The patient believes that the knowledge is correct, but an element of doubt exists that may cause the patient not to act on that knowledge.

(d) Fully informed quality (mastery):

The patient confidently knows that the knowledge is correct and it is actually correct.

Table 3.1 illustrates the correlation of knowledge and confidence and the four knowledge quadrants in the Learning Behavior Model (Hunt & Furustig, 1989). This method of assessment tells something about the patient and not just about the chance factors and it brings substantial improvements and savings in the performance of the assessment.

As discussed above the true knowledge of a patient can be measured by considering the value of confidence. Commonly in computerized adaptive testing systems, the next question is selected based on the score of the patient to the previous question. However, this score does not reflect the true knowledge. We believe by integrating the value of confidence in the adaptive testing a more effective assessment can be performed.

Table 3.1 Correlation between Knowledge and Confidence

		Knowledge	
		no	**yes**
Confidence	**no**	Uninformed	Doubt
	yes	Misinformed	Informed

3.4 The Proposed Selection Algorithm

In an adaptive test, when a patient answers a question correctly, a question with a higher difficulty rating is presented to them. If their answer is not correct, an easier question is asked.

We believe that confidence in the knowledge is an important parameter that if evaluated with knowledge, leads to achievement of better results in the assessment. We adopt Bruno's method for measuring confidence and we propose an approach for selecting the questions based on the Score comprised of the Confidence and Pursuit Level of patients.

We propose that for each answer, not only the patient's knowledge but also the patient's confidence in the answer is evaluated and the next appropriate question to be asked is then determined accordingly. Thus, the patient is scored, not on the number of right or wrong answers given, but rather on the average difficulty level of the questions answered correctly. In this way, the test is dynamically adapted to each individual patient's knowledge and Confidence Level.

A single answer to each question generates simultaneously correctness and confidence metrics based on value of confidence. The Knowledge and confidence of the patient can be categorized into one of the four quadrants, namely, informed, misinformed, partially informed and uninformed.

We create an item bank with different types of questions on our testing subject which is numeracy based on the four categories presented by Goldbeck et al. (Goldbeck, et al., 2005). The questions in the item bank have different difficulty levels. We assign different levels of difficulty to the questions on the basis of the understanding-rote parameter. The test starts with a question with difficulty level (DL). This is our initial value of DL and is usually assigned as the medium level. A question is retrieved from our item bank with the difficulty level of DL.

When the patient answers this question, a Confidence Level is calculated for the patient on that question. The Confidence Level falls into one of the four knowledge quadrants. We determine the difficulty level of the next question to be asked by the value of the Confidence Level. If the Confidence Level is either informed or partially informed the difficulty level will be respectively, increased or remain intact until DL reaches a threshold. Otherwise, the difficulty level will decrease. The flowchart of the proposed selection algorithm model is shown in Figure 3.7.

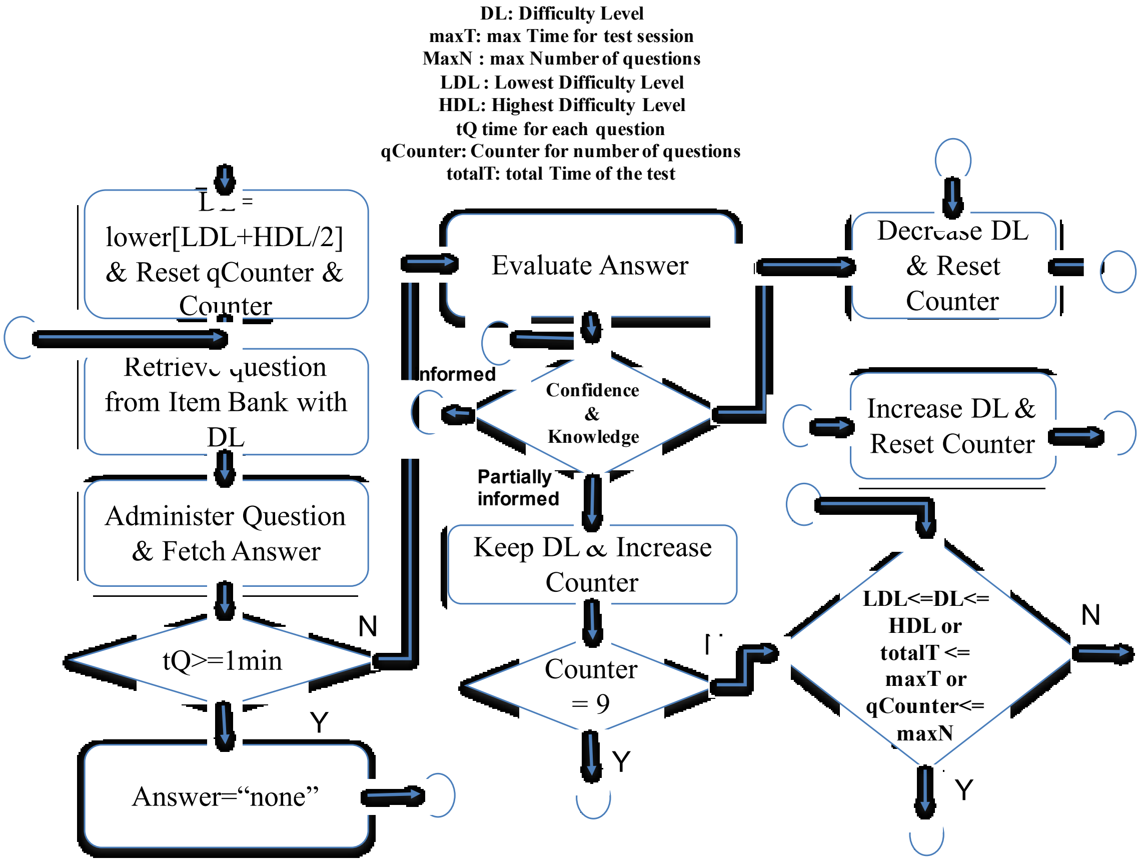

DL: Difficulty Level
maxT: max Time for test session
MaxN : max Number of questions
LDL : Lowest Difficulty Level
HDL: Highest Difficulty Level
tQ time for each question
qCounter: Counter for number of questions
totalT: total Time of the test
DL = lower[LDL+HDL/2] & Reset qCounter & Counter
Retrieve question from Item Bank with DL
Administer Question & Fetch Answer
tQ>=1min
N
Y
Answer="none"
Evaluate Answer
Informed
Confidence & Knowledge
Partially informed
Keep DL & Increase Counter
Counter = 9
Y
Decrease DL & Reset Counter
Increase DL & Reset Counter
LDL<=DL<= HDL or totalT <= maxT or qCounter<= maxN
N
Y

Figure 3.7 Proposed Selection Algorithm for C-PNA

3.5 Scoring Protocol

On the basis of Bruno's measurement, there are three types of answers for each question: one-choice, two-choice and no-answer. The answers are in the form of multiple choices, one-choice answer can be A, B or C. Two-choice answer can be AB, AC or BC and there is also no-answer choice. The type of answer an examinee chooses regarding its correctness or incorrectness has different scores. Based on the examinee's answer to the question, a weighted score according to a predefined scoring protocol with the comparison of the examinee's answer with the correct answer is calculated (Bruno, et al., 2005). Table 3.2 shows the instruction for scoring.

Table 3.2 Scoring Protocol

Type of Answer	**Correct**	**Incorrect**	**Score**
One-choice	X		30
One-choice		X	-100
Two-choice	X		10
Two-choice		X	-100
No-answer or (?)			0

In general, the point score for a correct answer (reward points) can be chosen within the range of +20 to +50 and the point score for an incorrect answer (penalty points) can be chosen within the range -80 to -150, thus the ratio of the absolute values of the point scores for a correct answer to point scores for an incorrect answer is approximately 13.3 % to 62.5%. Single letter answers indicate a high confidence whether the answer is correct or incorrect, though incorrect answers are considered as misinformed.

Under the above mentioned scoring protocol, the examinee is assumed to be "honest" and not to overdo their information. This is very true in the assessment of numeracy for patients. We believe patients are the most honest examinees and they have no incentives to guess at answers.

For each examinee a knowledge profile is created. This profile includes the questions the examinee answered, their score on each question along with their confidence on each answer. For each question, an actual percentage score on the basis of the weighted scoring protocol and a confidence percentage score are calculated. The difference between the actual score and the confidence score shows a degree of misinformation or information gap.

We suppose a test has four questions (items) q1, q2, q3, q4 with correct answers B, A, C, and A respectively. As shown in Table 3.3, if an examinee's answers to these questions happen to be B, ?, BC, B, then based on the examinee's choice type, we can conclude that the first and last questions are replied confidently, the third is replied with partial confident (50%) and the second is unknown (Take1 Column). Furthermore, by comparing the examinee's answers with the correct answers for each question, the correct and incorrect examinee's answers are specified (Take2 Column). So, the knowledge quality level for each question of the examinee is recognized. If an answer is correct with 100% confident, the examinee is considered as informed for that question, if the answer is correct with 50% confident, the examinee is considered as partially informed for that question. Moreover, if the examinee's answer is incorrect with 100% confident, they are considered misinformed. Otherwise, it is unknown. By applying the scoring protocol, the raw score for each question is marked.

For each examinee an aggregate score can be calculated with the total raw score by using the following formula:

Equation 3.1

*Aggregate Score = ((Number of Questions * 100) + Total Raw Score) / (Number of Questions * 130)*

For our example with the number of questions of (4) and the total raw score of (-60), the aggregate score is (4*100+ (-60))/4*130 which is equal to 65%. The aggregated score is correlated to five levels of numeracy skill: very high, high, medium, low and very low numerate as shown in Table 3.4. This grouping will be applied for the purpose of personalization of preferences elicitation process.

Table 3.3 Example of Scores for Different Answers of Questions

Question	Correct Answer	Examinee Answer	Take1	Take2	Confidence Score	Raw Score
q1	B	B	100% confident	Correct	Informed	+30
q2	A	?	Unknown	Incorrect	Uninformed	0
q3	C	BC	50% partially confident	Correct	Partially informed	+10
q4	A	B	100% confident	Incorrect	Misinformed	-100

Table 3.4 Different Levels of Numeracy based on Aggregate Score

No.	Numeracy Level	Aggregate Score
1	very high	92-100
2	high	86-91
3	medium	77-85
4	low	69-76
5	very low	0-68

3.6 Goal-driven Hierarchical Model for C-PNA

Here, we formalize the ideas presented in the previous sections and merge them into one cohesive goal-driven model.

A number of models and frameworks have been developed to support measurement processes based on goals. Basili and his colleagues developed the Goal Question Metrics (GQM) de facto standard (Basili, 1992) (Basili, et al., 1994) (Berander & Jönsson, 2006) (Scholtz, 2004).

The Goal Question (Indicator) Model (GQ(I)M) is an extension of the GQM approach, where the indicators are composed of a set of measures and provide a quantitative answer for the questions (Basili, et al., 1994). We apply the Goal Question (Indicator) Metric approach (Berander & Jönsson, 2006) for C-PNA model. In GQ(I)M approach, which is a top-down approach, the overall goals of the entire organization, corporation, or a single project or group are identified (see Figure 3.8). With respect to the goals that are set up, some questions are generated. Then each question is analyzed in order to identify measurements (indicators and measures) that are needed to answer them. Indicators can be composed of multiple measures that provide quantification and an interpretation of the status of a designated aspect of the assessment.

We explicitly defined our assessment goals and refined them into quantifiable questions and consequently, refined them into a set of indicators and measures for the data to be collected. The quantifiable questions and the related indicators will be used to help the tester achieve the assessment goals. In this way, we built up a patient numeracy assessment model that covers the issues related to numeracy and a set of questions that specifies each issue in a meaningful and quantifiable way. Our aim is to design a patient numeracy assessment, which is more accurate and reliable than the existing ones.

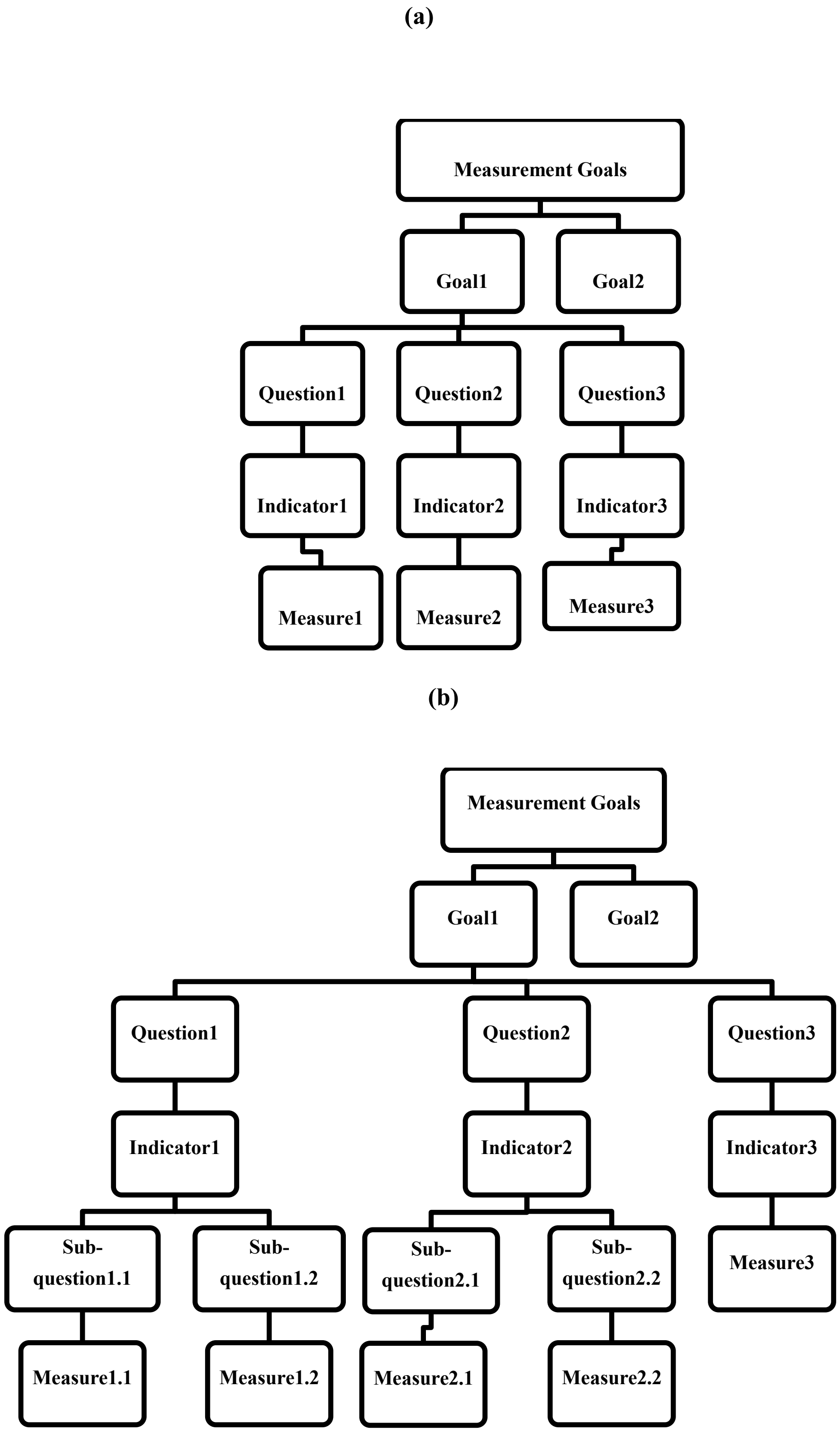


Figure 3.8 GQ(I)M Hierarchical Model (a) (Basili, et al., 1994);(b)Our Adaptation

Table 3.5 and Table 3.6 list the GQ(I)M definitions used in the Confidence-based Patient Numeracy Assessment Model. The goal, questions, sub-questions, indicators and the related measures are clearly defined based on our adapted GQ(I)M structure (as depicted in Figure 3.8).

Table 3.5 GQ(I)M Definition for Patient Numeracy Assessment

Goal	Acronym	To Assess Numeracy Skill of Patients
Question	QSc	What is the numeracy level of the patient?
Indicator	MSc	Aggregate Score
Question	QB	What is the knowledge and confidence of patient in basic numeracy skill?
Measure	MB	Basic knowledge confidence
Question	QC	What is the knowledge and confidence of patient in computational numeracy skill?
Measure	MC	Computational knowledge confidence
Question	QA	What is the knowledge and confidence of patient in analytical numeracy skill?
Measure	MA	Analytical knowledge confidence
Question	QS	What is the knowledge and confidence of patient in statistical numeracy skill?
Measure	MS	Statistical knowledge confidence
Question	QD	How difficult are the questions? (What is the difficulty level of each question?)
Measure	MD	Level of difficulty
Question	QN	How many questions are asked?
Measure	MN	Number of Questions

Table 3.6 Sub-questions for Patient Numeracy Assessment

Goal	Acronym	To Assess Numeracy Skill of Patients
Sub-question	QK.B	Does the patient have the knowledge of basic numeracy?
Measure	MK.B	Basic Numeracy knowledge
Sub-question	QK.C	Does the patient have the knowledge of computational numeracy?
Measure	MK.C	Computational Numeracy Knowledge
Sub-question	QK.A	Does the patient have the knowledge of analytical numeracy?
Measure	MK.A	Analytical Numeracy Knowledge
Sub-question	QK.S	Does the patient have the knowledge of statistical numeracy?
Measure	MK.S	Statistical Numeracy Knowledge
Sub-question	QCo.B	What is the patient confidence in basic numeracy?
Measure	MCo.B	Confidence in Basic Numeracy knowledge
Sub-question	QCo.C	What is the patient confidence in computational numeracy?
Measure	MCo.C	Confidence in Computational Numeracy Knowledge
Sub-question	QCo.A	What is the patient confidence in analytical numeracy?
Measure	MCo.A	Confidence in Analytical Numeracy Knowledge
Sub-question	QCo.S	What is the patient confidence in statistical numeracy?
Measure	MCo.S	Confidence in Statistical Numeracy Knowledge

To measure the numeracy knowledge of patients, we need to assess their numeracy skills in basic, computational, statistical and analytical groups. The type groups basic, computational, analytical and statistical consist of sub-groups. Specifically, the basic group consists of number recognition, fraction, decimal and sequencing. The computational group includes addition, subtraction, multiplication, division, conversion and comparison. The analytical group encompasses inference, estimation, percentage, and frequencies. Statistical literacy is concerned with an understanding of concepts such as chance and uncertainty (Lipkus, et al., 2008), sampling variability, margins of error, and randomization in clinical trials, and the ability to use such concepts to evaluate scientific information (Ancker & Kaufman, 2007). The GQ(I)M hierarchical model of C-PNA Model along with its sub-trees for knowledge and confidence in Numeracy are illustrated in Figure 3.9 to Figure 3.13.

We describe below how the data is interpreted based on GQM. In our GQ(I)M, we defined measures MK.B, MCo.B for quantifying MB which is an indicator for the measurement of the knowledge and confidence of patients in basic numeracy skills, each measure answering questions QK.B and QCo.B (Figure 3.9).

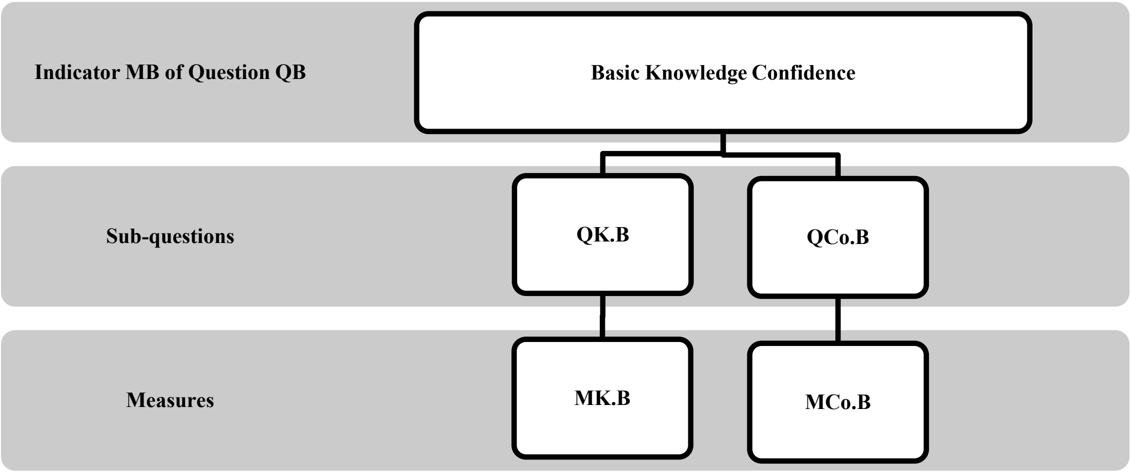


Figure 3.9 Partial Structure for Basic Numeracy Skill

Likewise measures MK.C and MCo.C are defined for quantifying the indicator MC for assessing the computational numeracy skills, answering questions QK.C and QCo.C. (Figure 3.10).

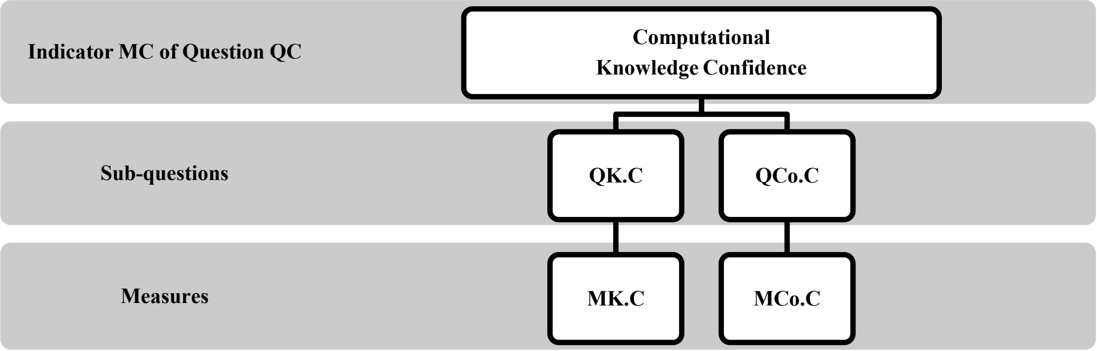

Indicator MC of Question QC
Computational Knowledge Confidence
Sub-questions
QK.C
QCo.C
Measures
MK.C
MCo.C

Figure 3.10 Partial Structure for Computational Numeracy Skill

The measures MK.A and MCo.A are used to calculate the indicator MA, answering QK.A and QCo.A (Figure 3.11).

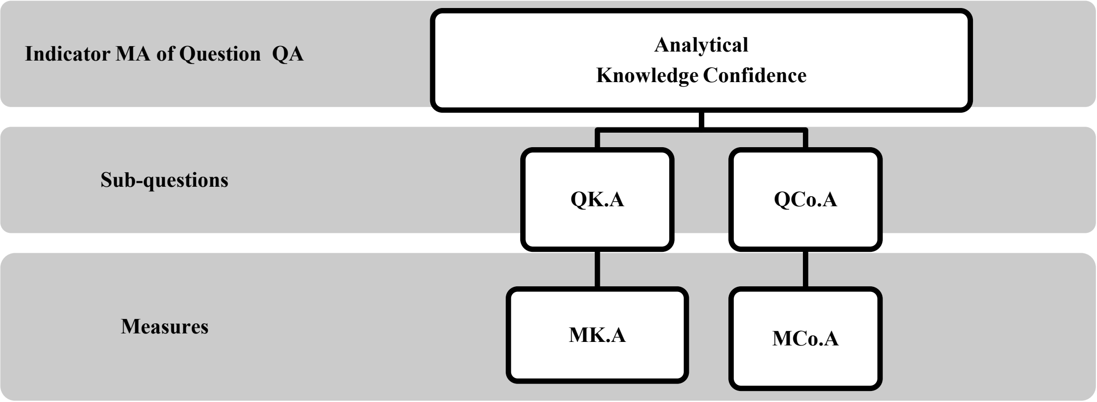


Figure 3.11 Partial Structure for Analytical Numeracy Skill

The measures MK.S and MCo.S are used to calculate the indicator MS, answering QK.S and QCo.S (Figure 3.12).

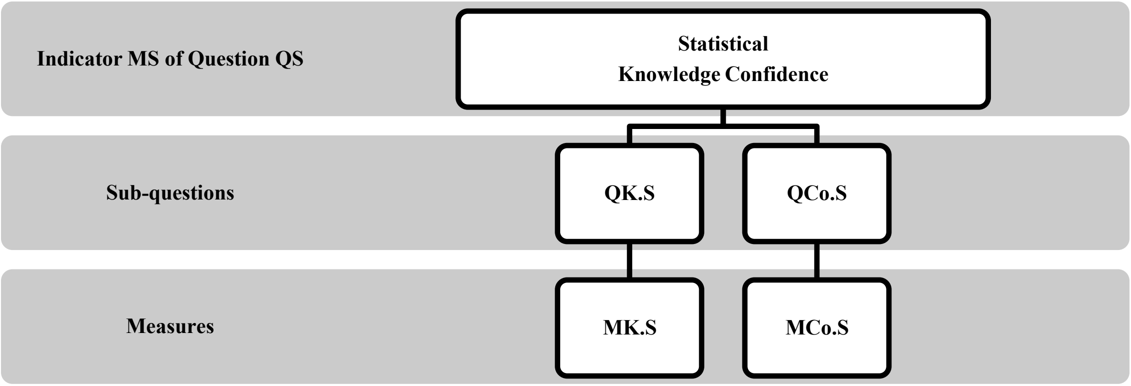


Figure 3.12 Partial Structure for Statistical Numeracy Skill

In turn, MK.B and MCo.B are measures that obtain their value from number recognition, fraction, decimal and sequencing questions, MK.C and MCo.C from addition, subtraction, multiplication, division, conversion and comparison questions, MK.A and MCo.A from percentage recognition, basic graph and risk recognition questions and MK.S and MCo.S from probability comparison/conversion, proportion comparison/conversion and percentage comparison/ conversion questions. Figure 3.13 depicts the root tree of C-PNA model including all the indicators and measures.

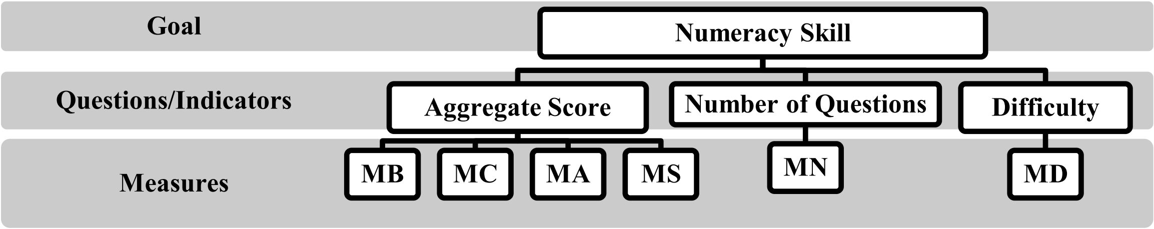


Figure 3.13 Root Tree of C-PNA Model

When MB, MC, MA and MS are each calculated for all questions with different difficulty levels, they are summed up and applied to obtain an aggregate score, which is an indicator of patient numeracy skill. Table 3.7 and Table 3.8 illustrate, respectively, the scale types of measures and indicators defined in our quality model

Table 3.7 Numeracy Skill Measures

Measure	Range of Value	Measurement Method	Scale Type	Subjective/ Objective
MK.A	{Correct, Incorrect}	Ranking	Ordinal	Objective
MK.B	{Correct, Incorrect}	Ranking	Ordinal	Objective
MK.C	{Correct, Incorrect}	Ranking	Ordinal	Objective
MK.S	{Correct, Incorrect}	Ranking	Ordinal	Objective
MCo.A	{misinformed, uninformed, doubt, master}	Ranking	Ordinal	Objective
MCo.B	{misinformed, uninformed, doubt, master}	Ranking	Ordinal	Objective
MCo.C	{misinformed, uninformed, doubt, master}	Ranking	Ordinal	Objective
MCo.S	{misinformed, uninformed, doubt, master}	Ranking	Ordinal	Objective
MD	0-n*	Ranking	Ordinal	Subjective
MN	Non Negative Integer	Counting	Absolute	Objective

*: number of Difficulty Levels

Table 3.8 Numeracy Skill Objective Indicators

Indicator	Values	Measurement Function	Scale Type
MSc	0..100	*	Ratio
MA	0,10,30,-100	Total MA=∑ MA for MD	Ordinal
MB	0,10,30,-100	Total MB=∑ MB for MD	Ordinal
MC	0,10,30,-100	Total MC=∑ MC for MD	Ordinal
MS	0,10,30,-100	Total MS=∑ MS for MD	Ordinal

*: *Total Raw Score = Sum Total (MB, MC, MA, MS),*

*Aggregate Score = ((Number of Questions * 100) + Total Raw Score) / (Number of Questions * 130)*
(See also Equation 3.1)

3.7 Summary

In this chapter, we proposed a goal-driven confidence-based model for patient numeracy assessment adaptive to each individual patient. Our Objective 1 is achieved on the basis of the work described in the chapter. The research contribution lies on the novelty of the assessment model that adapts to each individual patient, covers the full sets of numeracy skills, and considers confidence.

