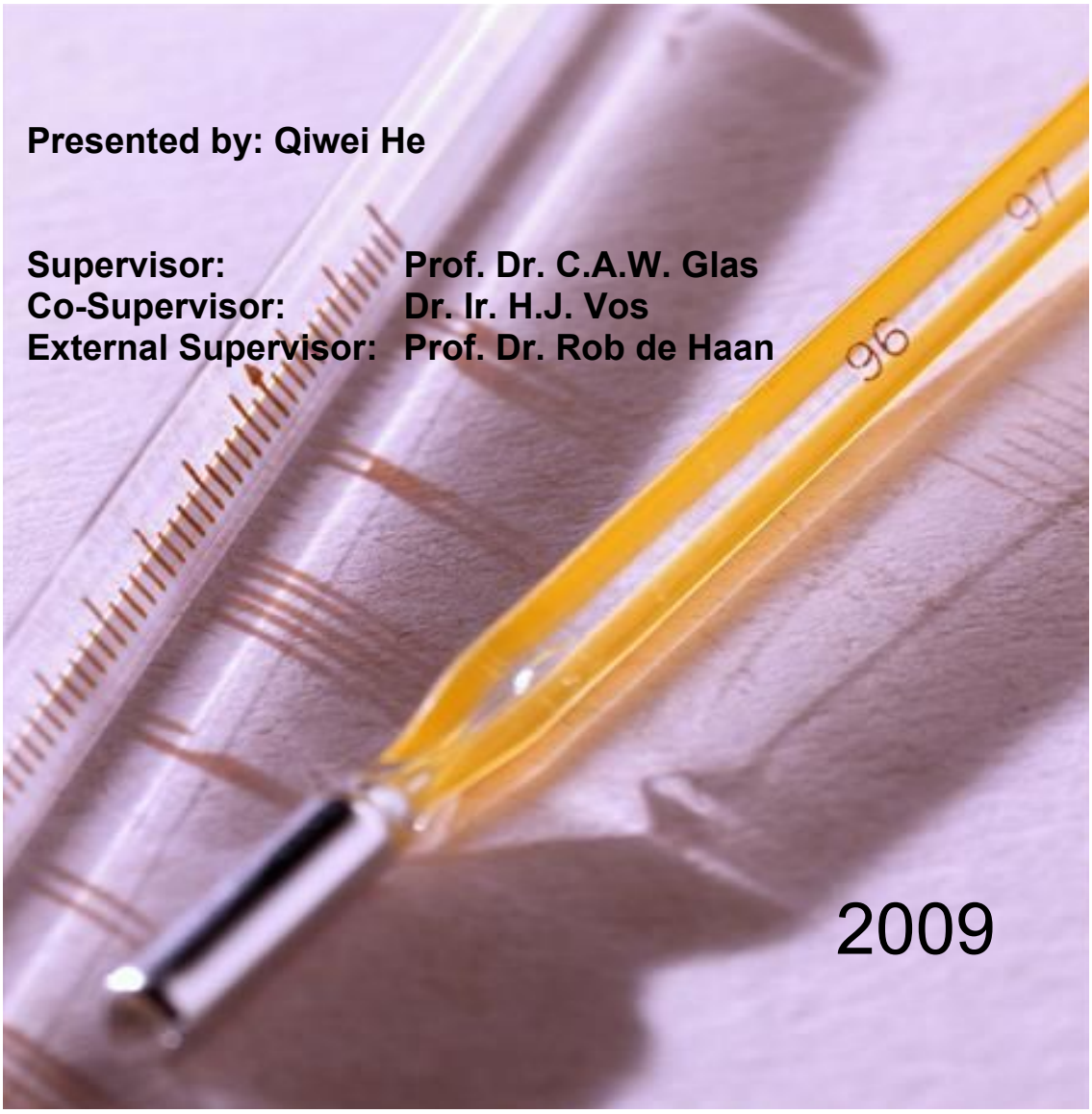
Twente University
Enschede – the Netherlands

INFORMATION ANALYSIS FOR WEBSITE OF AMC LINEAR DISABILITY SCORE



INFORMATION ANALYSIS FOR WEBSITE OF AMC LINEAR DISABILITY SCORE

BY

QIWEI HE

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE RESEARCH MASTER DEGREE PROGRAMME SOCIAL SYSTEM EVALUATION AND SURVEY RESEARCH

AT

DEPARTMENT OF RESEARCH METHODOLOGY, MEASUREMENT, AND DATA ANALYSIS

FACULTY OF BEHAVIORAL SCIENCE

UNIVERSITY OF TWENTE

2009

Acknowledgement

This thesis is the result of 18 months spent in the Master Program of Social System Evaluation and Survey Research of Twente University, including over three months enjoyable internship with the department of Clinical Epidemiology and Biostatistics at the Academic Medical Center of Amsterdam University. Many people have supported me in this period and I would like to thank some of them personally.