Here, we conclude that there is a need for a quality model for comparing our model with the existing models so the statistical tests are applied on meaningful data. We introduced our quality model in Chapter 4. When using a quality model, the data gathered would be more meaningful and the data correspond to our thesis objectives.

In the next chapter, we present the quality model for the evaluation of different methods for the assessment of numeracy in patients.

Chapter 4 Measuring the Quality of Numeracy Skill Assessment in Health Domain

Numeracy assessment in healthcare domain is noticeably an attractive topic which concerns the evaluation of the level of patients' numerical skill enabling them to understand and perceive the information related to their health.

Although a number of numeracy assessment methods have been used in the health domain, a key limitation of selecting the right method is that no quality model for evaluating numeracy assessment methods is available.

This chapter describes the development of a novel model to measure different objective and subjective quality characteristics of numeracy assessment methods, inspired by the recent standard ISO/IEC 25022 (ISO/IEC DIS, 2016). It provides a framework for the comparison of our method with any other existing numeracy assessment method. In Chapter 6, we use the new quality model to compare numeracy assessment methods. The results of our study demonstrate that our confidence-based adaptive testing method for the assessment of numeracy level of patients, C-PNA, has higher patient Satisfaction, Discretionary Usage and Trust than existing related work along with the same Accuracy, but greater Usage Efficiency and remarkable Effectiveness.

The organization of this chapter is as follows. In Section 4.1, the research methodology is explained; we define our research problem, our objectives and the steps to achieve them. The quality model is introduced in Section 4.2. We describe the objective and subjective characteristics of the quality model. Finally, in Section 4.3, we conclude the chapter and outline the directions of our ongoing research.

4.1 Quality Modeling Methodology

To address our Objective 2 (see Section 1.2), here we present our new quality model, aligned with the ISO/IEC 25022 and adapted specifically to our numeracy assessment method. The model is then used to conduct experiments aiming at evaluating the quality of the new C-PNA model as compared to two classical NC-PNA models: Lipkus (Lipkus, et al., 2001) and NUMi (Schapira, et al., 2012).

To achieve the research objectives, we followed the steps as outlined below:

Step 1: Quality Model. In order to evaluate an assessment model, we had to determine the appropriate objective quality characteristics, which mostly influence the numeracy assessment, namely accuracy, effectiveness, productivity and usage efficiency. Furthermore, we identified the subjective characteristics related to the research problem such as satisfaction, discretionary usage and trust. The measurements designed to quantify these objective and subjective characteristics were also determined.

Step 2: Tool Support. Secondly, we designed and developed a web-based application for the quality evaluation of numeracy assessment to carry on the experiments (Omidbakhsh & Ormandjieva, 2015). After a series of experiments, the quality model was revised and then pilot tested using the Web-based application as described in the next chapter.

Step 3: Empirical Study. The last step of our methodology is concerned with the empirical validation. We selected two classical numeracy assessment methods, Lipkus and NUMi, to enable a pairwise comparative measurement of the quality characteristics, and then we designed and conducted the experiments. Data were collected and validated during the execution of the experiments. These data were then analyzed and a comparison was performed with the results obtained using the alternative methodologies Lipkus and NUMi. This step is discussed in the next chapter.

The empirical investigation provided the evidence about our theory and helped us establish the place of our confidence-based numeracy assessment method among the other numeracy assessment methods. Our novel quality model is introduced next.

4.2 Quality Model

In developing numeracy assessment methods, not only high quality, reliable and efficient assessment is required, but also high personal satisfaction of the users should be taken into the consideration (ISO/IEC DIS, 2016). ISO/IEC 25022 standard provides a quality model definition, which could serve as a customer satisfaction model to ensure that all characteristics of quality are covered from the perspective of each stakeholder.

Here, we introduce our quality model, which is designed specifically for the purpose of numeracy assessment. The quality model is tailored in a way that facilitates the evaluation of such assessment systems in terms of accuracy, effectiveness productivity, usage efficiency, satisfaction, discretionary usage and trust.

For our study, we employed the quality characteristics both objective and subjective. The former is associated to sets of data, which depend only on the object that is measured, however, the latter not only depends on the object that is measured, but also on the viewpoint from which it is taken. The former includes Accuracy, Effectiveness, Productivity, and Usage Efficiency that only depend on the object being measured. On the other hand, the latter includes Comfort, Pleasure, Understandability, Satisfaction, Discretionary Usage and Trust that rely on the user viewpoint from which they are taken as well.

In our hierarchical quality model, the quality characteristics are delineated through several layers. At the root of this structure, there is a division of characteristics into objective and subjective ones.

4.2.1 Objective Characteristics

The quantification of the objective characteristics is based on numerical rules to ensure fairness of the assessment. In other words, it is assured that users produce same measurement results every time the measurement is undertaken on the same source and in the same context. This consistency of measurement is considered very important (Fenton & Bieman, 2014).

Each of the objective characteristics is defined as below.

1) Accuracy: the percentile of the numeracy assessment test results that are similar to the threshold (standard) test results. In other words, accuracy is indicated in terms of similarity of the results.

Generally, accuracy is described by answering to the question of: "What percent of our prediction were correct?" So, if we base the definition on the truthfulness of the reality and the prediction, accuracy is calculated as the ratio of prediction values that are the same as reality values over the total values true or false (Bettenburg, et al., 2008).

For our study, we took Lipkus as a standard for numeracy skill assessment (Reality) and then we compared the results of method C-PNA as a variation for numeracy skill assessment (Prediction) with the results of Lipkus. We calculated the percentile of users who fall in the same numeracy skill level in C-PNA as in Lipkus.

For this purpose, we first obtained the scores of each user in the tests; we used box-plotting technique for categorizing their level of numeracy skill. There are three levels in this categorization: low, medium and high. We compared the results of each user in both tests and find the overall number of the similar results.

2) Effectiveness: Effectiveness is defined in terms of the coverage of categories of numeracy questions. Difficulty Level (DL) is a number assigned to each question in the question bank and it varies depending on the type of the question. It is calculated as the number of DLs covered without explicitly asking related questions to each DL. If all DLs are covered in the test, the test covers all types of questions, all categories of numeracy questions, and it means that the set of questions of the test is effective.

3) Productivity: the number of questions asked in a specified test relative to the time taken to answer them by users. Generally, productivity is the output over input which here is the number of questions answered over time. We say users are more productive using the test if they answer more questions per unit of time.

4) Usage Efficiency: The usage efficiency based on ISO/IEC DIS 25022 of the test is measured as an objective been achieved over a specific time. It is calculated as the average time to cover one DL. Our objective is to cover more DLs meaning obtaining more coverage on different types of questions. Usage Efficiency is the time required to cover one DL, one category of numeracy question types.

Table 4.1 shows the definition of each of the objective characteristics discussed above along with their indicators. Table 4.2 introduces the base measures required for calculation of the objective characteristics with their measurement formulas and the measurement data interpretation as represented in Table 4.3.

Table 4.1 Objective Characteristics

Objective Characteristic	Indicator	Definition
Accuracy	accuracy_ind	The percentile of our test results that is similar to the standard test results. The number shows the percentile of the users who fall in the same category in two different tests.
Effectiveness	effectiveness_ind	Number of DLs covered without explicitly asking related questions to each DL.
Productivity	productivity_ind	Number of questions answered in a specified test relative to the time taken by the user.
Usage Efficiency	usageEfficiency_ind	The usage efficiency of the test is measured as an objective been achieved over a specific time. Our objective is to cover more DLs meaning obtaining more coverage on different types of questions.

Table 4.2 Definitions of Base Measures

Base Measure	Definition
A	Answer to each question for each individual
DL	Number of DLs covered by each test for each individual
Q	Number of questions required to complete a test for each user
TNP	Total Number of Users
TNS	Number of Users in the same Category
T	Time required for the user to complete a test

Table 4.3 Objective Characteristics Measurement Formulas

Indicator	Measurement Formula	Interpretation
accuracy_ind	= (TNP-TNS)/ TNP*100	Results close to 100% are ideal. Higher values indicate more accurate results.
effectiveness_ind	= DL	Results close to 100% are ideal. Higher numbers indicate higher effectiveness.
productivity_ind	= Q / T	Higher numbers show higher productivity.
usageEfficiency_ind	= DL / T	Higher numbers show higher efficiency in terms of usage.

Figure 4.1 depicts the objective characteristics of our hierarchal quality model, which is composed of four characteristics: accuracy, effectiveness, productivity and usage efficiency.

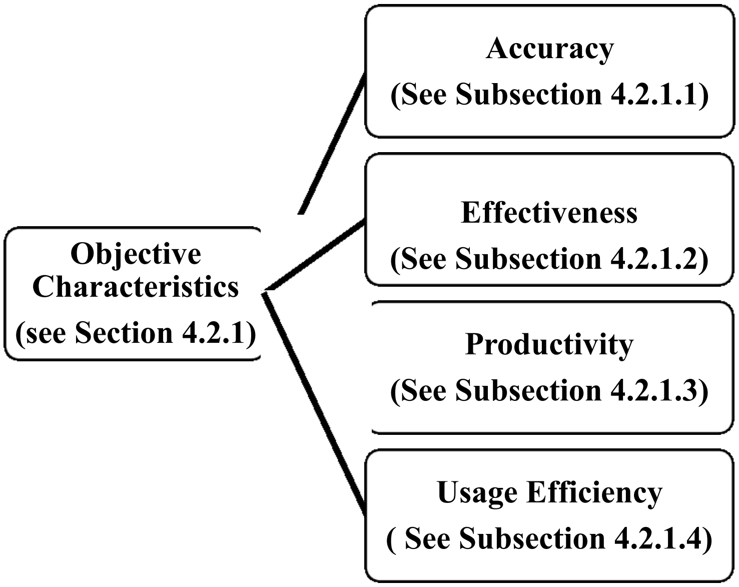


Figure 4.1 Objective Characteristics of Quality Model for Evaluation of Numeracy Assessment System

4.2.2 Subjective Characteristics

Subjective characteristic measurements reflect the viewpoint of whom it is measured by. Basically, the viewpoints of users are obtained from the questions on the questionnaires presented to them after their experience using the system. To collect this qualitative data, users indicate the ratings on an ordinal scale. Consequently, these subjective characteristics' quantification is engaged with human judgment (ISO/IEC DIS, 2016).

Our subjective characteristics include: (1) satisfaction characteristic which in turn concerns mainly on the comfort in answering the questions, the pleasure in writing the test, the understandability of the questions on the test, (2) the discretionary usage between two tests performed in one session, and (3) the trust on the test results.

The subjective characteristics are measured on a Likert scale; the users are asked to rate their reaction to a statement along a scale for a type of survey question from a range of responses often from a positive rating to a negative rating with a neutral score in between. These subjective characteristics are listed as:

4.2.2.1. Satisfaction Measures

Satisfaction measures based on ISO/IEC DIS 25022 assess the degree to which user needs are satisfied when a system is utilized in a specified context of use. The value of satisfaction can be an overall measure of satisfaction produced by combining measures of individual sub-characteristics, which could be in turn weighted according to the importance of them to the overall satisfaction. Users answer each question on the questionnaire by choosing one of the values on a scale ranging from strongly agree to strongly disagree. The sum of all sub-characteristics could be also transformed into a percentage.

Here, we in turn defined the users' level of satisfaction as a result of the pleasure in writing the test, the comfort in answering the questions, and the understandability of the test questions in each session. Table 4.4 shows the definition of satisfaction measures and Table 4.5 summarizes our satisfaction measure for the purpose of our study.

Table 4.4 Satisfaction Measures

Measure	**Description**	**Measurement Function**	**Method**
User Satisfaction	The overall Satisfaction of user	X=S(Xi) Xi sub-characteristics of Satisfaction	Questionnaire

Table 4.5 Satisfaction Indicator

Measure	**Description**	**Measurement Function**	**Method**
satisfaction_ind	The overall satisfaction of the user	X = Pleasure + Comfort + Understandability	Questionnaire

2) Comfort Measures:

Comfort measures based on ISO/IEC DIS 25022 assess the degree to which users' needs for physical comfort are satisfied. Physical comfort can be influenced by position or actions that the user has to make to use the computer system, and by the environment in which the system is used. It is shown as Table 4.6.

Table 4.6 Comfort Measures

Measure	**Description**	**Measurement Function**	**Method**
Physical Comfort	The extent to which the user is comfortable compared to the average for this type of system	X = A A = Psychometric scale value from a comfort questionnaire (See Table 4.11)	Questionnaire

3) Pleasure Measures:

Pleasure measures based on ISO/IEC DIS 25022 assess the degree to which user needs for pleasure are satisfied. The needs of users encompass their desire to obtain new knowledge and skills, to communicate their personal identity, to provoke new pleasant memories and to be involved in the interaction. Table 4.7 shows the definition of pleasure measures.

Table 4.7 Pleasure Measure

Measure	**Description**	**Measurement Function**	**Method**
User Pleasure	The extent to which the user obtains pleasure compared to the average for this type of system	X = A A = Psychometric scale value from a pleasure questionnaire (See Table 4.11)	Questionnaire

4) Understandability Measures:

Understandability measures assess the degree to which user understands the content of the questions on the test as defined in Table 4.8.

Table 4.8 Understandability Measures

Measure	**Description**	**Measurement Function**	**Method**
Understandability	The satisfaction of the user with Understandability of system	X = A A= Response to a question relate to understandability (See Table 4.11)	Questionnaire

4.2.2.2 Trust Measures

Trust measures based on ISO/IEC DIS 25022 assess the degree to which a user has confidence that a product or system will behave as intended. It is shown as Table 4.9.

Table 4.9 Trust Measures

Measure	**Description**	**Measurement function**	**Method**
User Trust	The extent to which the user trusts the system	X = A A = Psychometric scale value from a trust questionnaire (See Table 4.11)	Questionnaire

4.2.2.3 Discretionary Usage

Discretionary Usage on the basis of ISO/IEC DIS 25022 is defined as the proportion of users who prefer one method over the other one as depicted in Table 4.10.

The templates of Table 4.4 to Table 4.10 are inspired by ISO-IEC25022.

Table 4.10 Discretionary Usage Measures

Measure	**Description**	**Measurement Function**	**Method**
Discretionary Usage	The proportion of potential users choosing to use a system	X = A/B A= Number of users using a specific system B = Number of potential users who could have used the specific system	Measure user behavior or automated data collection

Table 4.11 shows the corresponding statements on the questionnaire for each of these subjective characteristics and Figure 4.2 demonstrates the subjective characteristics of our hierarchical quality model which is composed of satisfaction, discretionary usage and trust at one layer and satisfaction, itself, is included of comfort, pleasure and understandability at the next layer.

Table 4.11 Subjective Base Measure Definitions

Base Measure	Definition
Pleasure	The whole test was a pleasant experience to me.
Comfort	I felt comfortable going through the sequence of the questions in the test.
Understandability	It was easy to understand the questions.
Discretionary Usage	Personally, on the result of which method you prefer to have your numeracy skill assessed?
Trust	I trust the result of C-PNA.

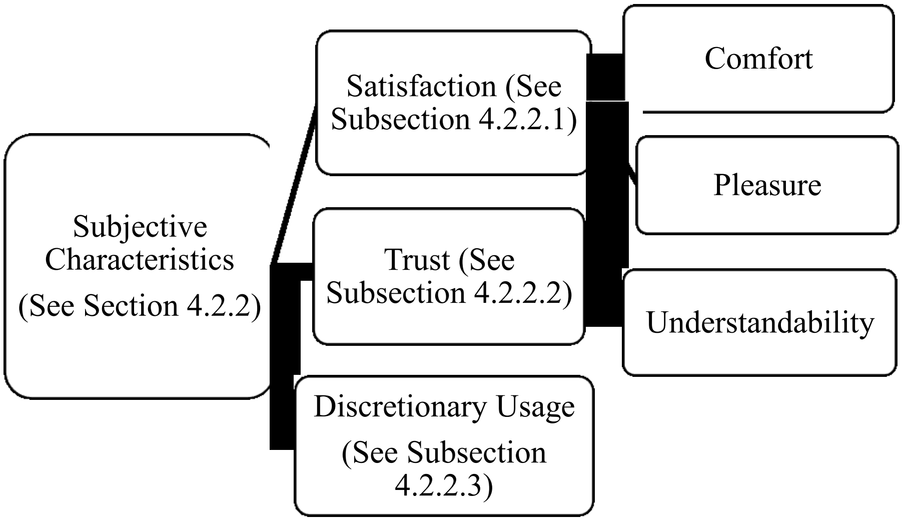


Figure 4.2 Subjective Characteristics of Quality Model for Evaluation of Numeracy Assessment System

In order to evaluate the model empirically, we designed a tool and performed some empirical studies. Step2, tool support and Step3 controlled experiments are discussed in Chapter 5 and Chapter 6 respectively (see Section 4.1 for more details).

4.3 Summary

We designed a quality measurement model for the evaluation of our C-PNA model and then we described the objective and subjective characteristics of the quality measurement. For the purpose of conducting controlled experiments, we developed a web-based/portal application to assess numeracy level of patients which withholds information about the patients and results of the surveys that is discussed in the next chapter.

Chapter 5 Online Platform for Conducting Healthcare Controlled Experiments

For the purpose of evaluating the quality of our C-PNA method, we designed an online web application, which enables us to create, run test sessions, and then save the results of the test sessions for further analysis. The system facilitates the process of designing different test sessions with C-PNA and NC-PNA methods and the process of adjusting the questionnaires based on the type of the tests and the comparison of results of different methods.

Figure 5.1 illustrates the general connections between the views of the online system developed as a proof of concept and used in the empirical studies.

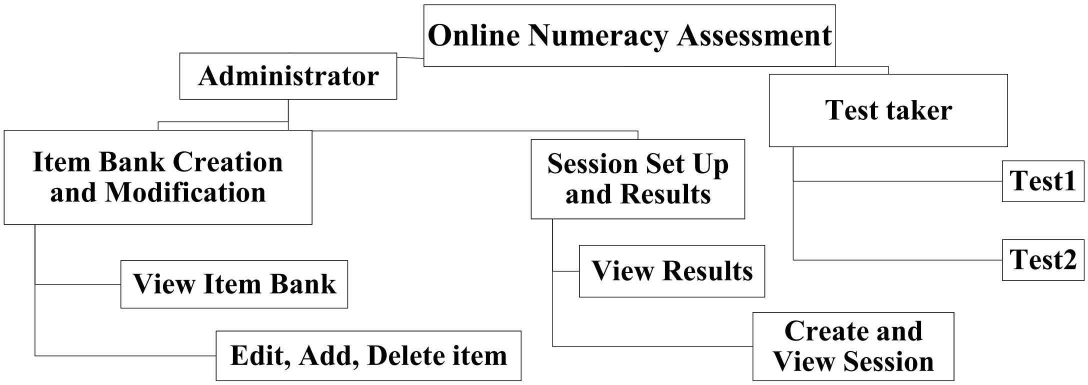


Figure 5.1 Structure of online C-PNA System

In this chapter to address our Objective 3, we provide the architecture of our patient numeracy assessment system. We discuss the significant design decisions underlying the architecture of the application, and explain the features and functionalities of the system.

5.1 C-PNA System Architecture

The architecture of a system is the fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principles guiding its design and evolution (Amery & Rich, 2008).

A software structure consists of the interrelated elements that perform their specific tasks to fulfill various functionalities and describe the whole architecture of the system and how it performs. The software system is differentiated with the tasks it performs and the specific set of domain it caters to. The software architecture provides a comprehensive architectural overview of the software. It presents a number of different architectural views to depict different aspects of the system. It is intended to capture and convey the significant architectural decisions, which have been made on the system to fit the requirements.

We adopted a 4+1 view model to present the architecture of our C-PNA system as accurately as possible, shown in Figure 5.2. The 4+1 view model which is salient for presenting large and complex architecture, emphasizes on the concerns of all the stakeholders. The model represents the system as several concurrent views with different UML representations each one addressing specific set for concerns (Kruchten, 1995).

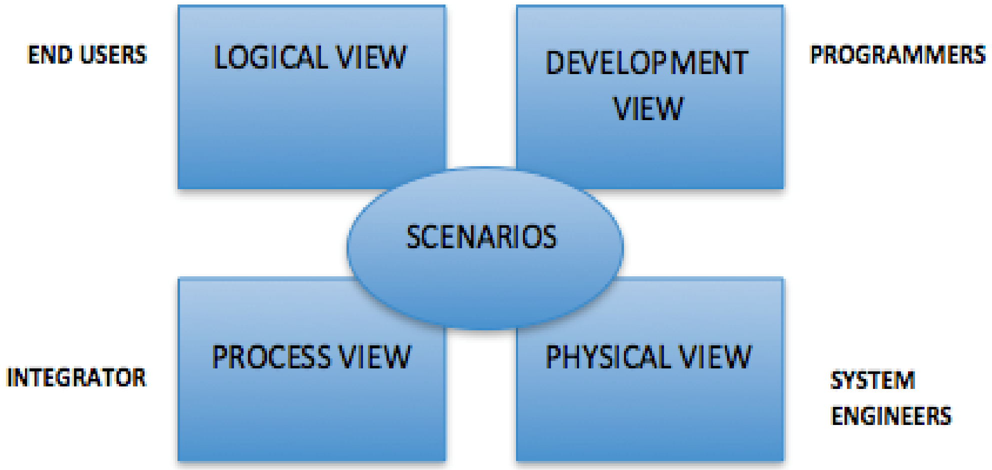


Figure 5.2 The 4+1 View Model (Kruchten, 1995)

5.1.1 Logical View

Logical view focuses on functionality and is responsible for the conceptual organization of layers and high-level functionality of components in each layer.

All end users, patients and test administrators in our system, are considered stakeholders. The representation of this view is by means of class diagram or domain model diagram. Figure 5.3 illustrates the class diagram of our C-PNA system.

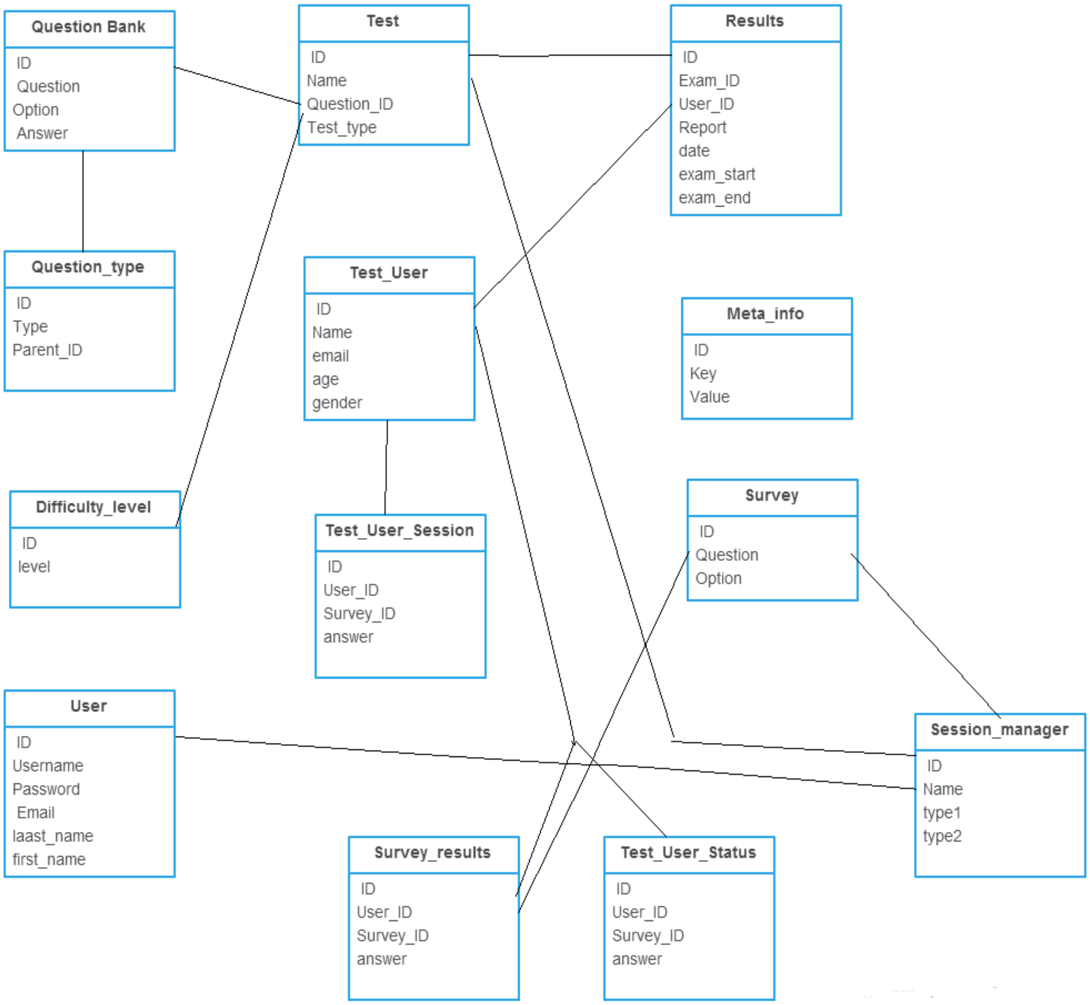


Figure 5.3 C-PNA Class Diagram

5.1.2 Process View

Process view deals with the dynamic aspects of the system and explains the system processes and their communication. Concurrency and synchronization is described on this view. The process view is indeed the runtime and the execution view.

This view is for illustrating non-functional requirements like concurrency, distribution, integrators, performance, and scalability. System Engineers are the stakeholders and activity diagrams are used for the representation of this view. Figure 5.4 and Figure 5.5 depict the activity diagrams for test session and registration of C-PNA system.

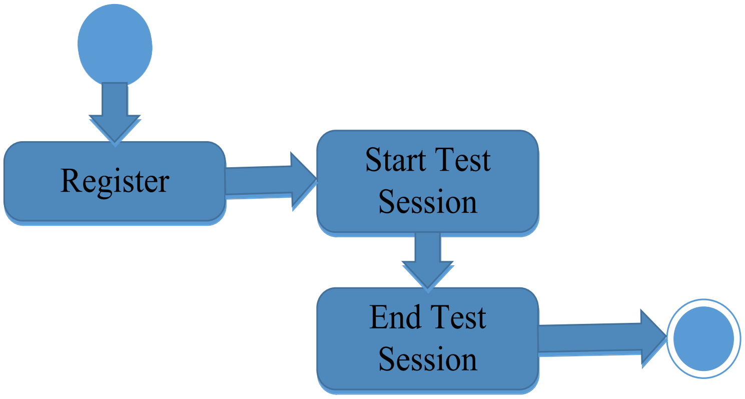


Figure 5.4 C-PNA System Test Session Activity Diagram

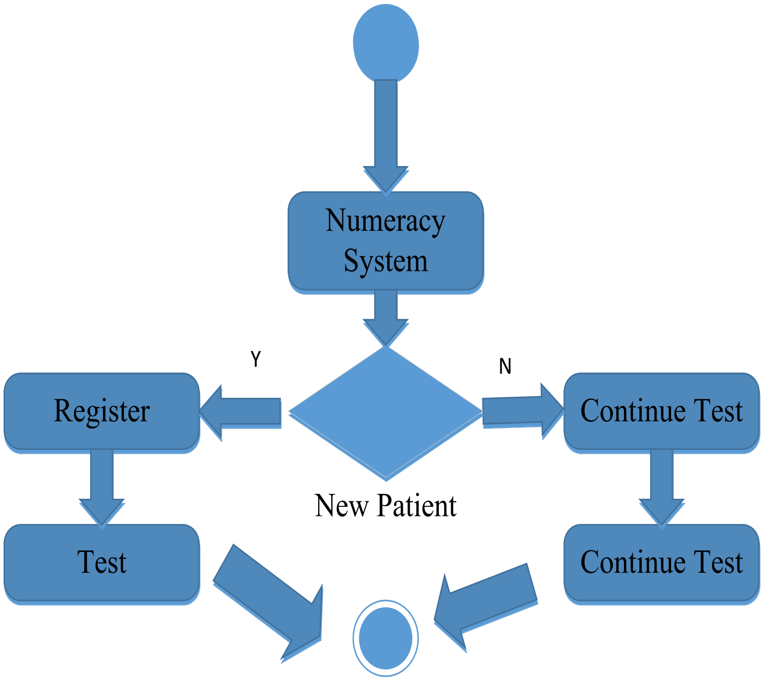


Figure 5.5 C-PNA System Registration Activity Diagram

5.1.3 Implementation View

The implementation view illustrates the system from a programmer's perspective and is concerned with software management. This view focuses on the organization of the actual software modules in the software development environment. The software is packaged in small chunks (program libraries or subsystems) that can be developed by one or more developers. The subsystems are organized in a hierarchy of layers, each layer providing a narrow and well-defined interface to the layer above it. The implementation view is represented by module and system diagrams that show the systems export and import relationships.

This view focuses on the actual source code, data files, and executables and the stakeholders are developers, managers, maintainers and testers. The view is represented as package diagram.

5.1.4 Deployment (Physical View)

Deployment view (physical view) shows what hardware components exist and how the different pieces are connected and what the relationship between them is.

Deployment diagram is created to explore the architecture of the system and Mapping the software to the hardware. The physical architecture takes into account primarily the non-functional requirements of the system such as availability, reliability (fault-tolerance), performance (throughput), and scalability. The software executes on a network of computers, or processing nodes (or just nodes for short). The various elements identified such as networks, processes, tasks, and objects—need to be mapped onto the various nodes. We expect that several different physical configurations will be used: some for development and testing, others for the deployment of the system for various sites or for different customers. The mapping of the software to the nodes therefore needs to be highly flexible and have a minimal impact on the source code itself.

The stakeholders are system engineers and UML Deployment is applied for the representation of this view. The deployment diagram of our C-PNA system is illustrated in Figure 5.6.

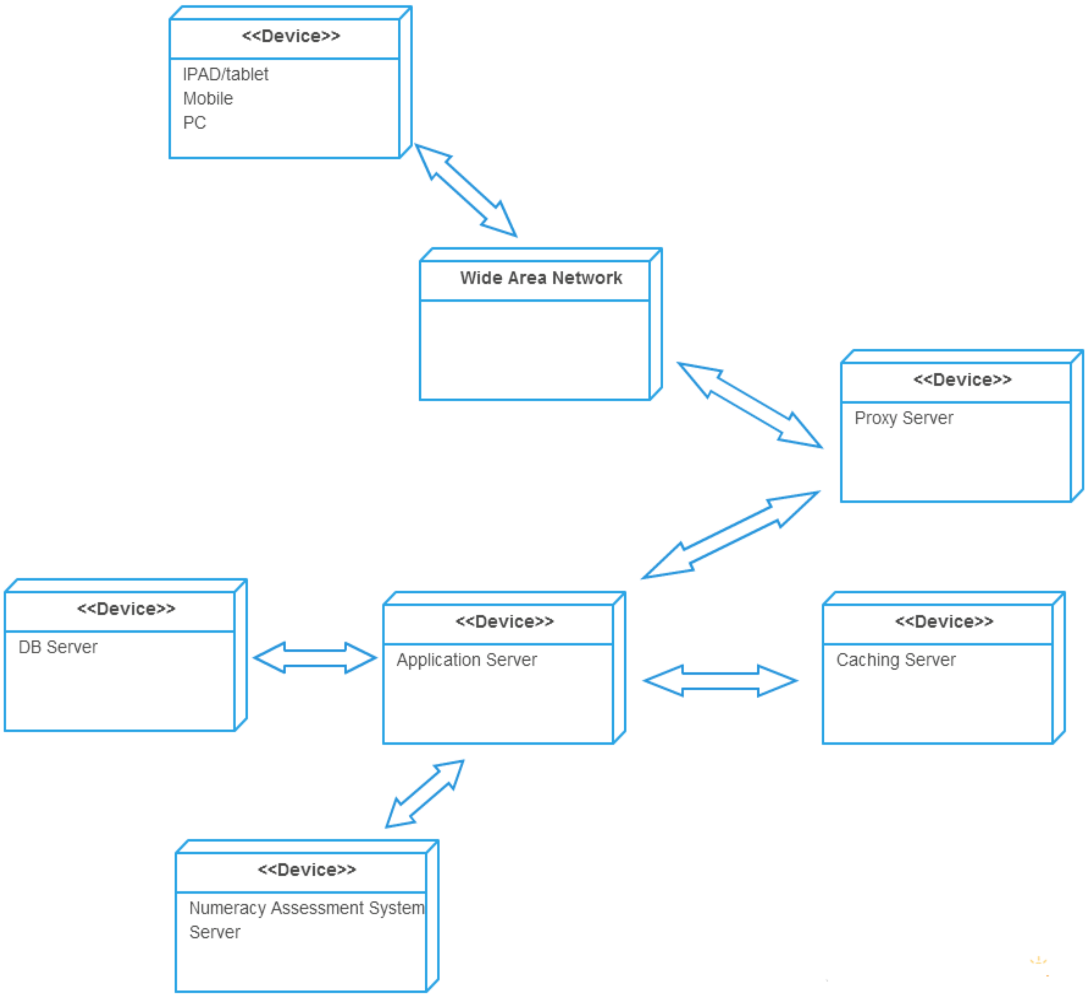


Figure 5.6 C-PNA Deployment Diagram

5.1.5 Use Case View or Scenario View

Use case view captures system functionality for end users. This view is built in early stage of development and represents the 'System Behaviour'.

This view aims at explaining system's intended functions in the form of architecturally important use cases, its users as Actors and relationship between use cases and actors. It also helps requirement engineers in prioritizing requirements. The stakeholders are requirement engineers and testers. Use case diagram is used for the representation of this view. Figure 5.7 shows the use case diagram for our system. In this diagram two main users of the application and their basic functions are shown.

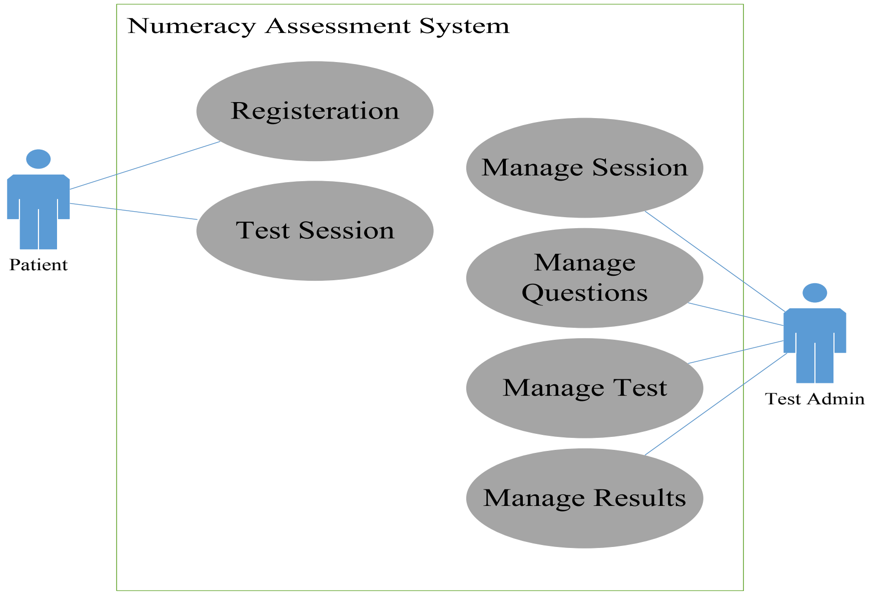


Figure 5.7 C-PNA System Use Case Diagram

5.2 Layers of C-PNA System

In our patient numeracy assessment system, we selected Model–view–controller (MVC) architectural pattern. The application is divided by three interconnected parts, which separate internal representations of information from what is presented to the user.

MVC architectural design pattern is suitable for the architecture of our web application: firstly, because MVC provides a very secure and reliable architecture, specifically for an online assessment system. Secondly, there is the separation of view and logic in MVC, which facilitates the modification of the front-end design without any modification of the logic; we can easily change or upgrade design or view of the system. Finally, MVC supports responsive design, allowing desktop webpages to be viewed in response to the size of the device one is viewing with. Figure 5.8 shows the structure of MVC.

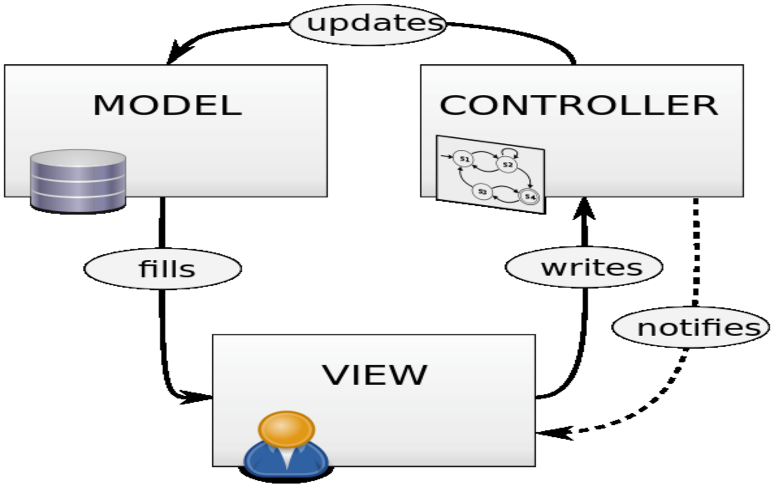


Figure 5.8 Model Controller View Structure (Reenskaug & Coplien, 2009)

Model: It contains the main logical model of our patient numeracy assessment system such as patient profiles, hospital information, history, doctor profiles, login information, database connectivity, database access objects and so on.

View: This layer just works with pure data and does not control the validity of the input data from the user or even control the validity of the data that is showing. It contains views for both mobile or tablet's browsers and android/iOS applications.

Controller: It manages the input of the user, answers them and interacts with the user. The controller layer manages the inquiry statements of the database and sends them to Model, and the Model implements the inquiry.

Implementation: We applied the singleton and observer software design patterns for the implementation of the MVC architectural pattern of our C-PNA system.

5.3 C-PNA Online System

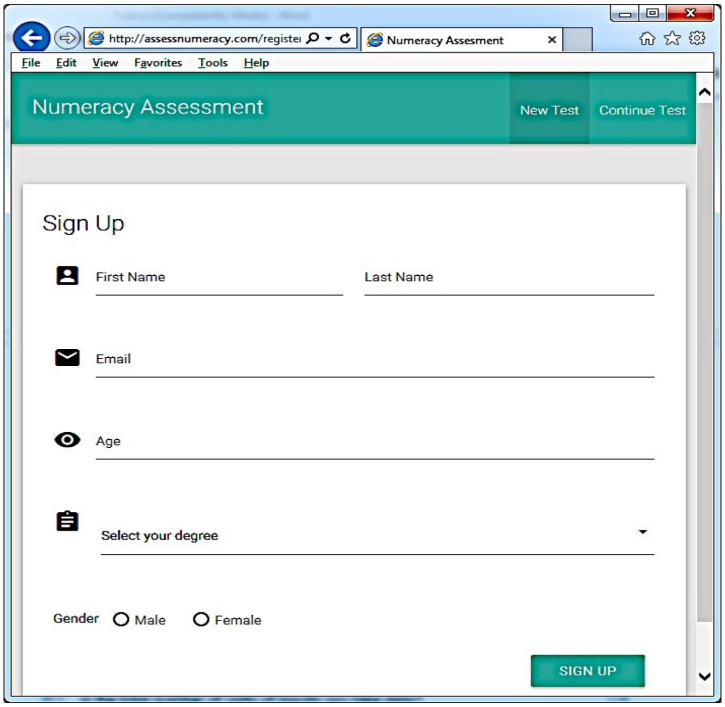
Our web application is built by PHP language and hosted on godaddy's server 'PhpMyAdmin' and the database imported from SQL database.php file.

Our C-PNA system is accessible using the following URL (Omidbakhsh & Ormandjieva, 2015):

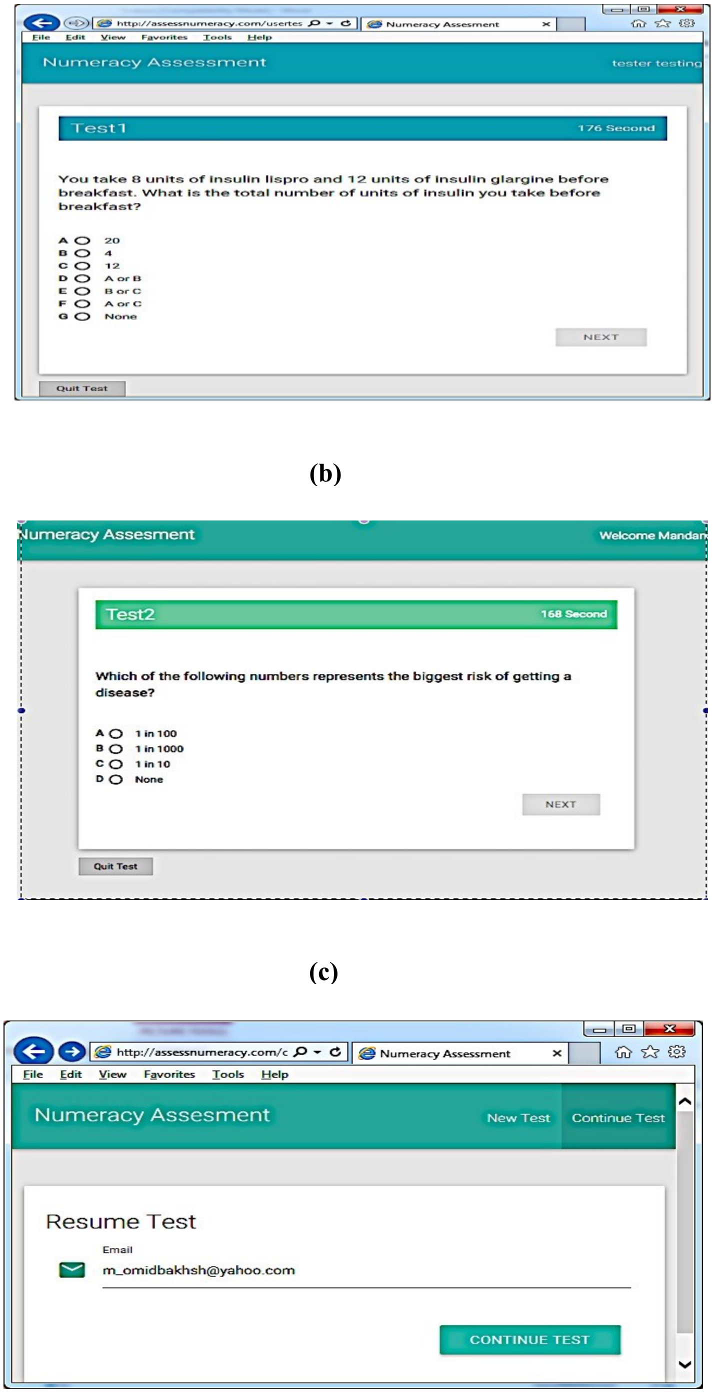
- User Level: http://assessnumeracy.com
- Admin Level: http://assessnumeracy.com/admin

Our system has user and administrator levels.

At the user level, it is possible for patients to create an account to sign in and also to continue the test sessions if already started and signed in. Our system patient interface is illustrated in Figure 5.9 (a, b, c, d).



(a)



You take 8 units of insulin lispro and 12 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?

A ○ 20
B ○ 4
C ○ 12
D ○ A or B
E ○ B or C
F ○ A or C
G ○ None

(b)

Which of the following numbers represents the biggest risk of getting a disease?

A ○ 1 in 100
B ○ 1 in 1000
C ○ 1 in 10
D ○ None

(c)

Resume Test

Email

m_omidbakhsh@yahoo.com

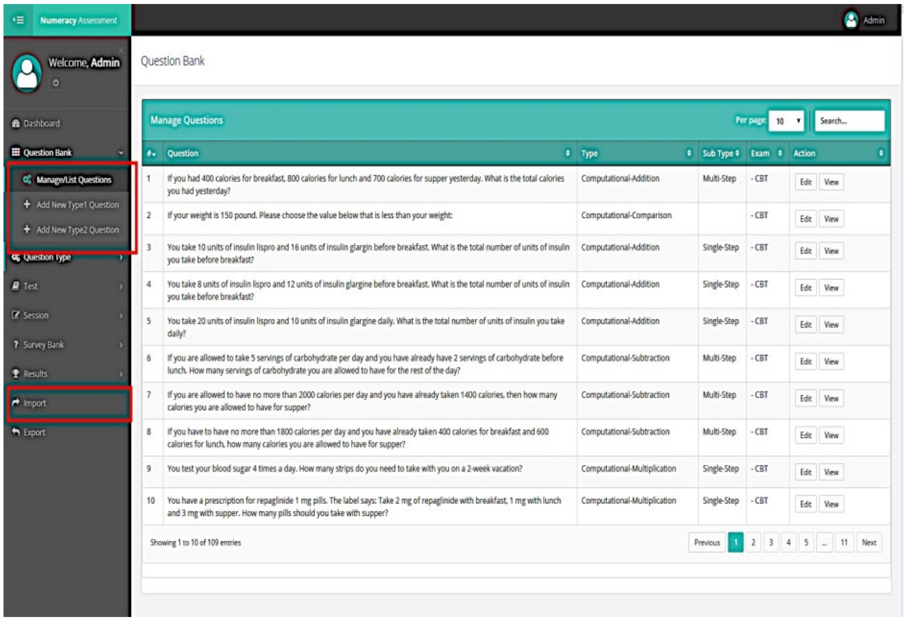
(d)

Figure 5.9 Patient Interface of C-PNA Website: (a), (b), (c), (d)

At the administrator level, the system has the following functionalities. When logging into the admin panel, the options are shown on the menu:

i) Manage/list/add/import questions to question bank

We add questions using ‘Add New Type 1 or 2 Question’ or Import function. If we want to use import function, then the same excel file format should be used. Questions can be updated, viewed or deleted but cannot be deleted if assigned to any tests. Figure 5.10 shows this feature of the system.



Question Bank

Manage Questions

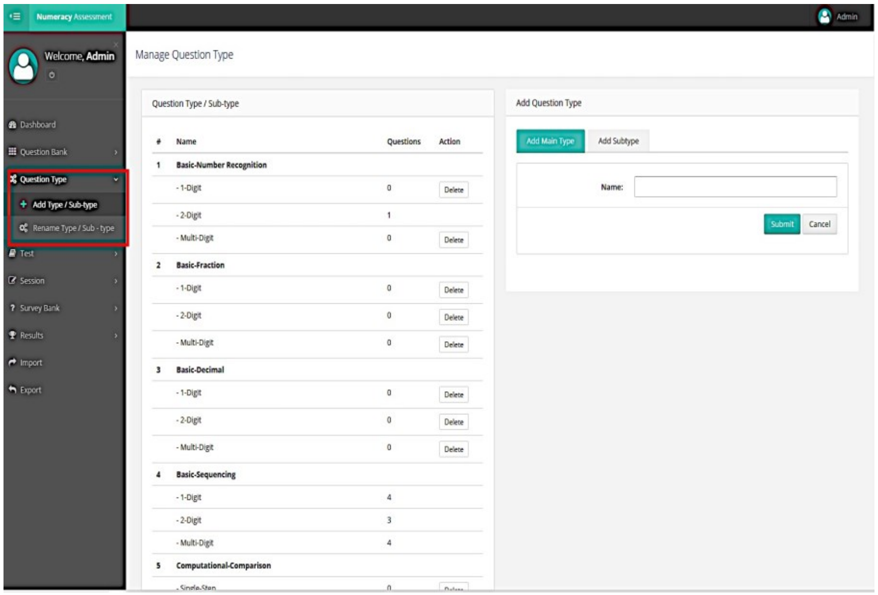
#	Question	Type	Sub Type	Exam	Action
1	If you had 400 calories for breakfast, 800 calories for lunch and 700 calories for supper yesterday. What is the total calories you had yesterday?	Computational-Addition	Multi-Step	- CBT	Edit View
2	If your weight is 150 pound. Please choose the value below that is less than your weight:	Computational-Comparison		- CBT	Edit View
3	You take 10 units of insulin lispro and 16 units of insulin glargin before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step	- CBT	Edit View
4	You take 8 units of insulin lispro and 12 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step	- CBT	Edit View
5	You take 20 units of insulin lispro and 10 units of insulin glargine daily. What is the total number of units of insulin you take daily?	Computational-Addition	Single-Step	- CBT	Edit View
6	If you are allowed to take 5 servings of carbohydrate per day and you have already have 2 servings of carbohydrate before lunch. How many servings of carbohydrate you are allowed to have for the rest of the day?	Computational-Subtraction	Multi-Step	- CBT	Edit View
7	If you are allowed to have no more than 2000 calories per day and you have already taken 1400 calories, then how many calories you are allowed to have for supper?	Computational-Subtraction	Multi-Step	- CBT	Edit View
8	If you have to have no more than 1800 calories per day and you have already taken 400 calories for breakfast and 600 calories for lunch, how many calories you are allowed to have for supper?	Computational-Subtraction	Multi-Step	- CBT	Edit View
9	You test your blood sugar 4 times a day. How many strips do you need to take with you on a 2-week vacation?	Computational-Multiplication	Single-Step	- CBT	Edit View
10	You have a prescription for repaglinide 1 mg pills. The label says: Take 2 mg of repaglinide with breakfast, 1 mg with lunch and 3 mg with supper. How many pills should you take with supper?	Computational-Multiplication	Single-Step	- CBT	Edit View

Showing 1 to 10 of 109 entries

Figure 5.10 Questions of the Question Bank

ii) Add /rename question types to questions in question bank.

We can add the type and subtype of the question to each question and also rename them afterwards as illustrated in Figure 5.11.



Manage Question Type

Question Type / Sub-type

#	Name	Questions	Action
1	**Basic-Number Recognition**		
	- 1-Digit	0	Delete
	- 2-Digit	1	
	- Multi-Digit	0	Delete
2	**Basic-Fraction**		
	- 1-Digit	0	Delete
	- 2-Digit	0	Delete
	- Multi-Digit	0	Delete
3	**Basic-Decimal**		
	- 1-Digit	0	Delete
	- 2-Digit	0	Delete
	- Multi-Digit	0	Delete
4	**Basic-Sequencing**		
	- 1-Digit	4	
	- 2-Digit	3	
	- Multi-Digit	4	
5	**Computational-Comparison**		
	- Single-Step	0	Delete

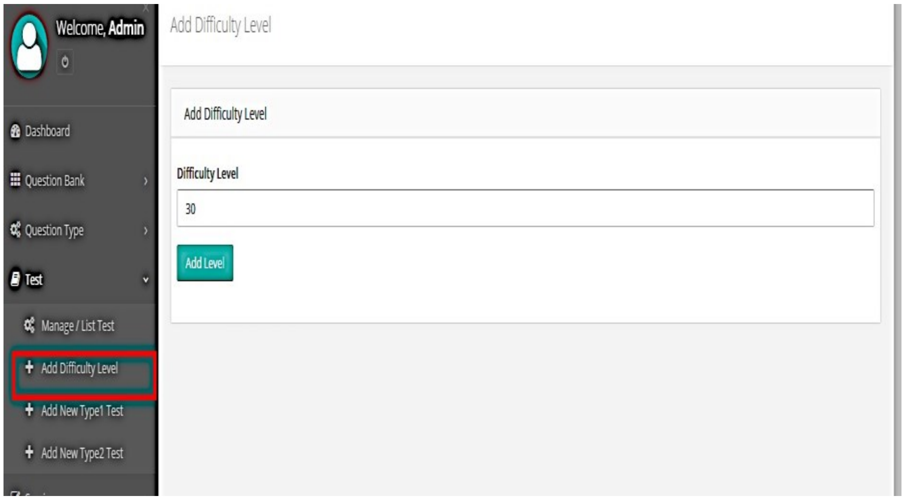
Add Question Type

Name:

Figure 5.11 Types of Questions

iii) Manage/list/add new type of tests

We built two types of tests, namely: Type1Test and Type2Test. Type1Test is for our C-PNA method, which requires the assignment of difficulty levels for each question. We also need to set difficulty level for each question in Type1Test as shown in Figure 5.12.



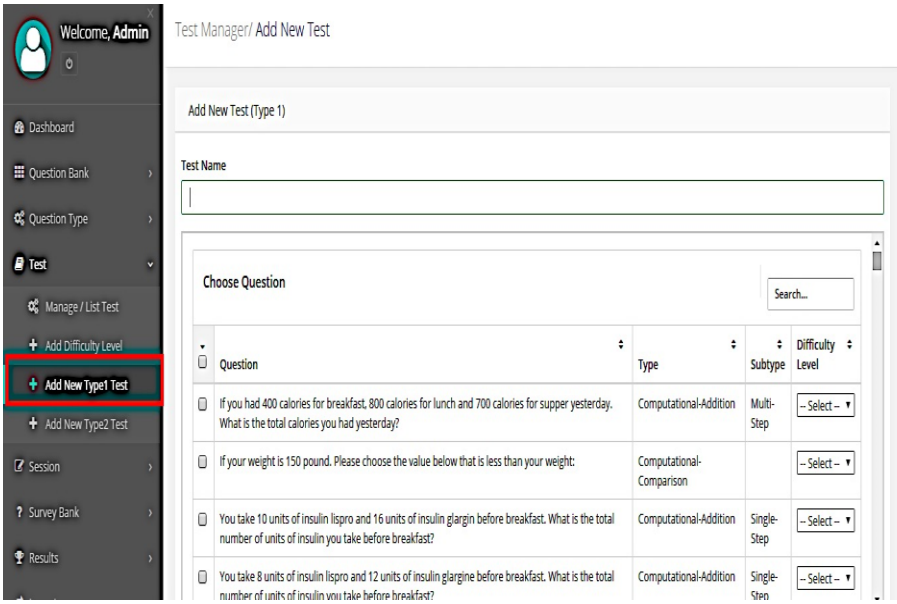
Add Difficulty Level

Add Difficulty Level

Difficulty Level: 30

Figure 5.12 Difficulty Levels

Type2Test is for NC-PNA methods which no difficulty level is needed. Figure 5.13 and Figure 5.14 represent these features of the system.



Test Manager/ Add New Test

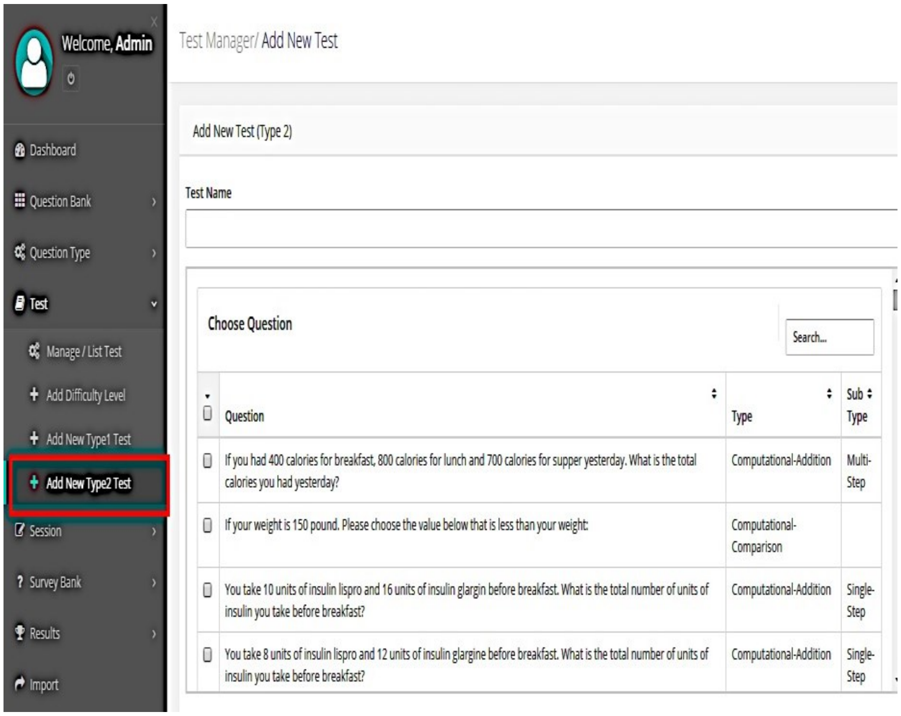
Add New Test (Type 1)

Test Name

Choose Question

□	Question	Type	Subtype	Difficulty Level
□	If you had 400 calories for breakfast, 800 calories for lunch and 700 calories for supper yesterday. What is the total calories you had yesterday?	Computational-Addition	Multi-Step	-- Select --
□	If your weight is 150 pound. Please choose the value below that is less than your weight:	Computational-Comparison		-- Select --
□	You take 10 units of insulin lispro and 16 units of insulin glargin before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step	-- Select --
□	You take 8 units of insulin lispro and 12 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step	-- Select --

Figure 5.13 Test Type 1



Test Manager/ Add New Test

Add New Test (Type 2)

Test Name

Choose Question

□	Question	Type	Sub Type
□	If you had 400 calories for breakfast, 800 calories for lunch and 700 calories for supper yesterday. What is the total calories you had yesterday?	Computational-Addition	Multi-Step
□	If your weight is 150 pound. Please choose the value below that is less than your weight:	Computational-Comparison	
□	You take 10 units of insulin lispro and 16 units of insulin glargin before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step
□	You take 8 units of insulin lispro and 12 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?	Computational-Addition	Single-Step

Figure 5.14 Test Type 2

i) Manage/list/add test sessions

When a test is created, a session can be added by clicking 'Add New Session'. We need a session name and type either type 1 or type 2 of the test for any particular session to be created. The sessions can be updated or deleted at any time. Figure 5.15 presents this feature of the system.

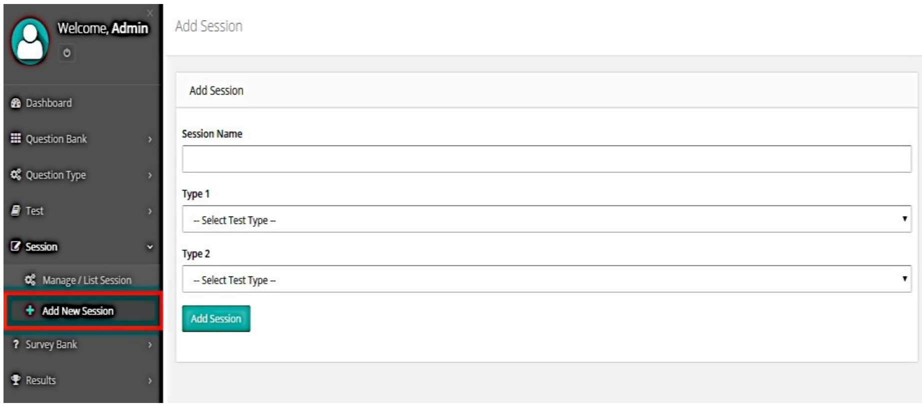


Figure 5.15 Test Sessions

v) Manage/list/add survey questions to survey bank

Similar to adding a question, we add a survey question as shown in Figure 5.16.

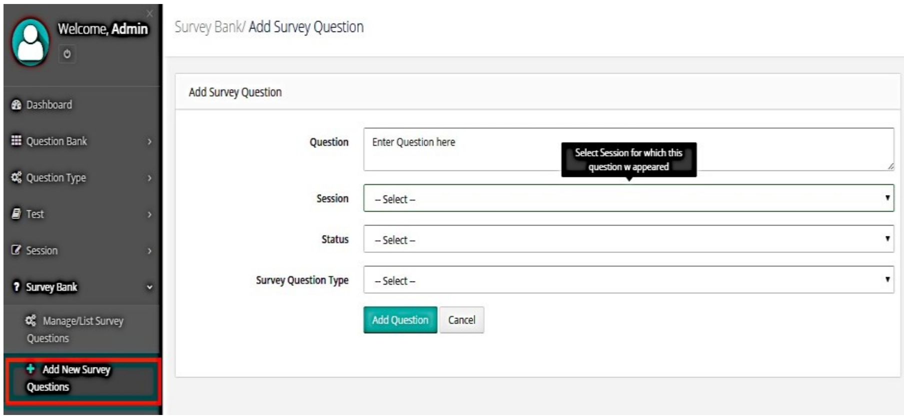
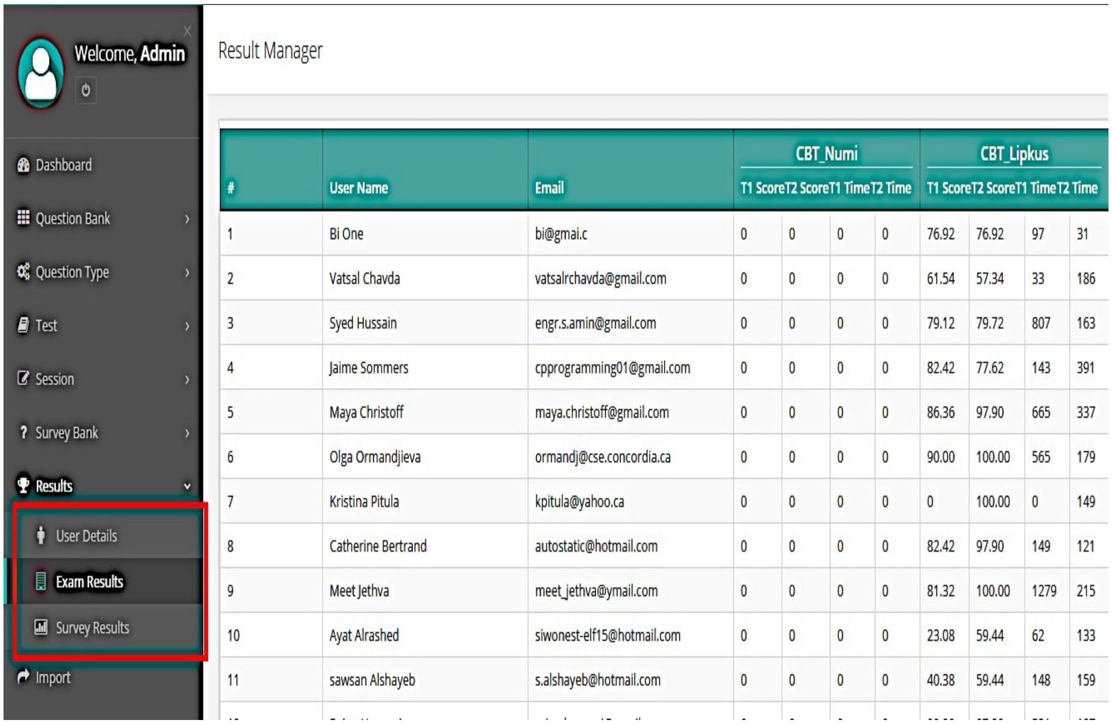


Figure 5.16 Survey Questions

vii) Present result information about patients, sessions and surveys

To view the results of tests taken and compare them, we could either view them or export them as illustrated in Figure 5.17.



Result Manager

			CBT_Numi				CBT_Lipkus			
#	User Name	Email	T1 Score	T2 Score	T1 Time	T2 Time	T1 Score	T2 Score	T1 Time	T2 Time
1	Bi One	bi@gmai.c	0	0	0	0	76.92	76.92	97	31
2	Vatsal Chavda	vatsalrchavda@gmail.com	0	0	0	0	61.54	57.34	33	186
3	Syed Hussain	engr.s.amin@gmail.com	0	0	0	0	79.12	79.72	807	163
4	Jaime Sommers	cpprogramming01@gmail.com	0	0	0	0	82.42	77.62	143	391
5	Maya Christoff	maya.christoff@gmail.com	0	0	0	0	86.36	97.90	665	337
6	Olga Ormandjieva	ormandj@cse.concordia.ca	0	0	0	0	90.00	100.00	565	179
7	Kristina Pitula	kpitula@yahoo.ca	0	0	0	0	0	100.00	0	149
8	Catherine Bertrand	autostatic@hotmail.com	0	0	0	0	82.42	97.90	149	121
9	Meet Jethva	meet_jethva@ymail.com	0	0	0	0	81.32	100.00	1279	215
10	Ayat Alrashed	siwonest-elf15@hotmail.com	0	0	0	0	23.08	59.44	62	133
11	sawsan Alshayeb	s.alshayeb@hotmail.com	0	0	0	0	40.38	59.44	148	159

Figure 5.17 User Results

viii) Export results

As discussed in (i) the questions can be imported to the system as a file. Likewise, the results of tests can be exported from our system.

5.4 Summary

In this chapter, we presented our online platform for conducting controlled experiments in the healthcare. We depicted the architecture of the system and the design patterns applied. By means of this system, we carried out three controlled experiments as discussed in the next chapter.

Chapter 6 Empirical Investigation

To address our Objective 3 (see Section 1.2), to evaluate our C-PNA model, we conducted three empirical studies. We first designed and built an online web application as discussed in the previous chapter and then we designed our test sessions and carried out three controlled experiments.

In order for us to conduct an evaluation of C-PNA model, we need to know what kind of study is appropriate, and what are the key elements involved in designing and conducting empirical studies. Empirical studies are means to test the theory that is developed to explain a phenomenon and predict some consequences (Fenton & Bieman, 2014).

In this chapter, we focus on these empirical studies and highlight the results obtained from them. They do not prove if a theory is true, but they provide further evidence to support or refute the theory. We discuss here, the type of study, study goals and hypotheses, threats to validity, and the use of human subjects in our empirical study.

6.1 Controlled Experiment 1

Our empirical study is a controlled experiment, which investigates alternative ways to perform a specific job. We decide in advance what we want to investigate and how to obtain data for that investigation. There is a high level of control over the variables affecting the result and replication is also possible with low cost.

Our goal is to show that the result of C-PNA is consistent with the results of the existing testing models in terms of the patients' level of numeracy skill. We investigate the consistency of our C-PNA model by conducting the study on the same patients, by comparing the results of the proposed confidence-based model versus non-confidence-based one.

For our goal, we define a null hypothesis and an alternative hypothesis. The null hypothesis is an assumption that the hypothesis is not true unless the evidence is very strong. The null hypothesis defined for our goal is stated as below:

Hyp0: There is no significant difference between the aggregate score of patient numeracy skill using a confidence-based model and the aggregate score of the patient numeracy skill without using the model.

Only if there is strong evidence that the null hypothesis is rejected, we evaluate the alternative hypothesis. Our alternative hypothesis is defined as below:

Hyp1: There is a significant difference between the aggregate score of patient numeracy skill using a confidence-based model and the aggregate score of the patient numeracy skill without using the model.

6.1.1 Experimental Design

In our controlled experiment, we selected at random ten adults from general public in downtown Montreal in winter 2011. The participants in the experiment took the role of patients. The independent variables in the study are: Basic knowledge confidence (MB), Computational knowledge confidence (MC), Analytical knowledge confidence (MA), Statistical knowledge confidence (MS), Level of difficulty (MD) and Number of Questions (MN) of the two methods, one with the confidence-based model and one without the model for the numeracy assessment. The dependent variables are the aggregate scores (MSc) for both methods. To conduct the controlled experiment, there is a list of materials needed for the experiment, which is as the following:

- **Subjective Measures (Profile Information):** General information is gathered by asking the patients to fill out a profile form. This form includes information about patient name, age, gender, and education level of the patients.

- **Question bank:** A bank of numeracy questions is provided. Each question in the bank is representing a type of numeracy question and is assigned a difficulty level to. Our questions in item bank are classified based on the type. For example: if you consider the following questions:

  >You test your blood sugar 4 times a day, each with a separate strip. How many strips do you take with you on a 2-week vacation?

8, 40, 14, 80

It is a computational numeracy question.

>Your target blood sugar is between 60 and 120. Please choose the value below that is in the target range:

89, 142, 13, 56

It is a basic numeracy question.

- **Objective measures:** The base measurements for each patient for the two methods one with the confidence-based model and one without the model for the numeracy assessments are recorded.

6.1.2 Test Environment

We conducted the test for each one of the patients at a time. We asked them to do the assessment once using with the confidence-based model and once without. For this purpose, we used the designed website and we went through the following steps to conduct the experiment: We welcomed and prepared the patients for the experiment. We introduced the process and explained the purpose of the study and their role in the study. We asked the patients to go to the website and complete the profile form. We explained the two assessment methods and we explained that the assessments are in question-and-answer format. We described how to record answers for each method on the system. We informed them that there is no time limit for completing the assessments.

We administered one assessment at a time. We asked the patients to complete the questions and record their answers after each assessment on the system. We also offered assistance for reading and understanding the instructions. We run each test on the website and we thanked the patients for participating in our study and we obtained the results of both assessments from the system.

6.1.3 Results

We tabulated the raw data for the empirical study in MS Excel for each of the patients and we then, calculated the dependent variables for each patient. The scale type of the collected measurement data is ratio, thus arithmetic mean is applied as a measure of central tendency of the data set (Fenton & Bieman, 2014). Mean scores of patients once with applying the confidence based model and once without applying the method are calculated and compared as shown in Table 6.1. We ran a t-test in IBM SPSS Statistics 24 and we obtained the results summarized in Table 6.1 to Table 6.4 below.

Table 6.1 shows the number of patients in each sample (N) and the means of each sample. To check for homogeneity of variances, the performed Leven's test is illustrated in Table 6.2. Here, the value of Sig. is less than 0.05 (significance level), so the assumption of equal variances is verified from Table 6.3 and Table 6.4. Table 6.3 represents a P-value of 0.232, meaning that with the significance level of 0.25 the null hypothesis (Hyp0) can be rejected. In other words, we can conclude that the means of two methods confidence-based and non-confidence based are statistically significant with confidence interval of 78%. Thus, the result of the work supports the alternative hypothesis (Hyp1).

Table 6.1 Results of Application of C-PNA and NC-PNA Tests

Patients Score	**Type of Method**	**N**	**Mean**	**Std. Deviation**
	Non Confidence-based	10	48.10	20.229
	Confidence-based	10	59.40	20.609

Table 6.2 Independent Sample Tests

Patients Score		Levene's Test for Equality of Variances		t-test for Equality of Means
		F	Sig.	t
	Equal variances assumed	.037	.850	-1.237
	not assumed			-1.237

Table 6.3 Independent Samples Test

Patients Score	t-test for Equality of Means		
	dif.		
	Sig.	Mean	Difference
Equal variances assumed	18	.232	-11.300
not assumed	17.994	.232	-11.300

Table 6.4 Independent Sample Tests

Patients Score	t-test for Equality of Means		
	Std. Error		
	Difference	Lower	Upper
Equal variances assumed	9.132	-30.485	7.885
not assumed	9.132	-30.486	7.886

6.1.4 Conclusion

The first controlled experiment revealed that the scores of patients in the two assessment methods are correlated and the means of two assessments are not statistically different. To achieve results with higher confidence interval, we increased the sample size in our next studies.

Furthermore, on the basis of the outcome analysis of Controlled Experiment 1, we defined more specific hypotheses, related to objective and subjective characteristics. We designed and ran Controlled Experiment 2 to test the validity of the hypotheses and then analyzed the results.

6.2 Controlled Experiment 2

In order to evaluate our patient numeracy assessment method, we performed an empirical investigation and conducted a controlled experiment. We adapted the process model described in (Fenton & Bieman, 2014) for this investigation. There are six phases in the model for conducting a controlled experiment, namely: conception, design, preparation, execution, analysis and dissemination.

In the conception phase, the objectives of the study are described. The objective of our controlled experiment was to determine, how differences in the numeracy skill assessment methods could affect the result of the assessment.

In the design phase, we defined the hypotheses for our study. A null hypothesis assumes that there is no significant difference between two methods: C-PNA and NC-PNA with respect to the dependent variables we are measuring. An alternative hypothesis posits that there is a significant difference between the two methods. Hypothesis definition is followed by the generation of a formal design to test the hypothesis.

Then, we prepared and organized the experimental tests in the preparation phase. In the execution phase, we ran the test sessions and collected the measurement data. We analyzed the data collected during previous phase in analysis phase. And finally, in the dissemination phase, we verified the study and determined the modification needed for the improvement of the study.

6.2.1 Hypotheses

We defined the goals of our controlled experiment as sets of hypotheses. We formulated and empirically investigated sets of hypotheses: null and alternative for objective characteristics (Accuracy, Effectiveness, Productivity, and Usage Efficiency) and for subjective characteristics (Pleasure, Comfort, Understandability, Satisfaction, Discretionary Usage and Trust) (see Chapter 4). Below we list them all in terms of null and alternative hypotheses. The first set of hypotheses for the Accuracy is defined as:

HypA0: The results obtained from Patient Numeracy Assessment tool with the confidence-based method are not as accurate as those from Patient Numeracy Assessment tool with the non-confidence-based model with the threshold of 70%.

HypA1: The results obtained from Patient Numeracy Assessment tool with the confidence-based method are as accurate as those from Patient Numeracy Assessment tool with the non-confidence-based model with the threshold of 70%.

The hypotheses for Effectiveness are defined as follows:

HypE0: There is no significant difference between Effectiveness of Patient Numeracy Assessment Tool using a confidence-based method and Effectiveness of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

HypE1: There is a significant difference between Effectiveness of Patient Numeracy Assessment Tool using a confidence-based method and Effectiveness of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

Furthermore, the hypotheses for Productivity are defined as:

HypPR0: There is no significant difference between Productivity of Patient Numeracy Assessment Tool using a confidence-based method and Productivity of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

HypPR1: There is a significant difference between Productivity of Patient Numeracy Assessment Tool using a confidence-based method and Productivity of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

And the hypotheses for the Usage Efficiency are defined as the following:

HypUE0: There is no significant difference between Usage Efficiency of Patient Numeracy Assessment Tool using a confidence-based method and Usage Efficiency of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

HypUE1: There is a significant difference between Usage Efficiency of Patient Numeracy Assessment Tool using a confidence-based method and Usage Efficiency of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

The hypotheses for Pleasure are defined as:

HypPL0: There is no significant difference between the patients' Pleasure using a confidence-based Patient Numeracy Assessment method and the patients' Pleasure of the same Patient Numeracy Assessment Tool without using confidence-based method.

HypPL1: There is a significant difference between the patients' Pleasure using a confidence-based Patient Numeracy Assessment method and the patients' Pleasure of the same Patient Numeracy Assessment Tool without using confidence-based method.

And the hypotheses for the Comfort are defined as the following:

HypCom0: There is no significant difference between the patients' Comfort using a confidence-based Patient Numeracy Assessment method and the patients' Comfort of the same Patient Numeracy Assessment Tool without using confidence-based method.

HypCom1: There is a significant difference between the patients' Comfort using a confidence-based Patient Numeracy Assessment method and the patients' Comfort of the same Patient Numeracy Assessment Tool without using confidence-based method.

The hypotheses for Understandability are described as:

HypU0: There is no significant difference between the patients' Understandability using a confidence-based Patient Numeracy Assessment method and the patients' Understandability of the same Patient Numeracy Assessment Tool without using confidence-based method.

HypU1: There is a significant difference between the patients' Understandability using a confidence-based Patient Numeracy Assessment method and the patients' Understandability of the same Patient Numeracy Assessment Tool without using confidence-based method.

Also, the hypotheses for Satisfaction are formulated as below.

HypS0: There is no significant difference between the patients' overall Satisfaction level using a confidence-based Patient Numeracy Assessment method and the patients' overall Satisfaction level of the same Patient Numeracy Assessment Tool without using confidence-based method.

HypS1: There is a significant difference between the patients' overall Satisfaction understanding level using a confidence-based Patient Numeracy Assessment method and the patient's overall Satisfaction level of the same Patient Numeracy Assessment Tool without using confidence-based method.

And the hypotheses for Trust are defined as:

HypT0: There is no Trust in patients on confidence-based Patient Numeracy Assessment System for the evaluation of their numeracy skill level.

HypT1: There is Trust in patients on confidence-based Patient Numeracy Assessment System for the evaluation of their numeracy skill level.

Finally, the hypotheses for Discretionary Usage are as follows:

HypDU0: There is no significant difference between the patients' Discretionary Usage using a confidence-based Patient Numeracy Assessment method and the patients' Discretionary Usage of the same Patient Numeracy Assessment Tool without using confidence-based method.

HypDU1: There is a significant difference between the patients' Discretionary Usage using a confidence-based Patient Numeracy Assessment method and the patients' Discretionary Usage of the same Patient Numeracy Assessment Tool without using confidence-based method.

Before executing the tests, instructions on the organization of the experimental tests, running tests and collecting the measurement data are set up (preparation level).

At the execution level, we applied the *assessments* (C-PNA, NC-PNA) to the *experimental object* (Website) by the *experimental subjects* (Patients). We recorded the time, date, location and type of patients for the data collection.

For the analysis, we, first, reviewed all the measurements taken to make sure that they are valid and useful. We eliminated the outliers of the data set form our data gathered through empirical studies. Then, we organized the measurements into sets of data that are examined as part of the hypothesis-testing process. We analyzed the sets of data according to the statistical principles. These statistical tests proved if the hypotheses are supported or refuted by the results of the experiment.

At the end of the analysis phase, we reached a conclusion about how the different characteristics we examined affected the outcome.

6.2.2 Experimental Design

In Controlled Experiment 2, we selected at random 24 adults from general public in downtown Montreal in the summer 2015 who took the role of patients. We offered them gift cards to reward them for their participation in the study.

6.2.3 Results

We investigated our patient numeracy assessment system based on the study characteristics defined in Chapter 4. We intended to compare our method with two existing ones: NUMi and Lipkus. To this aim, we designed two different test sessions. First session included NUMi Test and the second one included Lipkus. The same patients went through a pair of numeracy tests in each test session. We collected our data through our website: assessnumeracy.com. The results are discussed in terms of objective and subjective characteristics as below:

6.2.3.1 Objective Characteristics

The objective characteristics of our quality model provided measures relating to and Accuracy, Effectiveness, Productivity, and Usage Efficiency of the patients in the two sessions. The raw data for the empirical study for all patients were saved in our database system. We calculated the value of the characteristics for each patient separately. Then we used the mean of all characteristics, for all patients, to compare each characteristic for C-PNA and NUMi, in Session1 and C-PNA and Lipkus, in Session2.

We presented the data obtained from our experiment in separate graphs for each session. Figure 6.1.1 depicts the results for the objective characteristics: Effectiveness, Productivity and Usage Efficiency of C-PNA and NUMi tests for all the patients in Session1. Figure 6.1.2 shows the results for the objective characteristics: Effectiveness, Productivity and Usage Efficiency of C-PNA and Lipkus tests for all the patients in Session2.

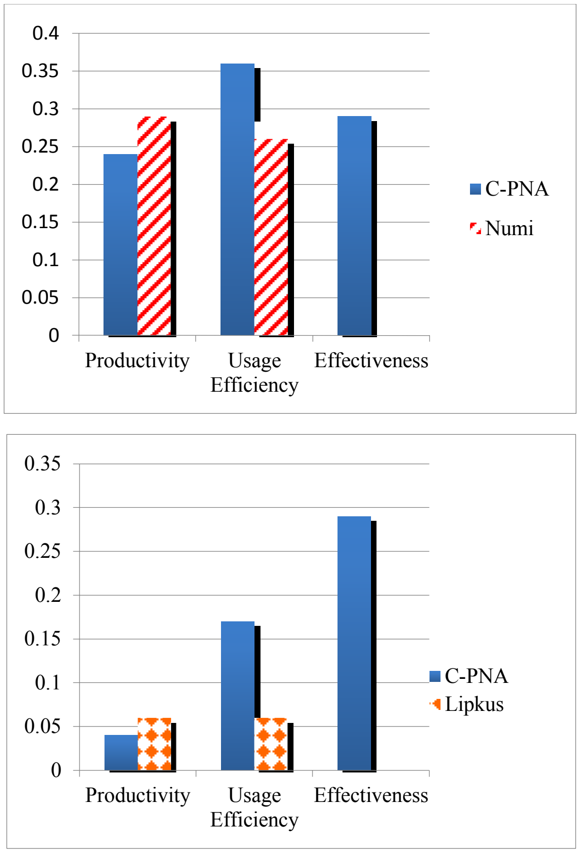


Figure 6.1 Objective Characteristics (1) Session1 (2) Session2

The Effectiveness is higher for C-PNA in both sessions. We concluded that HypE0 (Effectiveness) is refuted and HypE1 (Effectiveness) is supported by the result of the study. However, the value of Productivity for both sessions is quite comparable, so HypPR0 (Productivity) is supported and HypPR1 (Productivity) is refuted. Furthermore, the value for Usage Efficiency of C-PNA is higher for both sessions, resulting in supporting HypU1 (Usage Efficiency) and refuting HypU0 (Usage Efficiency). As Figure 6.2 represents the result of C-PNA is 80% accurate to Lipkus.

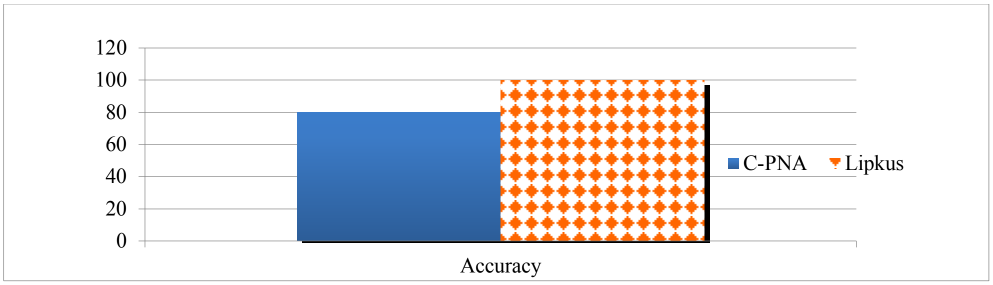


Figure 6.2 Accuracy Characteristic Session2

For statistical validation Table 6.5 and Table 6.6 show the t-Test values and the P-values for all the objective characteristics respectively for Session1and Session2 of this controlled experiment. The hypotheses are verified for each characteristic based on the t-value and the P-values. Our decision rule is to reject the null hypothesis if the computed P-value is not greater than 0.05, meaning that there is no statically significant difference between the two methods.

Table 6.5 Paired t-Test for the Objective Characteristics of Session1: C-PNA and Numi

Objective Characteristics	**P-value**	**t-value**
Accuracy	0.0	8.115
Productivity	0.226	-1.281
Effectiveness	0.001	4.387
Usage Efficiency	0.03	3.819

Table 6.6 Paired t-Test for the Objective Characteristics of Session2: C-PNA and Lipkus

Objective Characteristics	**P-value**	**t-value**
Accuracy	0.126	-1.627
Productivity	0.061	-2.034
Effectiveness	0.0	5.933
Usage Efficiency	0.008	1.836

6.2.3.2 Subjective Characteristics

We used the values: strongly agree, agree, neutral, disagree, and strongly agree for each of the subjective characteristics on the Likert scale. We assigned weights 5 to 1 for each value. Then, we calculated the median (for the ordinal scale) of each sub-characteristic of Satisfaction for all patients in each session.

In Session1 and Session2, as shown in Figure 6.3 the median for Pleasure of C-PNA is higher than both NUMi and Lipkus, however it is the same for Comfort and Understandability. The value of median of Pleasure is 4 and 3 for C-PNA and NUMi for Session1 and is 3.5 and 3 for C-PNA and Lipkus for Session2, respectively. It shows that patients found C-PNA test as a more pleasant experience than NC-PNA test.

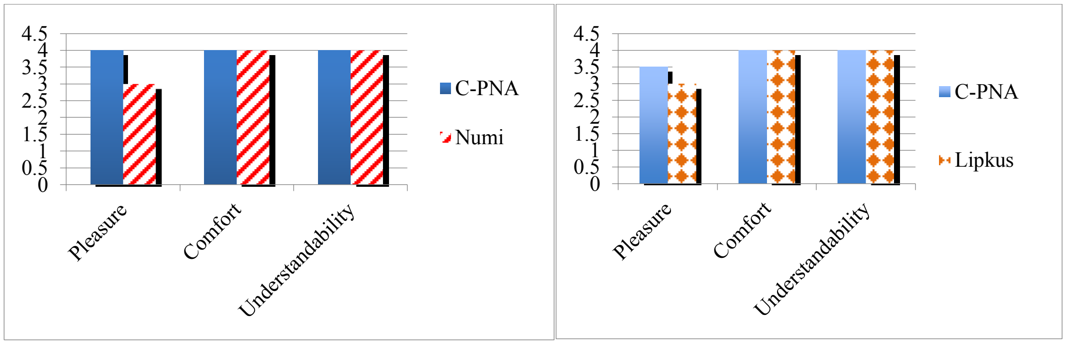


Figure 6.3 Subjective Characteristics (1) Session1 (2) Session2

We witness a significance difference between the value of Pleasure from C-PNA and NC-PNA. Therefore, HypPL0 (Pleasure) is refuted and HypPL1 (Pleasure) is supported by this study. We think the reason behind is that people feel more challenged when they write adaptive test with gradual rise on the difficulty level of the questions on the test. The median values for Comfort and Understandability are 4 for all tests in both sessions, showing that there is no difference between the tests in this respect. So, HypCom0 (Comfort) and HypU0 (Understandability) are accepted and HypCom1 (Comfort) and HypU1 (Understandability) are refuted.

Furthermore, our system calculates the percentile of patients who chose strongly agree, agree, neutral, disagree and strongly disagree for each sub-characteristic. Figure 6.4 and Figure 6.5 show the percentiles according to the choices taken by all patients and specifically demonstrate that patients chose strongly agree and agree for Pleasure 75% in C-PNA in comparison with 83% in NUMi, for comfort 66% in C-PNA over 75% in NUMi, but for Understandability 75% in C-PNA over 66% in NUMi in Session1. Moreover, for Pleasure 53% in C-PNA in comparison with 46% in Lipkus, for comfort 60% in C-PNA over 66% in Lipkus and, for Understandability 66% in C-PNA over 60% in Lipkus in Session2.

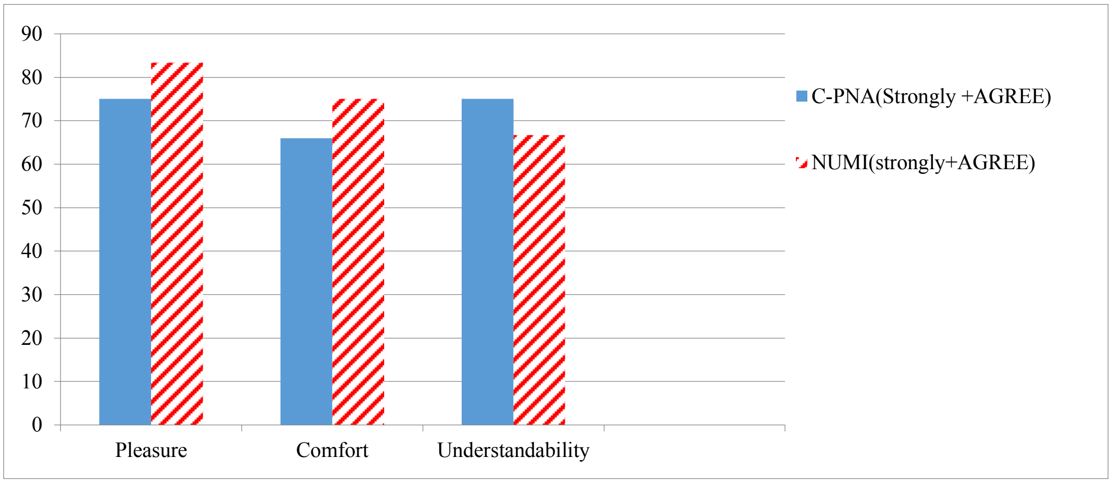


Figure 6.4 Percentile of Patients Who Chose Strongly Agree and Agree for Satisfaction- Session1

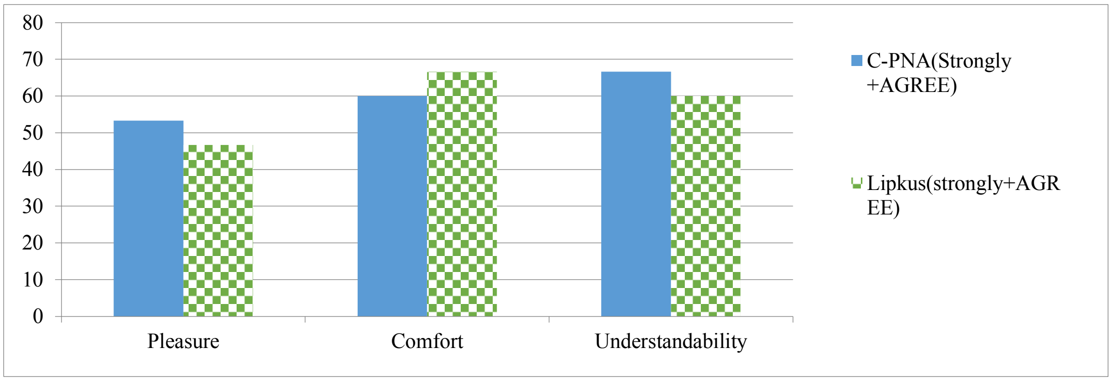


Figure 6.5 Percentile of Patients who Chose Strongly Agree and Agree for Satisfaction-Session2

1. Satisfaction

We also calculated the overall Satisfaction based on the formula defined in Chapter 4. We summed the median of each sub-characteristics in formula for getting the value of satisfaction_ind (see Chapter 4)

The graphs in Figure 6.6.1 and Figure 6.6.2 demonstrate that the median value of Satisfaction is 12 and 11 for C-PNA and NUMi for Session1 and 11 for both C-PNA and Lipkus in Session2, respectively. This allows us to refute HypS1 (Satisfaction) for which there is not much a significant difference between C-PNA and NC-PNA.

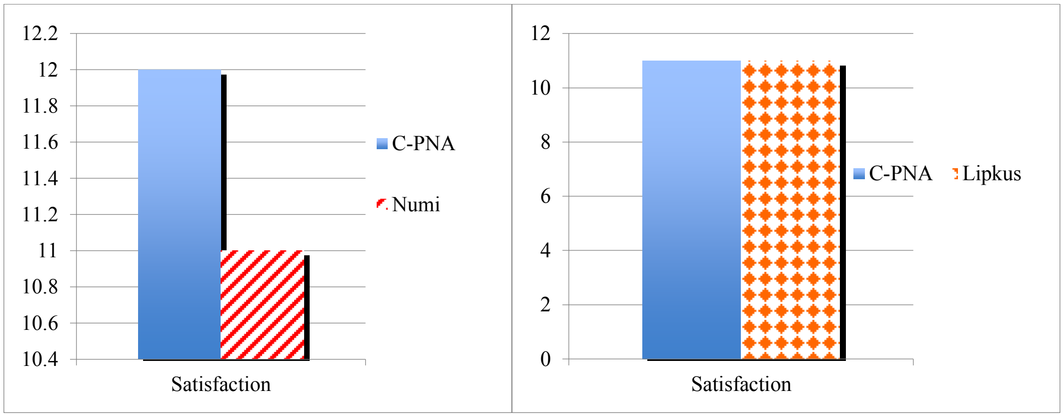


Figure 6.6 Satisfaction (1) Session1 (2) Session2

2. Discretionary Usage

For Discretionary Usage, as the graphs in Figure 6.7 indicate, 75% of patients preferred C-PNA over NUMi in Session1 and 60% of patients preferred C-PNA over Lipkus in Session2. This is also a support for HypDU1 (Discretionary Usage), as there is a significant difference between the values. We believe this is also due to the nature of our adaptive testing that gains more Discretionary Usage in comparison to NC-PNA.

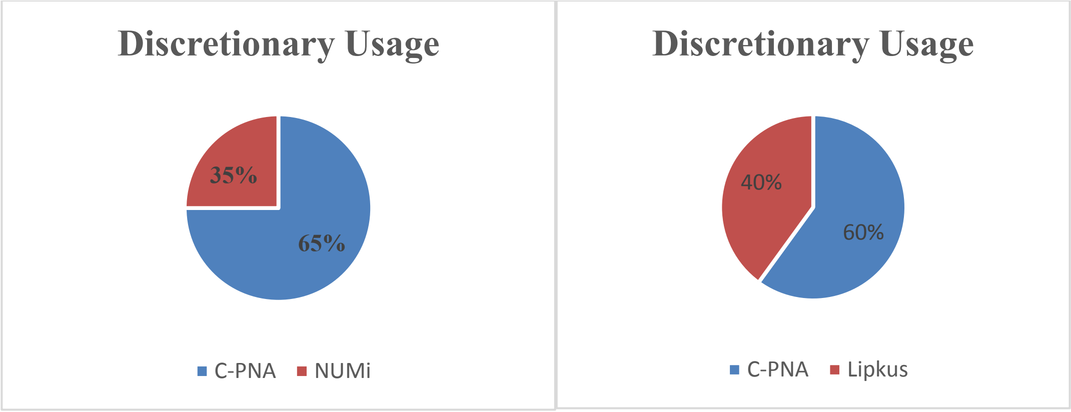


Figure 6.7 Discretionary Usage (1) Session1 (2) Session2

We mapped the values obtained for Discretionary usage in percentile to Likert scale values. The mapping was performed based on Table 6.7. The value of our Discretionary usage of C-PNA and NUMi are 4 and 2 respectively in Session1 and that of C-PNA and Lipkus are 4 and 3 respectively in Session2.

Table 6.7 Mapping of Likert Scale to Percentile

strongly disagree	**disagree**	**neutral**	**agree**	**strongly agree**
1	2	3	4	5
0-19%	20%-39%	40%-59%	60-59%	80-100%

3. Trust

For the subjective characteristic Trust, we added the following statement to our online questionnaire: "I trust the result of C-PNA." Figure 6.8 depicts the percentage of patients who chose strongly agree, agree, neutral, disagree and strongly agree for all patients.

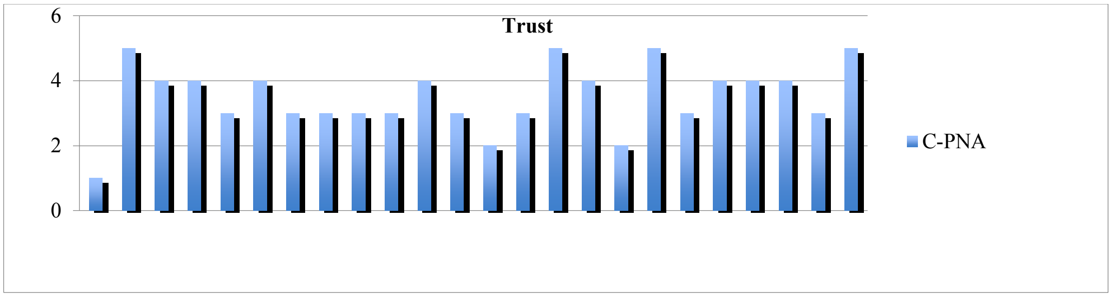


Figure 6.8 Total Trust on C-PNA

Moreover, we calculated the median of value of Trust and depicted in Figure 6.9. Hence, HypT1 (Trust) could be supported by the results obtained from this study and HypT0 (Trust) is refuted by the study data.

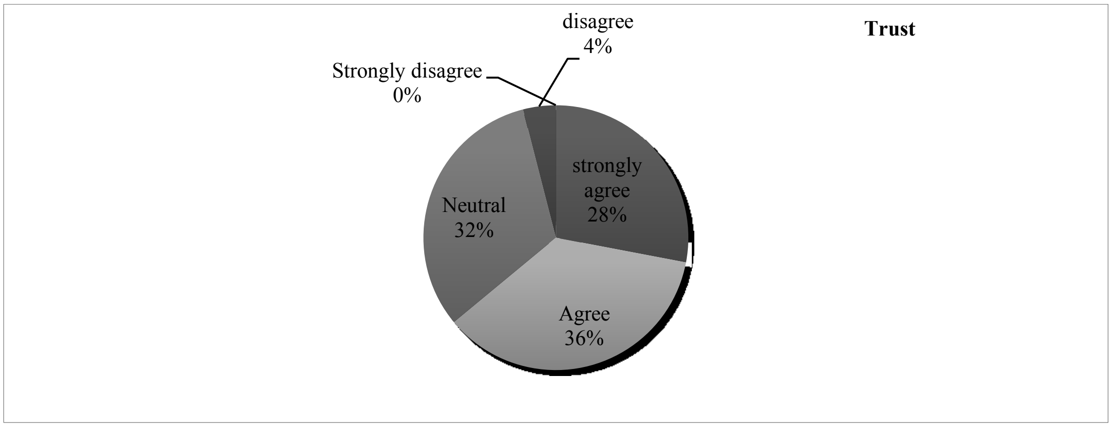


Figure 6.9 Distribution of Trust on C-PNA

To include Trust characteristic in our comparison, we assigned the values {-2, -1, 0, 1, 2} to the Likert scale values {“Strongly disagree”, “agree”, “neutral”, “agree”, “Strongly agree”} accordingly. As the total value of Satisfaction is 11, we added the value of Trust characteristic to it, and we got the value of 12 for C-PNA as Satisfaction and Trust together. Figure 6.10 shows these values and the comparison of them.

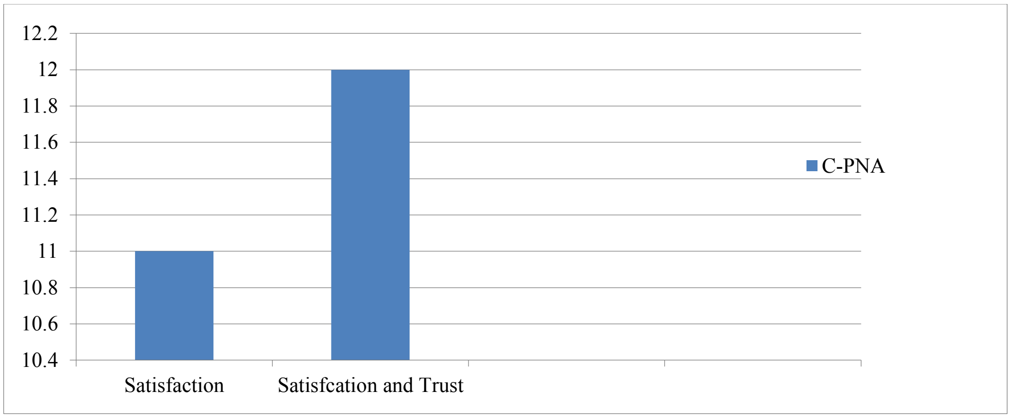


Figure 6.10 Satisfaction and Trust on C-PNA

We summarized the results of the hypothesis testing of Controlled Experiment 2 as shown in Table 6.8.

Table 6.8 Results of Hypothesis Testing of Controlled Experiment 2

Characteristics	**Hyp1**
Accuracy	Refuted
Effectiveness	Supported
Productivity	Refuted
Usage Efficiency	Supported
Pleasure	Supported
Comfort	Refuted
Understandability	Refuted

6.2.4 Conclusion

Based on the results of Controlled Experiment 2, we determined that to attain our objective, there is a need to gather more trustable meaningful measurement data, which is feasible by 1) increasing the number of participants in the study 2) adding another security level by user email verification when they sign in for the study, and 3) adjusting the subjective characteristic Trust on the questionnaire.

We made some revisions and added the following modifications to our system to be able to run Controlled Experiment 3:

i) Allowing patients authentication by submitting a confirmation email after registering in the system

ii) Updating some questions on the item bank

iii) Adding a survey question about subjective characteristic: Trust

iv) Verifying the termination rules on the selection algorithm

6.3 Controlled Experiment 3

The purpose of this experiment is to improve the assessment of the hypotheses validation in Controlled Experiment 2 after the revisions listed in Section 6.2.4. In this experiment, the number of participants is augmented and we grouped them based on their age group and gender as illustrated in Figure 6.11.

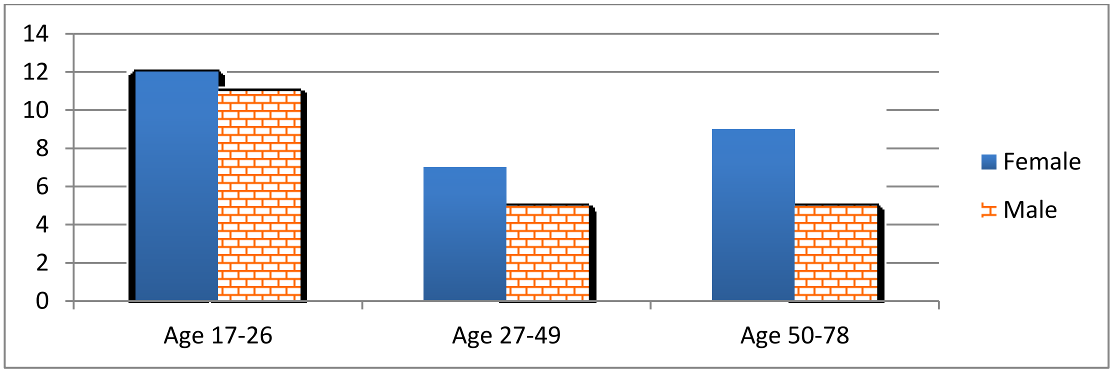


Figure 6.11 Patients Grouped by Age and Gender

6.3.1 Hypotheses

Here, we aim to investigate the consistency of the results obtained from Controlled Experiment 2 by providing more meaningful empirical data. Thus, the goal of the experiment is to distinguish the differences between results attained from C-PNA and NC-PNA models. The hypotheses of this experiment remain as the previous experiment (see Section 6.2.1), with the addition of the following hypotheses:

HypTrust0: There is no significant difference in patients' Trust in Patient Numeracy Assessment Tool using a confidence-based method and Usage Efficiency of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

HypTrust1: There is a significant difference between patients' Trust in Patient Numeracy Assessment Tool using a confidence-based method and Usage Efficiency of the same Patient Numeracy Assessment Tool using the non-confidence-based method.

6.3.2 Experimental Design

In Controlled Experiment 3, after setting up the email validation, we asked general public in the downtown area of Montreal to participate in the study and we were able to collect the data for 60 participants, in the spring 2016.

6.3.3 Results of the Study

There are two types of characteristics: subjective and objective that we measured in this study. We first focused on the objective characteristics and then on subjective characteristics.

6.3.2.1 Objective Characteristic

Accuracy, Productivity, Usage Efficiency and Effectiveness are the characteristics we measure as objective characteristics. Figure 6.12 shows the value of Accuracy, Figure 6.13, the value of Productivity, Figure 6.14, the value of Usage Efficiency and Figure 6.15, the value of Effectiveness for C-PNA and NC-PNA tests for different age groups. Figure 6.12 represents that the total result of C-PNA is 70% accurate to NC-PNA.

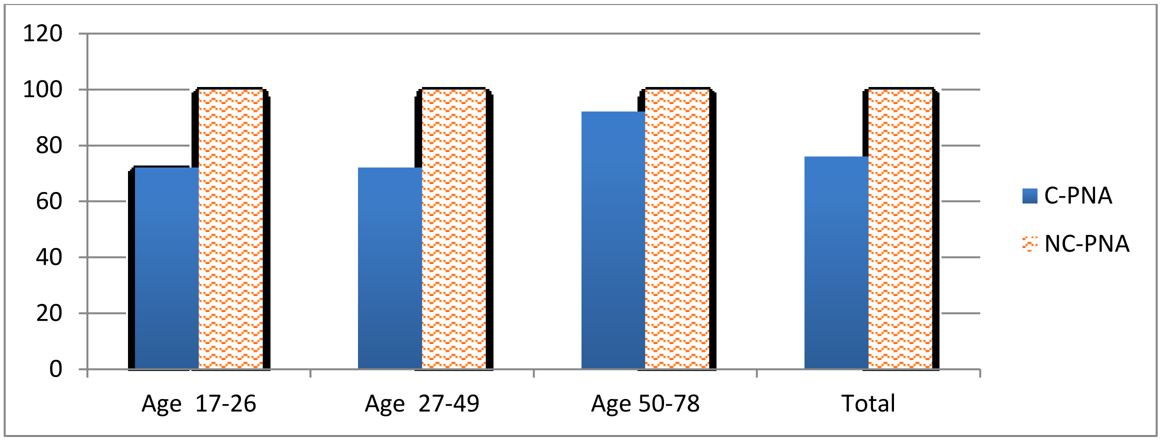


Figure 6.12 Accuracy in Different Age Groups

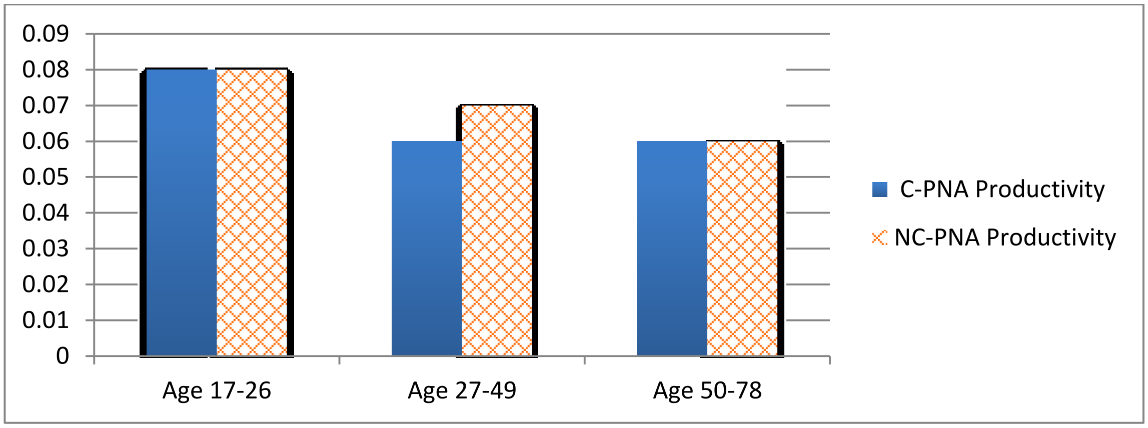


Figure 6.13 Productivity in Different Age Groups

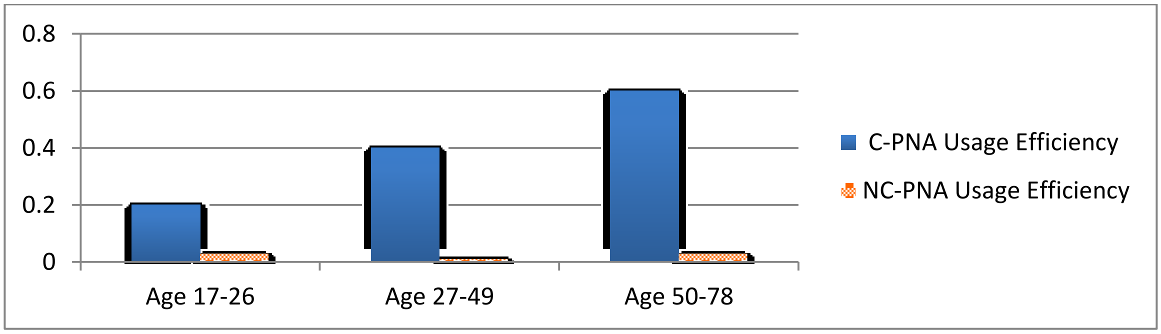


Figure 6.14 Usage Efficiency in Different Age Groups

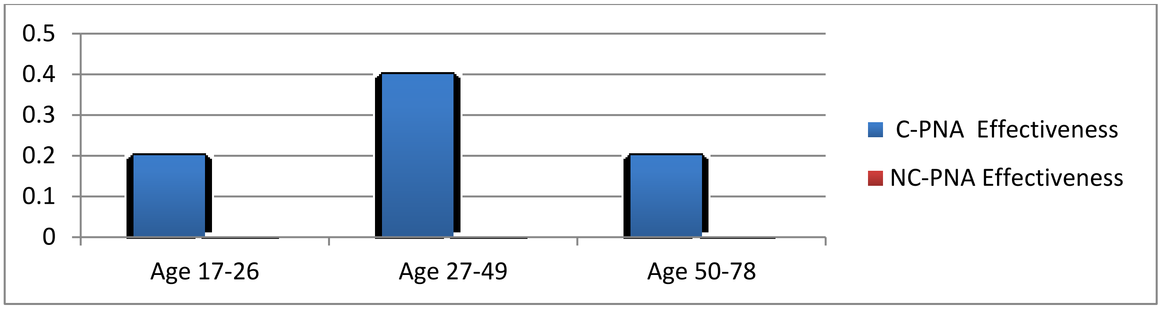


Figure 6.15 Effectiveness in Different Age Groups

The results obtained for this experiment is quite similar to the previous experiment with the exception of Productivity. The value of Effectiveness and Usage Efficiency are higher in C-PNA than NC-PNA for all age groups. We conclude that HypE0 (Effectiveness) and HypU0 (Usage Efficiency) are refuted and HypE1 (Effectiveness) and HypU1 (Usage Efficiency) are supported by the result of the study.

The value of Productivity for age group 27-49 is lower in C-PNA than NC-PNA, which is similar to Controlled Experiment 2, though this value is the same in the other age groups. So HypPR0 (Productivity) is refuted and HypPR1 (Productivity) is supported for the age group 27-49.

6.3.2.2 Subjective Characteristic

Satisfaction, Discretionary Usage and Trust are the characteristics, we measure as subjective characteristics. Moreover, Pleasure, Comfort and Understandability remain as sub-characteristics for Satisfaction characteristic.

1. Satisfaction

We calculated the median of the sub-characteristic of Satisfaction: Pleasure, Comfort, and Understandability for all patients as shown in Figures 6.16 to 6.18.

Figure 6.16 and Figure 6.17 illustrate that the majority (64%) of all the participants in C-PNA agreed or strongly agreed that C-PNA methods is pleasurable and comfortable compared to (50%) as in NC-PNA.

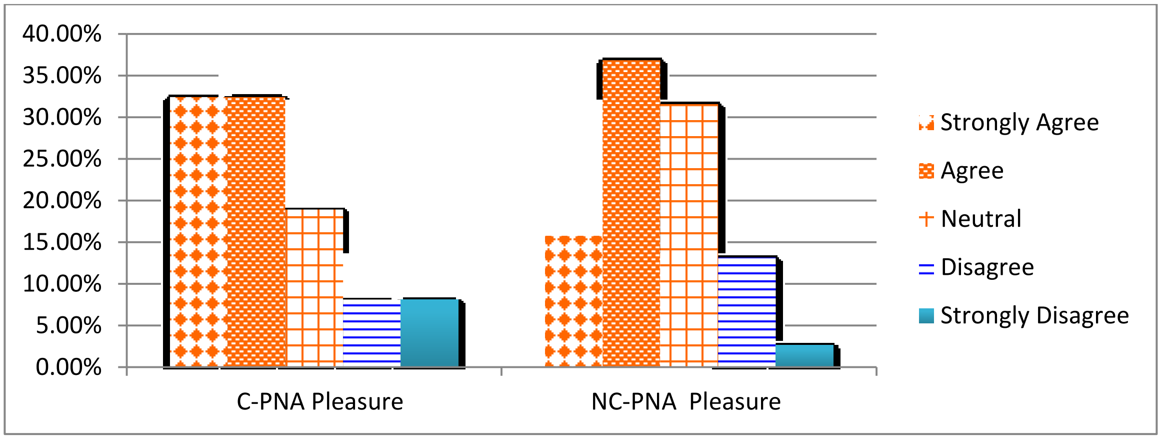


Figure 6.16 Pleasure of C-PNA in Comaprison to NC-PNA

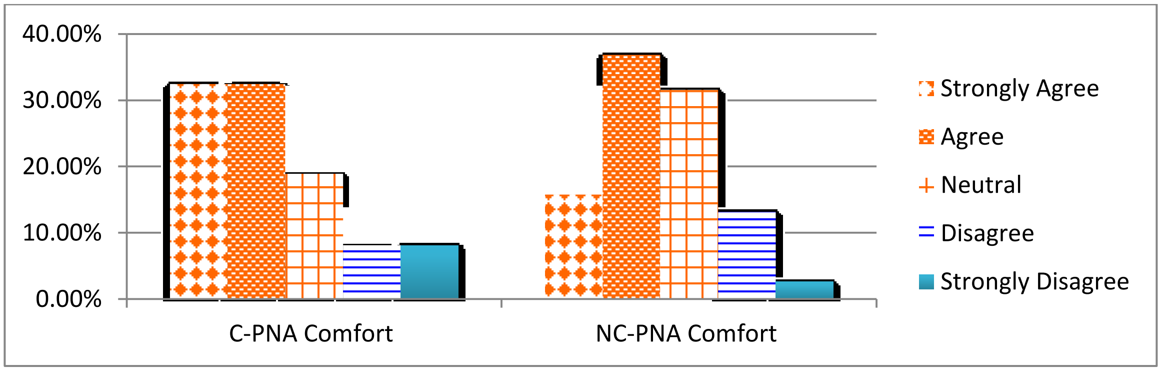


Figure 6.17 Comfort of C-PNA in Comaprison to NC-PNA

Likewise, in Figure 6.18 the majority (62%) of all the participants in C-PNA agreed or strongly agreed that C-PNA test is Understandable compared to (46%) as in NC-PNA.

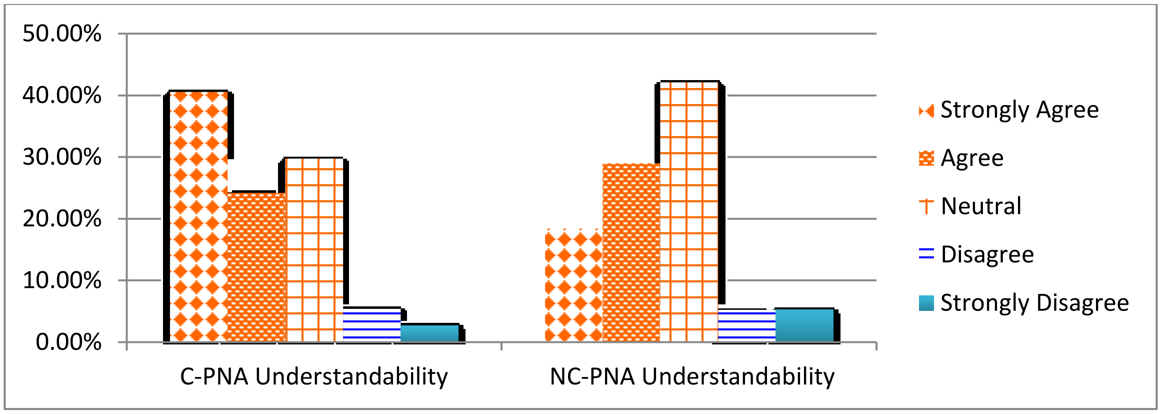


Figure 6.18 Understandability of C-PNA in Comaprison to NC-PNA

2. Discretionary Usage

The results demonstrate that 65% of the patients preferred C-PNA test to NC-PNA test as depicted in Figure 6.19. Therefore, the value of our Discretionary usage of C-PNA and NC-PNA are 4 and 2 respectively based on Table 6.5. This supports HypDU1 (Discretionary Usage) and refutes HypDU0 (Discretionary Usage), in other words, there is a significant difference between the patients' Discretionary Usage using a confidence-based Patient Numeracy Assessment method and the patients' Discretionary Usage of the same Patient Numeracy Assessment Tool without using confidence-based method.

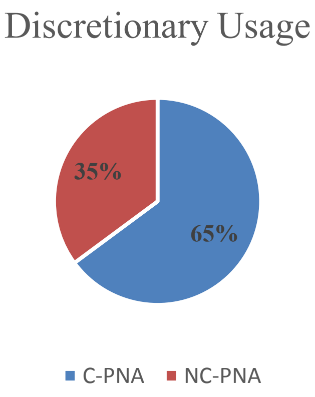


Figure 6.19 Discretionary Usage of C-PNA in Comaprison to NC-PNA

3. Trust

Figure 6.20 shows that the majority (59%) of all the participants in C-PNA agreed or strongly agreed that C-PNA is a trustable test compared to (38%) as in NC-PNA. Therefore, we ascertain that HypTrust0 is refuted and HypTrust1 is supported by the results of the study.

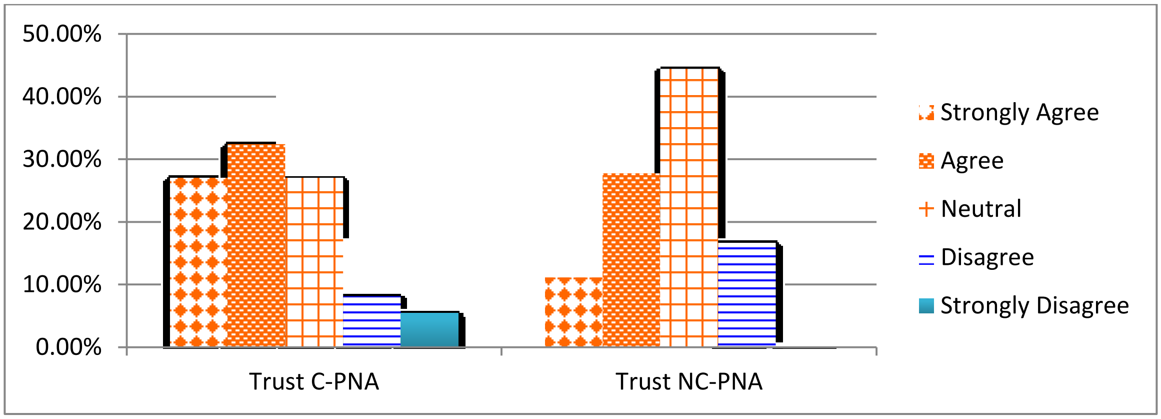


Figure 6.20 Trust of C-PNA in Comaprison to NC-PNA

We summarized the result of the hypothesis testing of Controlled Experiment 3 as shown in Table 6.9.

Table 6.9 Results of Hypothesis Testing of Controlled Experiment 3

Characteristics	**Hyp1**
Accuracy	Refuted
Effectiveness	Supported
Productivity	Refuted/supported age group 27-49
Usage Efficiency	Supported
Pleasure	Supported
Comfort	Supported
Understandability	Supported

6.4 Discussion

This empirical investigation validated the stated hypotheses and clearly demonstrated the method's usefulness in healthcare, providing details of the empirical study conducted with general public. The results of the formal empirical study revealed that our confidence-based numeracy assessment method attained better results as compared to the non-confidence assessment method. As for Effectiveness and Usage Efficiency characteristics, the values are higher in C-PNA than NC-PNA. For Satisfaction characteristics, there is 14% difference in the value of Pleasure and Comfort, and 16% higher in the value of Understandability in favor of C-PNA versus NC-PNA. Furthermore, the results show that the value of Trust is 21% higher in C-PNA than NC-PNA and the value of Discretionary Usage is 35% higher in C-PNA than NC-PNA.

Like in any empirical study, we considered different *threats to validity* on the basis of Wohlin et al. (Wohlin, et al., 2000) as follows:

- *Conclusion validity:*
  In the controlled experiments, we had two primary purposes: to confirm a theory by applying the paired Student t-test to our data, and to explore the relationships among datasets using correlation analysis to confirm whether or not there is a relationship between two attributes. By using this technique, we are generating measures of association that indicate the closeness of the behavior between the two variables. Thus, the conclusions of this research study are founded on an adequate analysis of the data.
- *Internal validity:*
  We minimized the chance of confounding since we drew the conclusions founded on direct manipulation of the independent variables.
- *External validity:* We can generalize the results of the controlled experiments to other numeracy assessment methods.

The results of the formal empirical studies demonstrated that our confidence-based numeracy assessment method excels the non-confidence assessment method in terms of objective and subjective characteristics.

6.5 Summary

In this chapter, we described three empirical studies. We defined our goals in terms of hypotheses and tested them by means of our online system. The result obtained from the empirical studies validate our C-PNA model.

In the next chapter, we discuss how the information attained from our C-PNA system is applicable for a patient preference elicitation system. We present the architecture of such a system and show how to develop the personalization of healthcare information in this regard.

Chapter 7 Patient Preference Elicitation

The elicitation of patient preferences has become prominent in healthcare, along with the increasing degree of participation by patients in their own treatment decision-making. Although some conventional patient preference elicitation techniques exist, their outcomes are error prone and unreliable. We believe that software engineering can have an important role in developing patient preference elicitation systems, which address some of the outstanding issues in this process (Omidbaksh, et al., Oct., 2010).

The purpose of this chapter is to address our Objective 4 (see Section 1.2) by reviewing the existing techniques, outlining the outstanding issues, and finally presenting a novel approach to develop a model-based preference elicitation. We represent our Patient Preference Elicitation (PPE) system and review the strategies for personalizing the elicitation process.

7.1 Patient Preference Elicitation

Shared decision making (Charles, et al., 1999) is based on the partnership between physicians and patients to contribute actively in decision-making concerning the patients' preferred treatment options. Better treatment decisions are made by understanding the consequences of the decisions, which subsequently results in lower decisional conflict, higher satisfaction with decisions, and improved health psychological outcomes. The need for development of appropriate software tools is clear from the fact that the availability of such software would certainly help towards making shared decisions better integrated into clinical practice.

One major problem to be addressed arises from the fact that patients should be fully informed about their health situation/condition which itself is a non-trivial task. On the one hand, patients need to acquire and understand sufficient health information to participate actively in decision making, On the other hand, determining patients' preferences and information needs is a very complicated procedure, requiring several consultation sessions that do not fit readily into physicians' busy schedules. One can easily see that it is indeed extremely difficult for physicians, working independently, to find the time to explore patients' preferences adequately.

Here, the decision-making refers to the action of making decisions on the treatment options available to a patient. The consequence of this action, or in other words, post-treatment state of patient health includes the side effects of treatment and the post-treatment pain, suffering and inconvenience. For instance, a patient diagnosed with one type of cancer has a decision to make about various treatment options, such as surgery, radiation therapy, chemotherapy, etc. (McNeil, et al., 1982).

In general, patients are usually not sufficiently familiar with the options available and although the empirical probabilities exist for some of different treatment outcomes, the consequences of these options cannot be generalized. Therefore, patients have to be informed, based on their diagnosis, about the following:

(i) the available course of actions,

(ii) their possible consequences, and

(iii) the probabilities of associated outcomes.

Patients' preferences become more prominent especially for health conditions in which there is no high quality clinical evidence available about outcomes. In these situations, preferences are dependent on patients' attitude toward uncertainty and their tradeoffs between benefits and harms. For example, the choices between watchful waiting, radiotherapy or radical prostatectomy as treatment options for prostate cancer are considered preference sensitive (McNeil, et al., 1982). Patients have to incorporate their personal preferences, characteristics and values to choose among the options. Their beliefs, attitudes and values influence their preferences for outcomes and risks of treatment.

There is a great demand for software tool support towards increasing patients' involvement in decision making which requires improving of patient knowledge and results in reducing decisional conflict and passivity in decision making (Brennan & Strombom, 1998). Some efforts have been reported in developing Web-based decision aids; however, they are not adequately personalized. They do not necessarily provide the best kind of interactivity. No decision aid has been tailored for individuals' unique risk profile that allows, say, engaging values clarification exercises, and other such assessments. Long-term persistence with decisions, health related quality of life and cost of decision aids are not yet established. There are unresolved issues on dissemination, coordination and standards.

7.2 Patient Preference Elicitation Techniques

Patient preference elicitation in health care is defined as the process consisting of at least the following steps:

(i) describing the care options,

(ii) gathering and framing evidence in a format comprehensible to patients, and subsequently

(iii) measuring of patients' preferences (Taylor, 2000).

Conventionally, there are three techniques in use for measuring patient preferences: Standard Gamble (SG), Time Trade-Off (TTO) and Rating Scale (Ruland, 1999). Below, we briefly describe these three and provide a comparison table, which reviews them on different aspects.

Standard Gamble is a well-known technique to value health states based on pair-wise comparisons between two alternatives with two possible outcomes: a certain outcome and a gamble. Patients indicate their preferences by choosing to be either ill or treated with uncertain outcome. This technique is derived from expected utility theory.

Time Trade-Off is another technique that measures the time a patient trades off to avoid a specific health outcome. There are two certain outcomes in this technique, but comparison is based on two times. Patients choose their preference between (i) number of years in a healthy state and (ii) number of years in a poorer state of health.

The third technique is categorical rating scale. Patients are asked to assign the rated health states that have numerals assigned to them; patients rate their desirability of a certain health state as a place between the two ends of the scale which are best and worst states. This technique has three variations: Magnitude Estimation, Equivalence, and Willingness-To-Pay. Magnitude Estimation is the technique in which one outcome is taken as the standard to which the other outcomes are compared (Stevens, 1971). Equivalence is the technique in which patients are asked how many patients in state A are equivalent to 100 patients in state B (Patrick, et al., 1973) and Willingness-To-Pay is the technique in which patients are asked what portion of their household income would they be willing to pay to get from state A to state B (Thompson, 1986).

Among these techniques, Standard Gamble is the most difficult to understand for both physicians and patients, internally inconsistent and biased in the risk aversion direction. Time-Trade-Off is more costly and difficult to administer. Category scaling is the most promising and straight forward. All of these techniques are very time consuming for implementation in practice (Taylor, 2000). Table 7.1 shows a comparison of these three different techniques.

Table 7.1 Comparisons of Preference Elicitation Techniques

Method	**SG**	**TTO**	**Rating Scale**
Scale	Medium	Small	Large
Administration	Easy	Difficult	Easiest
Approach Measure	Utility	Value	Value
Cost	Expensive	More Expensive	Least Expensive
Comment	- Llewellyn& Shoemaker - Life & death conditions	- Developed by Torrence - Promising - Life & death conditions - Good validity level - More easily understood as compared to SG	- Froberg first choice - Promising

7.3 Outstanding Issues

There is a great concern that patients' preferences may be strongly influenced by the process of elicitation (Lloyd, 2003). Patients might even be evolving their preferences in the course of this process. Below, we note some of the fundamental problems in regards to the preferences discovered by the elicitation techniques (Lloyd, 2003).

Firstly, preferences are not stable. Studies show that patients reflect on their responses and change them repeatedly, causing dramatic changes in the values elicited at different points of time (Lloyd, 2003). As preferences are not “relatively complete”, patients construct them at the time of the elicitation either on the basis of their previous information, attitudes, and emotional states or on the basis of the newly presented information; therefore, preferences are completely affected by context and are controlled by heuristics. This in turn contributes significantly to patients’ ignorance of much of the presented information. Time pressure, task complexity, response mode, and motivation leads to use of heuristics, which may result in bias or error.

Secondly, elicited preferences may often be inaccurate from a statistical point of view. Inaccuracy of the preferences is due to instability and incompleteness of the choices made by patients. When we say preferences are not accurate, we mean that what patients choose may not be precise at the first instance and may change with time, additional knowledge obtained and other such factors. As reported in the literature, preference elicitation techniques are based on the assumption that patients accurately reflect their actual preferences, resulting in accurate outcomes of the elicitation process. Furthermore, different techniques chosen for eliciting preferences produce different results. Thus, the results reflect patients’ preferences only in part and the rest is based on the manner in which the elicitation process is defined.

Thirdly, preferences are also dependent on how adaptively relevant information is presented. When patients are asked about their preferences in different loss and gain frames, there is a reversal of preferences for the same equivalent information. Thus the correctness of the responses in terms of being the closest to the individual's actual preference is questionable. Prospect theory explains the fact that when patients have to make decisions in the face of uncertainty, they not only have a generalized loss aversion, but also are risk averse in the domain of gains and risk seeking in the domains of losses (Lloyd, 2003). When they avoid more risky alternatives in favor of less risky alternatives, they are considered as being risk averse, and when they prefer risky situations, they are considered as being risk seeking. Thus, framing effect plays a major role in the construction of patients' preferences.

Furthermore, patients' perception of risk and uncertainty is another major issue. Patients are not always capable of understanding probabilities, uncertainties and risk information, and not capable of incorporating them into their decision making. The format in which uncertainties are presented has a great impact on how they comprehend numbers. Studies show that when patients are faced with information as absolute and relative risks, they react differently and the outcomes become different.

7.4 Patient Preference Elicitation Software

We contend that with a personalized interactive elicitation process, we can reasonably address the aforementioned issues and thus support patients in the decision-making concerning the outcomes of their treatment options. Our approach is based on: (i) creation and use of a patient model, and (ii) development of a model driven interactive elicitation process.

A patient model is considered as a composition of a user model and the underlying disease model. In general, a user model is achieved by adapting to user knowledge, cognitive properties, goals and plans, moods and emotions, characteristics, discernable values and beliefs (Quarteroni & Manandhar, 2007). Moreover, a user model can be built by identifying the stereotype or the class that a user belongs to and by identifying the general properties of the users of that stereotype. The key questions regarding this aspect of our work are the following:

- How do we represent user-modeling information?
- What is the information that needs to be included in the user model? (parameter values, networks presenting belief, plans and goals)
- How can the patient model be created so as to properly link the user model with the disease model?
- To what level of detail should we model the disease?

- How can patient's preferences be formally represented?

To obtain patient preference information, we need to determine how the interaction can be made to best suit patients' needs. The primary questions considered for the development of an effective elicitation process are:

- How do we generate and manage questions for the process? The questions can be either created based on the acquired knowledge or can be extracted from a predefined question bank.
- What algorithms are needed for this iterative process to proceed step-by-step? We have to determine how 'some parameters' relevant to patients' preferences can be extracted from patients' response in each iterative cycle. Also, we have to address the optimal termination of the question and answer cycle.
- How can we detect and resolve patients' numeracy and framing problems? Previous research has identified numeracy and framing as being two of the major issues with patients. Checking the levels of patients' comprehension, and ways of adaptation to those levels should be considered. Yet another important aspect is determining if framing effect is addressed by presenting the information in all different formats, or in only neutral format, or based on the patients' profile.

We attempt to develop a novel approach and architecture to empower patients to be involved in their own treatment decision making by creating a model-based question and answer dialogue system for eliciting patients' preferences and by personalization and adaptation of this elicitation process. The proposed elicitation process and the personalization of this process will be discussed in Sections 7.4.1 and 7.4.2 respectively.

7.4.1 Process of Preference Elicitation for Patients

One general strategy in preference elicitation consists of employing methods to extract information from the user and determine the user's preferences. Typically, by querying the user about preferences for various possible outcomes, different attribute values, utilities or risk attitudes. However, querying the user repeatedly until obtaining all preference information is not only bothersome to the user but also infeasible. It is reasonable to select a small number of questions to ask from the user. Therefore, one of the key components of an elicitation process is selecting the optimal set of questions that can yield maximum information about an individual user's preference.

Ideally, it is desirable to ask questions from users in such a way that their response reflects a maximal increase in the value of the chosen approach to the decision problem. However, there are a large number of possible questions and evaluating the approach is complicated. After each question, user's beliefs regarding the preferences are updated and the value of possible questions to be asked next must be computed based on the updates. Thus, gathering as much information as possible from the user, but minimizing the number of queries to be asked from the user, has to be a major objective.

We believe that eliciting preferences by exclusively asking questions from the user is not effective in our problem domain since patients might be totally unfamiliar with the medical terminology. They are also burdened by their physical and emotional conditions.

Another strategy in preference elicitation is indirect reasoning on the basis of the user's conditions, past behavior or stereotype classification methods. This strategy is likewise not sufficient for patients since each and every human being has a unique intricate health state which could also vary with the length of the treatment process and they cannot absolutely be categorized into similar classes.

We intend to apply the two mentioned strategies to effectively elicit patients' preferences both by questioning the decision maker and by suitably inducting from the models that we will build for this purpose. To this aim, a personalized adaptable interaction which provides question and answer dialogue is needed.

However, we are not restricted to a simple "one style fits all" approach. We look at the ways to address patients with different levels of education, learning styles, background, etc. and to detect and resolve issues regarding framing and numeracy. The questions will be selected based on partially filled parameter values of the patient model from a pre-defined set of questions hand-crafted to form a Question Bank. Initially, some of the parameters will be derived from the patient profile, and more will be filled during subsequent iterations of the elicitation process.

Each answer is used to form/update the preference model. Each iteration or 'Question-Answer' of the process will be used to reason the potential preference of the patient. The format of the answers will be restricted to machine digestible forms. The interaction cycle will continue until the elicited preferences reach an admissible level of the appropriate criteria for evaluating the true benefit of the elicitation process.

7.4.2 Personalization and Adaptation of Preference Elicitation Process

As previously mentioned, we strongly believe that the preference elicitation process needs to be specifically personalized and adapted for each individual patient. Based on Kobsa et al. (Kobsa, et al., 2001), personalization and adaptation can be categorized primarily on the content of information and secondly on the presentation of information. To this aim, both on the presentation and content levels, we focus on:

- Making Numbers more transparent with visual displays and presenting numbers in the most comprehensible manner; visual displays like cartoons, films, types of graphic display line graph, bar graph, pie charts, risk ladder, pictographs are good means for this purpose.
- Presenting an array of graphic formats and requesting their own choice.
- Framing medical information in different but equivalent ways; a range of complementary formats (e.g. descriptive, numerical, absolute and relative

risk, numbers needed to treat, and graphical representations) will be included.

- Adjusting the information to patients' reading level, requested depth and sequence of information and presentation style preferences.

7.4.3 Framing of Information for Patients

Patients need to understand the quantitive information without the impact of the format and the framing of the information presented to them in order to compare the treatment options (Peters & Levin, 2008). Specifically, the quantitive information about treatment benefis and risks for patient can be represented in three different formats (Kobsa, et al., 2001): NNT (Numbers needed to treat), ARR (Absolute Risk Reduction), and RRR (Relative Risk Reduction), or in combination, COMBO (Combination of all the three). NNT is an empirically derived estimte of the number of patients who must be treated to expect that one patient will avoid an adverse event or outcome over a defined period of time. ARR is the decrease in disease incidence due to treatment therefore provides an estimate of absolute patient benefit. RRR is the decrease in disease incidence relative to those who are not taking treatment. Table 7.2 illustrates different framing formats of risk information.

In Table 7.3, a hypothetical example is shown; patients were asked to imagine that 40 out of 1000 people just like them will develop Disease Y over the next 5 years and they were presented with the information on benefits of Treatment A and Treament B in one of four ways.

Studies show that patients with better numeracy skills mostly correctly calculated or deemed to have understood the treatment benefits (Kobsa, et al., 2001). It is suggested to use written quantitative information only with patients with higher numeracy skills and to present risk reduction information as ARR or RRR rather than as NNT or COMBO. New presentation formats and new ways for patients to interpret quantitative health information are needed (Brust-Renck, et al., 2013).

In a nutshell, presentation of information in different formats results differently in the patients' assessements of the benefits of the treatments. Patients with different numeracy levels and demographic charactericss need the information to be presented appriopately to them to avoid the framing issue.

Table 7.2 Different Framing Formats of Risk Information

No.	Framing Format	Description
1	NNT	Numbers Needed to Treat
2	ARR	Absolute Risk Reduction
3	RRR	Relative Risk Reduction
4	COMBO	1,2,3

Table 7.3 Examples for Different Framing Formats

No.	Framing Format	Example A	Example B
1	RRR	Treatment A reduces the chance that you will develop Disease Y by 25%.	Treatment B reduces the chance that you will develop Disease Y by 10%.
2	ARR	Treatment A reduces the chance that you will develop Disease Y by 10 per 1000 persons.	Treatment B reduces the chance that you will develop Disease Y by 4 per 1000 persons.
3	NNT	100 persons just like you would have to be treated with Treatment A for 5 years for a benefit against Disease Y to be evident in one of you.	250 persons just like you would have to be treated with Treatment B for 5 years for a benefit against Disease Y to be evident in one of you.
4	COMBO	1, 2, 3	1, 2, 3

7.4.4 Personalization of Healthcare Information for Patients

We focus on improving the patient experience in understanding the risk information. We propose the use of decision tables to represent the adaptation rules for this purpose. Our adaptation is based on a stereotype model. Eventually, the validation of this adaptation should be obtained by conducting a controlled experiment as a future work of this thesis.

We define three user stereotype models for the representation of risk information based on our module of numeracy assessment. Table 7.4 shows the stereotypes and the associated representation method.

Table 7.4 Different User Stereotypes

User Stereotype	**Representation Method**
Low Numerate	Information in the form of visual and gist visual.
Intermediate	Information in the form of modified written text (gist) and oral text.
High Numerate	Information in the form of written text.

Table 7.5 represents the different adaption forms needed to be applied for the presentation to different stereotypes.

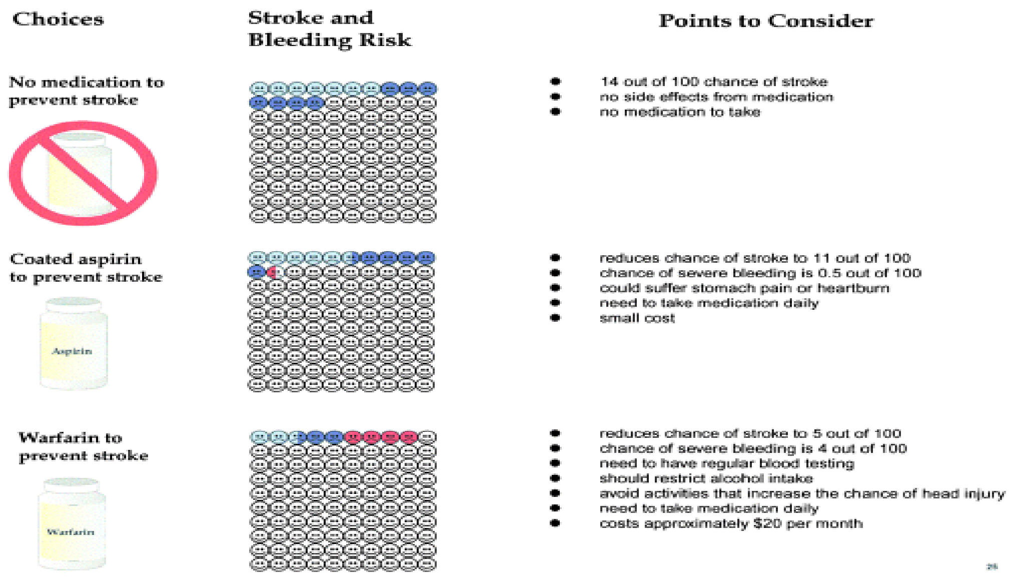
Table 7.5 Adaptation Forms

Features	**Value (Conditions)**
Display information	Text, Modified Text (gist), Visual, Oral
Framing format	NNT, ARR, RRR and COMBO

Table 7.6 and Table 7.7, show visual representation methods (Brust-Renck, et al., 2013) and furthermore, some examples of the different representation of risk information are illustrated in Figure 7.1.

Table 7.6 Different Methods of Visual Representation of Risk Information

No.	Information Representation Method
1	Euler diagram
2	100-square grid
3	2x2 table Venn
4	Icon array
5	Line graph
6	Pie chart
7	Venn diagram



- 14 out of 100 chance of stroke
- no side effects from medication
- no medication to take

- reduces chance of stroke to 11 out of 100
- chance of severe bleeding is 0.5 out of 100
- could suffer stomach pain or heartburn
- need to take medication daily
- small cost

- reduces chance of stroke to 5 out of 100
- chance of severe bleeding is 4 out of 100
- need to have regular blood testing
- should restrict alcohol intake
- avoid activities that increase the chance of head injury
- need to take medication daily
- costs approximately $20 per month

Figure 7.1 Visual and Text Representation of Information (Wilhelms & Reyna, 2013)

Table 7.7 Different Visual Representation Methods and their Purpose

Type	**Subtypes**	**Purpose**
Graphical Representation	Stacked bar graph	Avoid denominator neglect
Graphical Representation	Bar Graphs	Relative difference between two magnitudes
Graphical Representation	Pie Charts	Relative magnitude Incidents of adverse events that occur more than 1 percent of the time are effectively communicated through pie charts.
Graphical Representation	Icons	Relative magnitude: Events that occur less than 1 percent of the time are effectively communicated through icon-based pictographs.
Graphical Representation	Line Graphs	Change over time
Visual Formats	2x2 tables	Avoid probability judgments
Visual Formats	Venn Diagrams	Avoid probability judgments
Visual Formats	Euler Diagram	Avoid probability judgments
Visual Formats	Icon Arrays	Understanding of relative risk and relative magnitude while making denominator clear

7.5 Architectural Requirements

As discussed, it is clear that patient preferences are subjective in nature and could be significantly influenced by the elicitation process. Further, patients may construct their preferences during this process. Our research objective is to develop a system supporting interactive elicitation process driven by a collection of models. The system must address the following challenges:

Unlike the set of options in general recommender systems for movies, books or restaurants, the set of options over which patients choose treatments of diseases is a small set. However, in this problem domain where the decisions about personal health are much more serious than, say, going to a restaurant or buying a product, we are confronted with two facts: (1) there is a lack of certainty in the outcomes of the treatment options, which is the nature of the medical domain, and (2) there is a lack of certainty in patients' choices; when more information is available their preferences could change. The former is out of our scope and is typically the domain of researchers in health care. The latter is the focus of this work. If patients possess inadequate information for the decision making, they either do not have access to the relevant information required for their decision making or do not perceive the information presented to them correctly. Therefore, preferences tend to change and be unstable, and consequently, patients' confidence level becomes low.

If we assume that a patient has the preference values for two options A and B at time T1, they are different from their preference values for the same options at a later time T2 and so on. Many different events may occur between T1 and T2, and between T2 and T3, and so on. Hence there could be a temporal progression towards stability in the patient's preferences over the time span available to make a decision. This progression continues either until the patient reaches a decision and is satisfied with the choice made; or until the patient runs out of time.

Newly perceived information in between the two points of time, say T1 and T2, could be one major cause for the change of preferences (assuming that their state of health is stable). Some events act as "catalysts" for patients' constructing their relative preferences of the options by providing access to new information.

Furthermore, patients may not be able to comprehend the relevant information available for constructing their preferences. They may have difficulty in perceiving risk, uncertainty and probabilities. The format and the framing of the information have great impact on the perception of the newly presented information. Obviously, different patients have different level of perceiving the information; they may have framing and/or numeracy issues. In this research with our proposed preference elicitation process, we considered the above framing and numeracy issues and address them by techniques that we have termed "perception boosters". (A generic name referring to the techniques that address numeracy and framing issues in order to boost patients' perception of information presented to them.)

Figure 7.2 gives an overview of the architecture of our PPE system. The core components of our system are: Knowledge Base which consists of models and Question Bank, User Interface which manages the user interaction with the system and a Model-based-Reasoner which performs the reasoning. We have a knowledge base, an inference engine, and a user interface that form an expert system. The knowledge base includes facts and rules and the inference engine uses the rules to the existing facts to deduce more new facts.

The output of this system is an instance of patients' preferences that can be applied in their treatment decision-making. The responsibilities of the preference elicitation process are distributed among various modules of the system as shown in Figure 7.8 and further elaborated below.

- Dialogue Manager manages the question and answer interactions.
- Perception Booster is composed of

  - Detection Manager that detects framing and numeracy issues of a given patient. It results in execution under the control of Dialogue Manager and

  - Resolution Manager that resolves adaptively the detected framing and numeracy issues for a given patient.

- Catalyst Manager, as a monitoring agent, proactively displays information about the opinions of other patients, physicians, caregivers and family members.

- Convergence Manager detects the convergence based on halting criteria.

Preference Model is built by direct comparison of available options obtained both by questioning the patient and by the induction of preference relations derived from the patient model. For our problem domain, the set of possible options at any given state of the patient, given the current state of the progress of the disease is finite and discrete and the members of the set are assumed to be independent of each other. Patient preferences, obtained by the interactive elicitation process, will be represented by the order of the elements of this set as a directed acyclic graph with labeled edges. Labels associated with the edges are used to show quantitative or qualitative parameters such as patient's confidence level on each option, patient's perception of the information or "catalyst" effect on the patient's preference. Here, the challenge is in extracting patients' preferences under the condition that patients are well informed, have understood the associated consequences and are finally satisfied with their choices.

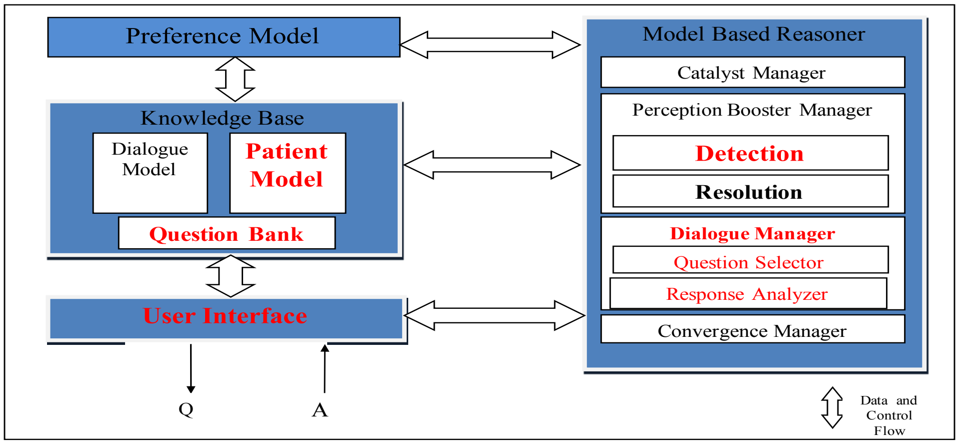


Figure 7.2 Architectural Components of PPE system

Our further work includes personalization and integration of the information in the elicitation process. Our ultimate goal is to evolve the design of PPE system that supports patients in their decision-making regarding their treatments.

7.6 Summary

In this chapter, we addressed the outstanding issues regarding patient preference elicitation by presenting the methods, algorithms and architecture that allows us to build PPE, a system that can support health care professionals to interactively elicit patients' preferences for the purpose of assisting them in their treatment decision making.

In the next chapter, we summarized the contributions and the future work of this thesis.

Chapter 8 Conclusions and Future Work Directions

In this thesis, we proposed a goal-driven confidence-based adaptive model for the assessment of numeracy skill of patients, which is a milestone in patient education and consequently eliciting patients' preferences in medical decision-making.

The novelty of our contributions lies in the fact that the assessment is not only based on the knowledge of each individual, but also on the confidence in that knowledge, it is adaptive to each individual patient, and covers the full sets of numeracy skills (Omidbaksh & Ormandjieva, Dec., 2015). The quality of C-PNA model proposed is assessed using a new numeracy assessment methods quality model proposed in this thesis and inspired by ISO/IEC 25022. Empirical studies provided the evidence for patients' high Satisfaction and Trust in C-PNA, as well as significant Effectiveness and Usage Efficiency of our patient numeracy assessment method.

In this research, we assume that patients are considered computer literate and are willing to use technology, since our model is highly dependent on the patient acceptance of technology and their ease of use. Our C-PNA model is not designed for the patients who are under stress, hospitalized or in emergency units with critical illness.

8.1 Contributions

As discussed above, there is no model which takes confidence as a parameter in consideration for adaptive assessment. We incorporated the parameter of confidence in the adaptive testing and we worked on two dimensions of knowledge and confidence for the assessment.

We developed a novel model for building Confidence-based Patient Numeracy Assessment (C-PNA) system that is based on this parameter. C-PNA reflects both knowledge and confidence in the assessment and obtains more efficient and productive results for the assessment. Consequently, this approach leads to behavioral outcomes and empowers patients to act.

In our approach, we adopted a conceptual model for assessment from educational psychology domain. Additionally, we borrowed relevant ideas from Confidence Based Learning (CBL) methodology and then extended them to the field of Computerized Adaptive Testing (CAT) to develop a confidence-based model of adaptive testing for numeracy assessment.

In this research, we assumed that patients are considered computer literate and are willing to use technology, since our model is highly dependent on the patient acceptance of technology and their ease of use.

To this end, the contributions of this thesis are as follows:

1. Characterization of an Adaptive Testing Model for assessment of patient numeracy in healthcare domain (Omidbaksh & Radhakrishnan, May, 2014).
2. Definition of a Goal-driven modeling for confidence-based patient numeracy assessment (Omidbaksh & Ormandjieva, Dec., 2015).
3. Modeling of Quality Measurement for Numeracy Assessment Methods (Omidbakhsh & Ormandjieva, 2015).
4. Proposing a Patient Preference Elicitation Model (Omidbaksh, et al., Oct., 2010).
5. Development of a comprehensive question bank for the subject of numeracy (Omidbaksh & Ormandjieva, Dec., 2015).
6. Development of an online platform for conducting controlled experiments in healthcare domain (Omidbakhsh & Ormandjieva, 2015).
7. Empirical Evaluation of Confidence based Model for numeracy assessment in Healthcare.

8.2 Limitations of the Approach

The proposed confidence-based numeracy method has some limitations that are as follows:

1) Language, context, and culture are not considered in the assessment.
2) Potential confounders such as distress due to illness or cognitive deficits are not considered in the assessment.

8.3 Future Work

In this thesis we were inspired by the conceptual math model (Strawderman, 2009) which encompasses four dimensions in math assessment. The fourth dimension, social/motivational dimension, with the continuum of behavior and the two extremes of pursuit and avoidance, is not covered in this work and could be added to the system as a master's thesis.

For the proposed PPE system two components are required: numeracy assessment and personalization. The former was covered in this thesis. We also provided the theory for the latter, however the implementation of the personalization and the related empirical studies could be considered as a future work for a master student.

As emotion plays an important role in the process of decision-making and has great impact on the learning process, there is a need to explore it for the topic of assessment. In this work, we focused on the factor of confidence, based on the choice of options a tester makes; a future PhD work could include the facial emotion recognition as the basis of confidence measurement.

Another future work could be the extension of the measurement model by adding other quality measures to our model such as usefulness and context coverage.

Usefulness is defined to which a user satisfied with their perceived achievement of pragmatic goals which includes the result of use and the consequences of the results.

Context coverage which encompasses both context completeness and flexibility, is described as the degree to which a product or system can be applied with effectiveness and efficiency and freedom of risk in both specified context and beyond the context based on ISO 25022:2016.

In this work, we focused on the subject of numeracy. Eventually, our model could be applied for any other subjects for assessment; this could be studied as another PhD thesis. In that case, issues such as testers' cheating and motivation should be taken into consideration. The testers' reviews and response time are still open challenges in the domain and algorithms for cheating detection are highly in demand.

Our work was focused on the healthcare domain due to high demand and sensitivity of the topic, however the work can be applied for other domains such as education and psychology as well.

References

Adams, T. M., 2009. *The Importance of Confidence in Improving Educational Outcome.* 25th, Annual Conference on Distance Teaching and Learning.

Amery, D. & Rich, H., 2008. *Updating IEEE 1471: Architecture Frameworks and Other Topics,* Boston: MIT.

Ancker, J. & Kaufman, D., 2007. Rethinking Health Numeracy: A Multidisciplinary Literature Review. *Journal of the American Medical Informatics Association,* 14(6), p. 713–721.

Apter, A., 2006. Asthma Numeracy Skill and Health Literacy. *Journal of Asthma,* 43(1), pp. 705-710.

Ashcraft, K. M., 2001. The Relationships among Working Memory, Math anxiety, and Performance. *Journal of Experimental Psychology: General,* 130(2), pp. 24-237.

Basili, V. R., 1992. *Software Modeling and Measurement: The Goal Question Metric Paradigm,* Maryland: Computer Science-TR-2956 (University of Maryland Institute for Advanced Computer Studies-Technical Report).

Basili, V. R., Rombach, H. D. & Caldiera, G., 1994. The Goal Question Metric. *The Electronic Journal of Information System of Evaluation,* 14(2), pp. 264-272.

Berander, P. & Jönsson, P., 2006. *A Goal Question Metric Based Approach for Efficient Measurement Framework Definition.* ISESE '06, Proceedings of ACM/IEEE International Symposium on Empirical Software Engineering.

Bettenburg, N., Premraj, R., Zimmermann, T. & Kim, S., 2008. *Extracting Structural Information from Bug Reports.* Leipzig, ACM, pp. 27-30.

Betz, N. E., 1978. Prevalence, Distribution, and Correlates of Math Anxiety in College Students. *Journal of Counseling Psychology,* 25(5), pp. 441-444.

Brennan, P. F. & Strombom, I., 1998. Improving Health Care by Understanding Patient Preferences: The Role of Computer Technology. *Journal of the American Medical Informatics Association,* 5(3), pp. 257-262.

Bruno, J. E., 1990. Design of Technology-Based Information Reference Grouping systems for Use in Large Urban Schools. *The Urban Review,* 22(3), pp. 163-181.

Bruno, J. E., 1993. Using Testing to Provide Feedback Support to Instruction: A Reexamination of the Role of Assessment in Educational Organizations. In: *Item Banking: Interactive Testing and Self-Assessment.* NATO ASI Series: UCLA, pp. 190-209.

Bruno, J. E. & Abedi, J., 1989. Concurrent Validity of Information Referenced Testing Format Using MCW-APC Scoring Methods. *Journal of Computer Based Instruction,* 20(1), pp. 21-25.

Bruno, J. E. et al., 2005. *Method and System for Knowledge Assessment and Learning Incorporating Feedbacks.* United States, Patent No. 20030190592.

Brust-Renck, P., Royer, C. & Reyna, V., 2013. Communicating Numerical Risk: Human Factors That Aid Understanding in Health Care. *Human Factors Ergon,* 8(1), pp. 235-276.

Carpenter, T. P. et al., 1981. *Mathematics Educations Research: Implications for the 80's,* Alexandria, VA: National assessment. In E. Fennema (Ed.).

Charles, C., Gafni, A. & Whelan, T., 1999. Decision-Making in the Physician–Patient Encounter: Revisiting the Shared Treatment Decision-Making Model. *Social Science & Medicine,* 49(5), pp. 651-661.

Cokley, K., 2000. An Investigation of Academic Self-Concept and its Relationship to Academic Achievement in African American College Students. *Journal of Black Psychology,* 20(2), pp. 148-164.

Davis, T. et al., 1991. Rapid Assessment of Literacy Levels of Adult Primary Care Patients. *Journal of Family Medicine,* 23(6), pp. 430-440.

Davis, T. et al., 2002. Health Literacy and Cancer Communication. *CA,* 52(3), pp. 134-149.

Donelle, L., Hoffman-Goetz, L. & Arocha, J. F., 2007. Assessing Health Numeracy among Community-Dwelling Older Adults. *Journal of Health Communication,* 12(7), pp. 651-659.

Easterbrook, S., Singer, J., Storey, M. A. & Damian, D., 2008. Selecting Empirical Methods for Software Engineering Research. In: *Guide to Advanced Empirical Software Engineering.* Toronto: Springer, pp. 285-311.

Eccles, J. S. & Jacobs, J. E., 1986. Social Forces Shape Math Attitudes and Performance. *Signs.,* 11(1), pp. 367-380.

Echternacht, G. J., 1972. The Use of Confidence Testing in Objective Tests. *Review of Educational Research,* 42(2), pp. 217-236.

Estrada, C., 2004. Literacy and Numeracy Skills and Anticoagulation Control. *American Journal Medical Science,* 328(2), pp. 88-93.

Fagerlin, A. et al., 2007. Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making,* 27(5), pp. 672-680.

Fenton, N. & Bieman, J., 2014. *Software Metrics: A Practical and Rigorous Approach.* 3rd ed. CRC Press: Taylor and Francis Group.

Goldbeck, A., Ahlers-Schmidt, C., Paschal, A. & Dismuke, E., 2005. A Definition and Operational Framework for Health Numeracy. *American Journal of Preventive Medicine,* 29(4), pp. 375-376.

Goldbeck, A., Ahlers-Schmidt, C., Paschal, A. & E. Dismuke, E., 2005. A definition and operational framework for health numeracy. *American Journal of Preventive Medicine,* 29(4), pp. 375-376.

Goldbeck, A., C. Ahlers-Schmidt, Paschal, A. & Dismuke, E., 2005. Health Numeracy in Adults: An Overview and Annotated Bibliography.

Goldstein, H. & Wood, R., 1989. Five Decades of Item Response Modelling. *British Journal of Mathematical and Statistical Psychology,* 42(1), pp. 139-167.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J., 1991. *Fundamentals of Item Response Theory.* Newbury Park(CA): Sage Press.

Health Canada Official, 2015. *Health Canada.* [Online] Available at: http://www.hc-sc.gc.ca/sr-sr/advice-avis/reb-cer/consent/index-eng.php[Accessed 01 10 2015].

Hodge, M. B., 1999. Do Anxiety, Math Self-Efficacy, and Gender Affect Nursing Students' Drug Dosage Calculations?. *Nurse Education,* 24(4), pp. 36-41.

Huizigna, M. et al., 2009. Literacy, Numeracy and Portion-size Estimation Skills. *American Journal of Preventive Medicine,* 36(4), pp. 324-328.

Huizinga, M., 2008. Development and validation of the Diabetes Numeracy Test (DNT). *BMC Health Services,* 48(10), pp. 737-746.

Hunt, D. P. & Furustig, H., 1989. *Being Informed, Being Misinformed and Disinformation: A Human Learning and Decision Making Approach,* Karlstad: Technical Report PM 56:238.

Hunt, D., 2003. The Concept of Knowledge and How to Measure it. *Journal of Intellectual Capital.*

ISO/IEC DIS, 2., 2016. *System and Software Engineering-Systems and Software Quality.* Geneva: ISO/IEC.

Kang, T., 2006. *Model Selection Methods for Unidimensional and Multidimensional IRT.* Madison: University of Wisconsin.

Kobsa, A., Koenemann, J. & Pohl, W., 2001. Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationship. *The Knowledge Engineering Review,* 16(2), pp. 111-155.

Kruchten, P., 1995. Architectural Blueprints- The 4+1 View Model of Software Architecture. *IEEE Software ,* 12(6), pp. 42-50.

Lipkus, M., Samsa, G. & Rimer, B. K., 2001. General Performance on a Numeracy Scale among Highly Educated Samples. *Medical Decision Making,* 21(1), pp. 37-44.

Lipkus, I. et al., 2008. Clinical Implications of Numeracy: Theory and Practice. *Annals of Behavioral Medicine*, 35(3), pp. 261-274.

Lloyd, F. & Reyna, V., 2001. A Web Exercise in Evidence-based Medicine Using Cognitive Theory. *Journal of General Internal Medicine,* 16(2), pp. 94-99.

Lloyd, A. J., 2003. Threats to the Estimation of Benefit: Are Preference Elicitation Methods Accurate? *Health Economics,* 12(5), pp. 393-402.

Lopez, F. G., Lent, R. W., Brown, S. D. & Gore, P. A., 1997. Role of Social-cognitive Expectations in High School Students' Mathematics-related Interest and Performance. *Journal of Counseling Psychology,* 44(1), pp. 44-52.

McNeil, B., Pauker, S., Sox, H. & Tversky, A., 1982. On the Elicitation of Preferences for Alternative Therapies. *New England Journal of Medicine,* Volume 306, pp. 1259-1262.

Murray, E., Pollack, L., White, M. & Lo, B., 2007. Clinical Decision Making: Patients' Prefrences and Experience. *Patient Education and Counseling,* Volume 65, pp. 189-196.

Omidbakhsh, M., Radhkrishnan, T. & Mudur, S., March, 2010. *Numeracy Assessment: A tool for Empowering Patients.* Lima, Proceedings of Pan American Health Care Exchanges Conference (PAHCE), pp.123-127.

Omidbaksh, M., Radhakrishnan, T. & Mudur, S., Oct., 2010. *Patient Preference Elicitation Empowerment.* New York, Proceedings of ACM/IEEE 32nd International Conference on Software Engineering (SEHC), pp. 19-23.

Omidbaksh, M. & Radhakrishnan, T., May, 2014. *An Adaptive Testing Model for Assessment of Numeracy in Patients.* New York, CBMS, pp. 475-476.

Omidbakhsh, M. & Ormandjieva, O., July, 2015. *Numeracy Assessment.* [Online] Available at: assessnumeracy.com [Accessed 3 8 2016].

Omidbaksh, M. & Ormandjieva, O., Dec., 2015. *Goal-driven Modeling for Confidence-based Patient Numeracy Assessment: C-PNA.* Berlin, Elsevier, pp. 213-220.

Omidbaksh, M. & Ormandjieva, O., July, 2016. *Measuring the Quality of Numeracy Skill Assessment in Health Domain.* Las Vegas, The 2nd International Conference on Health Informatics and Medical Systems (HIMS'16).

Omidbaksh, M. & Ormandjieva, O., Aug., 2016. *A Survey of Numeracy Assessment Approaches for Patient E-learning.* Bandung, American Scientific Publishers.

Patrick, D. L., Bush, J. & Chen, M. M., 1973. Methods for Measuring Levels of Well-being for a Health Status Index. *Health Services Research,* 8(3), pp. 228-245.

Peters, E. & Levin, I. P., 2008. Dissecting the Risky-choice Framing Effect: Numeracy as an Individual-difference Factor in Weighting Risky and Riskless Options. *Judgment and Decision Making,* Volume 3, pp. 345-448.

Quarteroni, S. & Manandhar, S., 2007. *User Modelling for Personalized Question Answering.* Rome, Springer, pp. 386-397.

Reckase, M. D., 1981. *Final Report: Procedures for Criterion Referenced Tailored Testing,* Columbia: University of Missouri.

Reenskaug, T. & Coplien, J., 2009. *The DCI Architecture: A New Vesrion of Object-Oriented Programming.* [Online] Available at: www.artima.com [Accessed 02 06 2015].

Rothman, R., Montori, V., Cherrington, A. & Pignone, M., 2008. Perspective: The Role of Numeracy in Health Care. *Journal of Health Communication,* 13(6), pp. 583-595.

Rozanski, N. & Woods, E., 2011. *Software Systems Architecture: Working with Stakeholders Using Viewpoints and Perspectives.* ACM : Addison-Wesley Professional.

Ruland, C., 1999. Decision Support for Patient Preference-based Care Planning Effects on Nursing Care and Patient Outcomes. *Journal of the American Medical Informatics,* 6(4), pp. 304-312.

Schapira, M. M. et al., 2012. The Numeracy Understanding in Medicine Instrument: a Measure of Health Numeracy Developed Using Item Response Theory. *Medical Decision Making,* 32(6), pp. 851-865.

Scholtz, J., 2004. *A Framework for Real-World Software System Evaluations.* New York, Proceeding CSCW’04 of the conference on computer supported cooperative work, pp. 600-603.

Schwartz, L. M., Woloshin, S., Black, W. C. & Welch, H., 1997. The Role of Numeracy in Understanding the Benefit of Screening Mammography. *Annals of Internal Medicine,* 127(11), pp. 966-972.

Shuford, E. H. & Brown, T. A., 1973. *Quantifying Uncertainty into Numerical Probabilities for the Reporting of Intelligence,* Advanced Research Projects Agency: Rand Report Prepared for the Defense.

Skemp, R., 1971. *The Psychology of Mathematics.* Baltimore(MD): Penguin Books.

Stevens, S., 1971. Issues in Psychophysical Measurement. *Psychological Review,* 78(5), pp. 426-450.

Strawderman, V. W., 2009. *Math Anxiety Model, PhD Dissertation.* Atlanta: Department of Math, Georgia State University.

Taylor, T., 2000. Understanding the Choices Patients Make. *Journal of the American Board of Family Medicine,* 13(2), pp. 124-133.

Thissen, D. & R. J. Mislevy, R. J., 2000. *Computerized Adaptive Testing: A Primer, Testing Algorithms..* Mahwah(NJ): Wainer, H. (Ed.) Lawrence Erlbaum Associates.

Thompson, M. S., 1986. Willingness to Pay and Accept Risks to Cure Chronic Disease. *American Journal of Public Health,* 76(4), pp. 392-396.

Tobias, S. & Weissbrod, C., 1980. Anxiety and Mathematics: An Update. *Harvard Educational Review,* 50(1), pp. 63-70.

US Department of Health and Human Services, 2000. *Healthy People with Understanding and Improving Health Objectives for Improving Health.* 2nd: US Government Printing Office.

Wainer, H. & Mislevy, R., 2000. Item Response Theory Caliberation and Estimation. In: H. Wainer, ed. *Computerized Adaptive Testing: A Primer.* Mahwah(NJ): Lawrence Erlbaum Associates, pp. 61-100.

Weintraub, S., 2000. Neuropsychological Assessment of Mental State. *Principles of Behavioural and Cognitive Neurology,* 12(5), pp. 121-173.

Weiss, B. et al., 2005. Quick Assessment of Literacy in Primary Care: The Newest Vital Sign. *Annals of Family Medicine,* 3(6), pp. 514-522.

Weiss, D. J. & Kingsbury, G. G., 1984. Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement,* 21(4), pp. 361-375.

Wilhelms, E. A. & Reyna, V. F., 2013. Effective Ways to Communicate Risk and Benefit. *American Medical Association Journal of Ethics,* 15(1), pp. 34-41.

Wohlin, C. et al., 2000. *Experimentation in Software Engineering: An Introduction.* Norwell(MA): Kulwar Academic Publishers.

Woloshin, S., 2005. Can Patients Interpret Health Information? An Assessment of the Medical Data Interpretation Test. *Medical Decision Making,* 25(3), pp. 290-300.

Woloshin, S. et al., 2000. A New Scale for Assessing Perceptions of Chance: A Validation Study. *Medical Decision Making,* 20(3), pp. 298-307.

Woloshin, S., Schwartz, L. M. & H.G.Welch, H. G., 2005. Patients and Medical Statistics: Interest, Confidence, and Ability. *Journal General Internal Medicine,* 20(1), pp. 996-1000.

Yin, R. K., 2009. *Case Study Research: Design and Methods.* Textbook: Sage.

Appendix A: (Controlled Experiment 1)

Numeracy Questions

1. Numeration 14. Your target blood sugar is between 60 and 120. Please choose the value below that is in the target range:

A.55 B.145 C.180

2. Numeration 141. If your weight is 150 pound. Please choose the value below that is less than your weight:

A. 115 B.250 C. 200

3. Addition 25. You take 10 units of insulin lispro and 16 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?

A. 6 B. 26 C. 160

4. Addition 251. You take 8 units of insulin lispro and 12 units of insulin glargine before breakfast. What is the total number of units of insulin you take before breakfast?

A. 20 B. 4 C. 12

5. Addition 252. You take 20 units of insulin lispro and 10 units of insulin glargine daily. What is the total number of units of insulin you take daily?

A. 10 B. 200 C.30

6. Multistep (addition) 311. If you had 400 calories for breakfast, 800 calories for lunch and 700 calories for supper. What is the total calories you had yesterday?

A. 500 B. 1900 C. 2200

7. Subtraction 111. If you are allowed to take 5 servings of carbohydrate per day and you have already have 2 servings of carbohydrate before lunch. How many servings of carbohydrate you are allowed to have for the rest of the day?

A. 10 B. 7 C.3

8. Subtraction 112. If you are allowed to have no more than 2000 calories per day and you have already taken 1400 calories, then how many calories you are allowed to have for supper?

A. 600 B. 3400 C. 1600

9. Multistep (addition subtraction) 211. If you have to have no more than 1800 calories per day and you have already taken 400 calories for breakfast and 600 calories for lunch, how many calories you are allowed to have for supper?

A. 200 B. 800 C.1200

10. Multiplication 16. You test your blood sugar 4 times a day. How many strips do you need to take with you on a 2-week vacation?

A. 96 B. 56 C. 8

11. Multiplication 18. You have a prescription for repaglinide 1 mg pills. The label says, “Take 2 mg of repaglinide with breakfast, 1 mg with lunch and 3 mg with supper.” How many pills should you take with supper?

A. 3 B. 6 C. 4

12. Multiplication 19. You have a prescription for metformin extended release 500 mg tablets. The label says, “Take 1 tablet with supper each night for the first week. Then, increase by 1 tablet each week for a total of 4 tablets daily with supper.” How many tablets should you take with supper each night the second week?

A.4 B.5 C.2

13. Multiplication 26. The doctor tells you to take 2 units of insulin for every 1 serving of carbohydrate you eat. How many units of insulin do you take for 6 servings of carbohydrate?

A. 12 B. 8 C. 4

14. Multiplication 27. 1 unit of insulin lowered your blood sugar by 30 points. How much does 4 units of insulin lower your blood sugar?

A. 34 B.120 C. 26

15. Multistep (addition, multi)20. You have only a few pills left in your pill bottle. Your doctor’s office needs 3 days to process a new prescription and your pharmacy needs 2 days to fill it. You take 2 pills a day. What is the least amount of pills that should be in your prescription bottle when you call for a renewal?

A. 10 B. 7 c.12

16. Division 21. For your diabetes, you take 1 pill two times per day. When you get your refill, the bottle has 60 pills. How many days supply do you have?

B. A.60 B.59 C. 30

17. Division 28- You are given the following instructions: “Take 1 unit of insulin for every 7 grams of carbohydrate you eat.” How much insulin do you take: When you eat 49 grams at Breakfast?

A. 42 B.7 C.6

18. Division 29- You are given the following instructions: “Take 1 unit of insulin for every 7 grams of carbohydrate you eat.” How much insulin do you take: When you eat 60 grams at Lunch?

A. 8 B.7 C.6

19. Division 30- You are given the following instructions: “Take 1 unit of insulin for every 7 grams of carbohydrate you eat.” How much insulin do you take: When you eat 98 grams at Supper?

A. 14 B. 12 C. 10

20. Division 31. You have been told to cut your insulin in half for a colon test. Your usual dose is 41 units. What amount should you take for the colon test?

A. 11 B. 20 C. 41

Question number	Topic	Difficulty level
1	numeration	1
2	numeration	1
3	addition	2
4	addition	2
5	addition	2
6	Multistep0	**2(start)**
7	subtraction	3
8	subtraction	3
9	Multistep1	**3(start)**

10	multiplication	4
11	multiplication	4
12	multiplication	4
13	multiplication	4
14	multiplication	4
15	Multistep2	**4(start)**
16	division	5
17	division	5
18	division	5
19	division	5
20	division	5

Appendix B: (Controlled Experiments: 2 and 3)

To conduct the controlled experiment, there are a couple of forms which are as the following:

- **Profile questionnaire:** General information is gathered by asking the test takers to fill out a form. This form includes information about username, age, gender, and education level of the test taker. This information will be gather through Form1.
- **Tests:** Two series of questions are presented to the test takers randomly each associated to the two types of methods.
- **Satisfaction and Emotional Response Questionnaires:** The Satisfaction Questionnaire includes questions regarding the test takers level of satisfaction through the experiment and the Emotional Response Questionnaire includes information regarding the enjoyment, announce, stress and frustration of the test takers during the tests. This information is gathered through Form2 and Form3.

Furthermore, these are the step we follow to conduct the experiment.

Step 1: Welcome and prepare the test taker.

We introduce the process and explain the purpose of the study and their role in the study.

Step 2: Ask the test taker to complete Form1.

Step 3: Choose assessments (CB or NCB) randomly and explain assessment process.

We explain that the assessment is a question-and-answer format, and we describe how to record answers. We inform them that there is no time limit for completing the assessments.

Step 4: Administer assessment(s).

We administer one assessment at a time. We ask the test taker to complete the questions and record their answers on the answer sheets. We also offer assistance for reading and understanding the instructions.

Step 5: Receive feedback.

We ask the test taker to fill out Form2.

Step 7: Record results.

We thank the test taker for participating in our study and records the results of both assessments.

CONSENT Form

A. Purpose

   I have been informed that the purpose of the research is as follows:

   Numeracy in healthcare domain is a measure of the ability of patients to understand and digest numerically presented information. A tool is built to assess the level of numeracy.
   This study is aimed to see if using this tool would be useful to assess numeracy level of patients in order for the clinicians to present information regarding their health care.
   The goal is to prove or disprove that this tool which is based on our numeracy assessment model is more effective than other existing methods.

B. Procedures

- I understand that I would be asked to sign up in a website at a time that is mutually agreeable to both myself and the principal investigator.
- I understand that after signing up on the website, I will be asked to answer a series of questions using the prototype of the aforementioned tool in two parts, namely TEST1 and TEST2.
- I understand that after completing the two tests, I will continue answering the survey questionnaire associated to each part.
- I understand that my participation will be recorded on the computer system for the purpose of extracting statistics to complete the goal of the study.
- I understand that the entire session will last between 30 to 45 minutes and not longer than an hour depending on the answers provided to the questions.
- I understand that my name will not be associated with any of the answers given to the questions asked.
- I understand that my age, gender, and education level will be used only for statistical purposes in conjunction with the answers given to the questions asked during the session.

C. Risks and Benefits

I understand that there are no risks in participating in this study. I will be viewing and using a prototype while answering series of questions during and after the testing related to this prototype.

D. Conditions of Participation

- I understand that I am free to withdraw my consent and discontinue my participation at any time without negative consequences.

- I understand that in the eventuality that I should choose to withdraw my consent and discontinue my participation at any point during the process, any data collected up until that point will still be used in this research.

- I understand that my participation in this study is CONFIDENTIAL (i.e. the researcher will know, but will not disclose my identity)

- I understand that the statistical data and verbal statements collected from this study may be published.

- ➢ I understand that the consent, data and prototype shown to me as part of this study is PROPRIETATRY and CONFIDENTIAL and as such that I will not divulge to anyone any information regarding what I have seen or discussed during this study. This will apply from when I sign this consent form up to and including the date of the publication of the results of this study.

□ I HAVE CAREFULLY STUDIED THE ABOVE AND UNDERSTAND THIS AGREEMENT. I FREELY CONSENT AND VOLUNTARY AGREE TO PARTICIPATE IN THIS STUDY.

Form 1: Profile Questionnaire

A. Please provide the following information:

Name:

Gender:

- Female
- Male

Age group:

Education level:

- Under diploma
- Some years of college
- College degree
- University degree

B:

Test type1:

Start time:
End time:

Test type2:

Start time:
End time:

Introduction for test Type 1:

You are now going to be asked some questions.

Please choose freely only one of the options provided as the answer to each question.

Thank you!

Introduction for test Type 2:

Now you are going through another series of questions. Please choose A, B, C or D as the best answer based on your knowledge. Thank you!

Form 2: Satisfaction Questionnaire

This questionnaire is aimed to represent how you feel about your numeracy skill assessment you performed today by two different methods.

Please mark the choice that most clearly expresses how you feel about a particular statement. Please provide us with any comments you have regarding each question.

A. Thinking about your experience with **TEST1** method, to what extent do you agree with the following statements?

	Strongly agree	Agree	N/A	Disagree	Strongly disagree
The whole test was a pleasant experience to me.	□ 1	□ 2	□ 3	□ 4	□ 5
I felt comfortable going through the sequence of the questions in the test.	□ 1	□ 2	□ 3	□ 4	□ 5
It was easy to understand the questions.	□ 1	□ 2	□ 3	□ 4	□ 5
I trust the result of the test.	□ 1	□ 2	□ 3	□ 4	□ 5

B. Thinking about your experience with **TEST2** method, to what extent do you agree with the following statements?

	Strongly agree	Agree	N/A	Disagree	Strongly disagree
The whole test was a pleasant experience to me.	□ 1	□ 2	□ 3	□ 4	□ 5
I felt comfortable going through the sequence of the questions in the test.	□ 1	□ 2	□ 3	□ 4	□ 5
It was easy to understand the questions.	□ 1	□ 2	□ 3	□ 4	□ 5
I trust the result of the test.	□ 1	□ 2	□ 3	□ 4	□ 5

C. Personally, on the result of which method you prefer to have your numeracy skill assessed?

- **TEST1** □
- **TEST2** □

D. Comments:

Form 3: Emotional response Questionnaire

*(I)Please read the following sentences and rate each sentence based on a 1 to 6 scale (1 representing the weakest and 6 representing the strongest for**Test1**.*

A. **I enjoyed answering the questions.**

1 2 3 4 5 6

B. **The questions in the study were annoying.**

1 2 3 4 5 6

C. **The questions in the study made me feel stressed.**

1 2 3 4 5 6

D. **The questions in the study were frustrating.**

1 2 3 4 5 6

*(II)Please read the following sentences and rate each sentence based on a 1 to 6 scale(1 representing the weakest and 6 representing the strongest for**Test2**.*

A. **I enjoyed answering the questions.**

1 2 3 4 5 6

B. **The questions in the study were annoying.**

1 2 3 4 5 6

C. **The questions in the study made me feel stressed.**

1 2 3 4 5 6

D. **The questions in the study were frustrating.**

1 2 3 4 5 6

End of the Session:

Thank you for participating in this study.

Your participation and feedback is greatly appreciated and very valuable to the goals of this study.

Please note that the contents of this study are confidential and that you should not discuss anything that you saw here today.

Thank you, once again, for your participation.

Appendix C: (Likpus Objective Numeracy Scale)

Imagine that we have a fair, 6-sided die (for example, from a board game or a casino craps table). Imagine we now roll it 1000 times. Out of 1000 rolls, how many times do you think the die would come up even (number 2, 4 or 6)?

In the Big Bucks Lottery, the chances of winning a $10.00 prize is 1%.What is the best guess about how many people would win a $10.00 prize if 1000 people each buying a single ticket to Big Bucks?

In the Acme publishing sweepstakes, the chance of winning a car is 1 in 1000. What percentage of tickets to Acme Publishing Sweepstakes wins a car?

Which of the following numbers represents the biggest risk of getting a disease?

1 in 100, 1 in 1000, 1 in 10

Which of the following numbers represents the biggest risk of getting a disease?

1%, 10%, 5%

If person A's risk of getting a disease is 1% in 10 years, and person B's risk is double that of A's, what is B's risk?

If person A's chance of getting a disease is 1 in 100in 10 years, and person B's risk is double that of A's, what is B's risk?

If the chance of getting a disease is 10%, how many people out of 100 would be expected to get the disease?

If the chance of getting a disease is 10%, how many people out of 1000 would be expected to get the disease?

If the chance of getting a disease is 20 out of 100, this would be the same as having a -----% chance of getting the disease.

The chance of getting a viral infection is 0.0005 out of 10,000 people, about how many of them are expected to get infected?

Appendix D: (Numi Questions)

(Numeracy Understanding in Medicine Instrument questions:

1. James has diabetes. His goal is to have his blood sugar between 80 and 150 in the morning. Which of the following blood sugar readings is within his goal?
    a. 55
    b. **140**
    c. 165
    d. 180
2. Nathan has a pain rating of 5 on a pain scale of 1 (no pain) to 10 (worst possible pain). One day later Nathan still has pain but it is better. Now, what pain rating might Nathan give?
    a. **3**
    b. 5
    c. 7
    d. 9
3. Natasha started a new medicine and was given a handout showing the chance that side effects will occur as in the table below. Which side effect is Natasha least likely to get?

Side Effect	**Chance of Occurring**
a Dizziness	1 in 5 people
b Nausea	1 in 10 people
c Stomach pain	1 in 100 people
d **Allergic reaction**	**1 in 200 people**

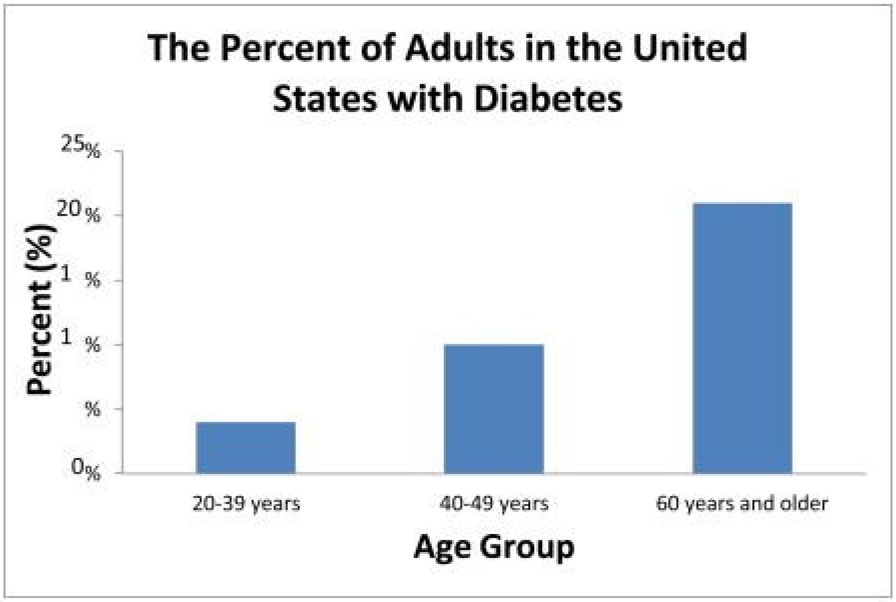
4.Frank has a test to look for blockages in the arteries of his heart. The doctor said that a person with a higher percent (%) blockage has a high chance of having a heart attack. Which percent (%) blockage has the highest chance of a heart attack?

a. 33%
b. 50%
c. 75%
d. **98%**

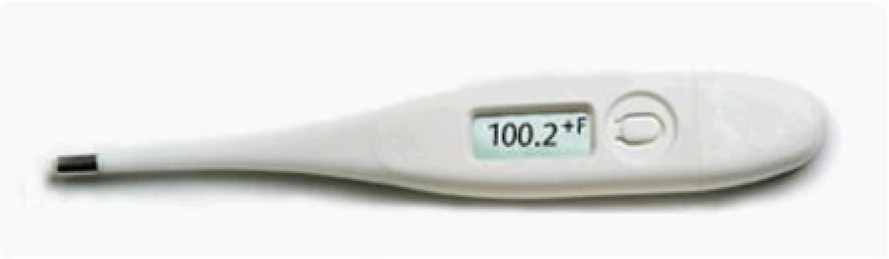
5. The doctor told Maria not to take more than 3 grams (g) of Tylenol a day. Each Tylenol pill is 500 milligrams (mg). What is the highest number of pills that Maria can take in one day?
    a. 3 pills
    b. **6 pills**
    c. 8 pills
    d. 12 pills
6. A medical study will randomly assign people so that people are equally likely to get medicine A or medicine B. If there are 300 people in the study, about how many are expected to get medicine A?
    a. 100 people
    b. **150 people**
    c. 200 people
    d. 250 people
7. David is 50 years old and smokes cigarettes. His doctor tells him that the chance of having a heart attack increases as people age and if they smoke. His current chance of a heart attack is 10% over the next 10 years. Which of the following is the best guess of David's chance of a heart attack in the next 20 years?
    a. 5%
    b. 10%
    c. **30%**
    d. 100%
8. James starts a new blood pressure medicine. The chance of a serious side effect is 0.5%. If 1000 people take this medicine, about how many would be expected to have a serious side effect?
    a. 1 person
    b. **5 people**
    c. 50 people
    d. 500 people
9. The PSA (prostate specific antigen) is a blood test that looks for prostate cancer. The test has false alarms so about 30% of men who have an abnormal test turn out not to have prostate cancer. John had an abnormal test. What is the chance that John has prostate cancer?
    a. 0%
    b. 30%
    c. **70%**
    d. 100%
10. Rebecca was treated for stage 2 breast cancer. The chance that the breast cancer will come back is 10% over the next 10 years. If Rebecca takes a new medicine, this chance will decrease by about 30%. Out of 100 women like Rebecca who take the medicine, how many will have breast cancer come back within 10 years?
    a. 3 out of 100 women
    b. **7 out of 100 women**

   c. 10 out of 100 women
   d. 30 out of 100 women

11. A study found that chemotherapy decreased the risk of dying from colon cancer by about 30%. The study was 95% sure that the real benefit was between 10% and 50%. Which of the following is not in the expected range of benefit?
    a. 11% decrease in risk
    b. 30% decrease in risk
    c. 45% decrease in risk
    d. **95% decrease in risk**
12. A study in arthritis patients found that medicine A decreased arthritis pain 10% more often than medicine B. The difference was not statistically significant. Which of the following best describes these results?
    a. **Medicine A and medicine B work equally well**
    b. Medicine A is proven to be better than medicine B
    c. Medicine B is proven to be better than medicine A
13. A study found that a new diabetes medicine led to control of blood sugar in 8% more patients than the old medicine. This difference was statistically significant (p=0.05). The likelihood that this finding was due to chance alone is:
    a. 1 in 5
    b. 1 in 10
    c. 1 in 15
    d. **1 in 20**
14. In general, the results of a randomized controlled trial will be more reliable if a larger number of people are in the study.
    a. **True**
    b. False
15. A survey asked a group of people about their exercise habits and followed them; over time. The study found that those who exercised 3 times a week or more lived an average of 2 years longer than those who did not. What did this study show?
    a. Exercising causes people to live longer
    b. **There is a relationship between exercising and living longer**
16. According to the graph below, what percent (%) of adults in the 40–49 year old age group have diabetes?
    a. 5%
    b. **10%**
    c. 15%
    d. 20%



17. John had a fever. The doctor told him to come to the hospital if his temperature was above 102.5 F. Otherwise, John should take Tylenol and rest. If John's temperature is as shown in the picture below, what should John do?
    a. **Take Tylenol and rest**
    b. Go to the hospital



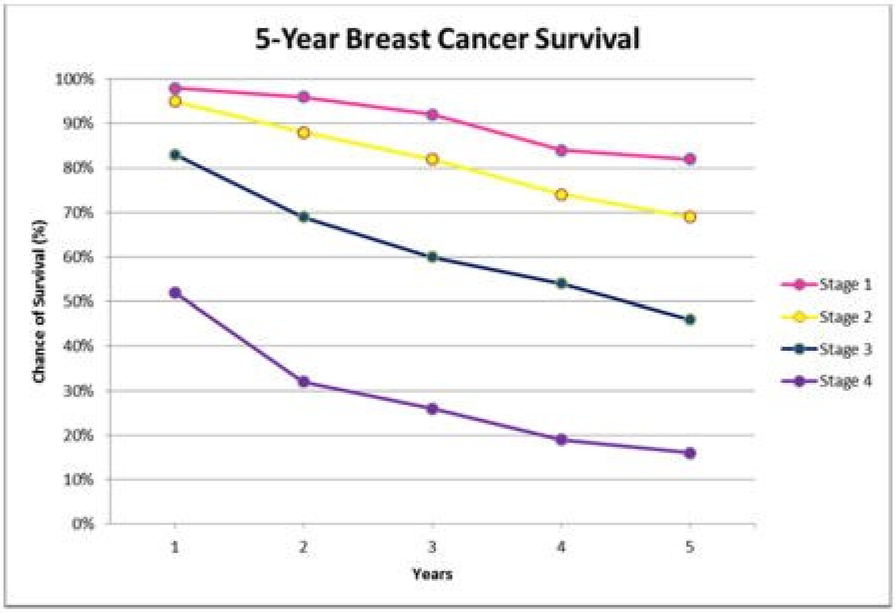
18. A nutrition label is shown below. How many calories did Mary eat if she had 2 cups of food?
    a. 140 calories
    b. 280 calories
    c. **560 calories**
    d. 680 calories

Nutrition Facts
Serving Size 1 cup (228g)
Servings per Container 2
Amount Per Serving

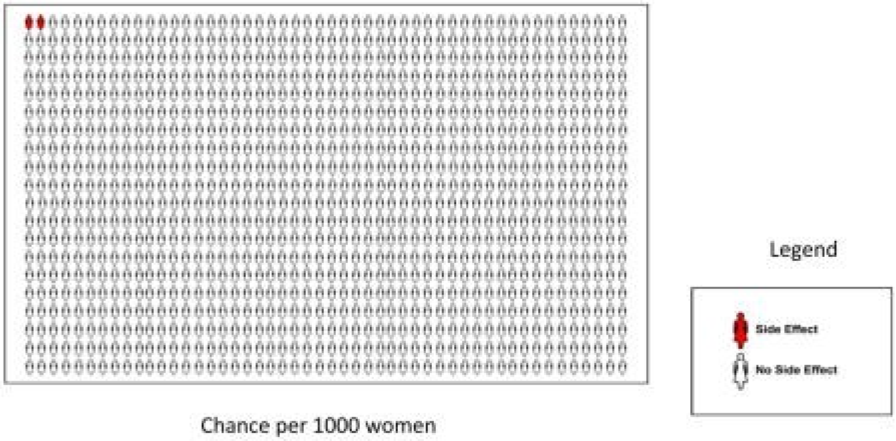
Calories 280	**Calories from Fat** 120
	% Daily Value[*]
Total Fat 13g	20%
Saturated Fat 5g	25%
Trans Fat 2g	
Cholesterol 2mg	10%
Sodium 660 mg	28%
Total Carbohydrate 31g	10%
Dietary Fiber 3g	
Sugars 5g	
Protein 5g	
Vitamin A 4%	Vitamin C 2%
Calcium 15%	Iron 4%

f. [*]Percent Daily Values are based on a 2,000-calorie diet. Your Daily values may be higher or lower depending on your calorie needs.

19. The graph below shows the outcomes of a group of women diagnosed with breast cancer. Andrea has stage 2 breast cancer. According to the graph, what is her chance of surviving 3 years after diagnosis?
   a. 56%
   b. **82%**
   c. 92%
   d. 100%



20. Carol is taking a new medicine. The chance of a side effect is very small as shown in the graph below. What number best shows her chance of having a side effect?
    a. 0.0002
    b. **0.002**
    c. 0.02
    d. 0.20



Chance per 1000 women

Appendix E: (Table of Abbreviations)

Abbreviation	Name
ARR	Absolute Risk Reduction
BMI	Body Mass Index
CAT	Computerized Adaptive Testing
CBL	Confidence Based Learning
C-PNA	Confidence-based Patient Numeracy Assessment
COMBO	Combination
DL	Difficulty Level
GQM	Goal Question Metric Model
GQ(I)M	Goal Question Indicator Model
HCI	Human Computer Interface
Hyp0	Null Hypothesis
Hyp1	Alternative Hypothesis
HypA0	Accuracy Null Hypothesis
HypA1	Accuracy Alternative Hypothesis
HypCom0	Comfort Null Hypothesis
HypCom1	Comfort Alternative Hypothesis
HypDU0	Discretionary Usage Null Hypothesis
HypDU1	Discretionary Usage Alternative Hypothesis
HypE0	Effectiveness Null Hypothesis
HypE1	Effectiveness Alternative Hypothesis
HypPL0	Pleasure Null Hypothesis
HypPL1	Pleasure Alternative Hypothesis
HypPR0	Productivity Null Hypothesis
HypPR1	Productivity Alternative Hypothesis
HypS0	Satisfaction Null Hypothesis
HypS1	Satisfaction Alternative Hypothesis
HypT0	Trust Null Hypothesis
HypT1	Trust Alternative Hypothesis
HypTrust0	Trust Null Hypothesis (in comparison between two types)
HypTrust1	Trust Alternative Hypothesis (in comparison between two types)
HypU0	Understandability Null Hypothesis
HypU1	Understandability Alternative Hypothesis
HyUE0	Usage Efficiency Null Hypothesis
HypUE1	Usage Efficiency Alternative Hypothesis
IRT	Item Response Theory
MDIT	Medical Data Interpretation Test
MVC	Model-View-Controller

Abbreviation	**Name**
NC-PNA	Non Confidence-based Patient Numeracy Assessment
NNT	Numbers Needed to Treat
NUMi	Numeracy Understanding in Medicine Instrument
PPE	Patient Preference Elicitation
REALM	Rapid Estimate of Adult Literacy in Medicine
RCTs	Randomized Controlled Trials
RRT	Relative Risk Reduction
S-TOFHLA	Shortened version of TOFHLA
SG	Standard Gamble
TOFHLA	Test of Functional Health Literacy in Adults
TTO	Time Trade-Off
URL	Universal Resource Locator
WRAT-3	Wide Range Achievement Test