Firstly, I would like to thank my supervisors and co-supervisors. Prof. Cees Glas, I am amazed at your knowledge of item response theory and statistical methodology. Whenever I ask a question, you immediately pluck the appropriate book from the shelf. You are a really strict, creative and energetic professor. You keep me inspired to develop creative ideas in data analysis and encourage me to step into new field in the further studies. Dr. Hans Vos, you are the most enthusiastic teacher I have ever met. I am really glad that I have followed six courses from you in one short year. I enjoyed your active presentation and appreciated your prompt feedbacks on my questions every time. Prof. Rob de Haan, I am particularly grateful that you have taught me how to present complex statistical ideas in a clinical context. I have never thought that I could do something in medical science until you led me into the ALDS project.

I would like to thank Dr. Marcel Dijkgraaf, the senior biostatistics expert in AMC, and Ms. Nadine Fleitour, the senior research nurse in Neurology Department in AMC, for taking the time to arrange my internship and give me lots of advice on the ALDS project.

And I would also like to thank Dr. Bernard Veldkamp and Dr. Jean Paul Fox, the associate professors in the Department of Research Methodology, Measurement and Data Analysis (OMD), Faculty of Behavioral Science, University of Twente. I am really impressed your knowledge of the computerized adaptive testing and mathematical statistics.

Furthermore, I am very grateful for the help from the Management of the Faculty of Educational Sciences. Special thanks go to Ms. Dionysia Loman, Ms. Frances Leusink and Ms. Astrid ten Peze.

I would also like to thank my friends and classmates. It is you who make me feel at home and enjoy the happiness in studying. Maaike van Groen, I am really pleased to be classmate with you. You have given me a lot of advice in educational assessment studies and shared the best recipe of cooking witlof with me!

Finally, I would like to thank my family for their great support during the 18 months. I am especially grateful to my mother and father. Although they are now living is China, more than 7,000 kilometers away, their strength is always an inspiration to me. Weihua Zhou, my dearest husband, you are the most fascinating person I have ever met. Your encouragement, supports, patience, considerate care and great love make me brave to confront various challenges and ultimately realize my ambition! Thank you dear, love you for ever!

Contents

Abstract

Taking the advantages in flexibility and accuracy, item response theory (IRT) has been popularly used in diversified fields besides educational tests. The Academic Medical Center (AMC) Linear Disability Score (ALDS) project is an efficient application in clinical measurements with IRT models. In contrast to numerous multi-item questionnaire constructed by using the classical test theory, the ALDS measures at the item level by placing the patients' ability on the same linear scale with the item difficulty.

This thesis introduces the framework and current measurement procedures of ALDS and proposes to build a website for ALDS aiming to help the researchers and doctors use the ALDS instrument in a more accurate, flexible and efficient way. On the base of information analysis, the website work flows, embedded computerized adaptive testing program, simultaneous statistical tool for power analysis and the automatic item selection modules are stressed in the ALDS website framework.

Key words: item response theory, computerized adaptive testing, linear disability scores, power analysis, website construction

Chapter 1 Introduction

An estimated 10% of the world's population experience some form of disability. The number of people with disabilities is increasing due to population growth, aging, emergence of chronic diseases and medical advances that preserve and prolong life (World Health Organization, 2007). The severity of illness can be measured with a wide range of physiological parameters, for example blood tests and imaging techniques. However, these parameters cannot tell the whole story about how the disease process affects patient and their life. Thus, new instruments to describe the disease outcomes in a systematic and hieratical manner have been paid much attention in the recent decade.

So far, numerous generic and diseases-specific instruments measuring disability have been developed, such as SF-36 (Brazier, Roberts & Deverill, 2002) and SIP (Bergner, Bobbitt, Carter & Gilson, 1981). Most of these instruments are multi-item questionnaires constructed by classical test theory (CTT). In spite of the popularity, there are several problems associated with their use. Firstly, responses to all items on a scale are required to calculate at a sum score. The long questionnaires cost patients, clinicians and researchers a great amount of time to complete. Secondly, since sum scores are dependent on the items included in the instrument, it is difficult to compare scores from different instruments, even if they measure the same disability concept (Lindeboom, Vermeulen, Holman & de Haan, 2003).

1.1 Item Response Theory

In contrast to the CTT sum score methods, item response theory (IRT) measures at the item level. This means that disability status can be assessed in a much more flexible way and that each patient can be presented with a smaller selection of items than is possible using sum score based methods. IRT models are similar to logistic regression models. Using this approach it is possible to place items on a hierarchical difficulty with linear measurement properties. The units of the scale are the regression coefficients and are expressed in logits (Weisscher, 2008).

The Academic Medical Center (AMC) Linear Disability Score (ALDS) project was an good example in implementation of IRT in clinical measurement. This project, developed by the AMC of Amsterdam University in the Netherlands, calibrated 77 items in the item bank to be used in daily patient care and clinical research. In the development phase of the ALDS, data was collected from over 4000 disabled patients with a broad range of conditions including stroke, Parkinson's disease and chronic pain (Weisscher, 2008).

1.2 Purpose and Scope of the Study

Although a great number of researchers, clinicians and patients have get benefits from the ALDS measurement system, there is still some inconvenience in application. For example, some researchers complain that a lot of repetitive work has to be done in ALDS project. Little guidance could be found for constructing randomized controlled clinical trials (RCT) in the context of IRT. Researchers proposed questions such as "how many items should be included in the trial", "how many patients are required as

sample", "how much statistical power can the clinical trial acquire"…and so on. Moreover, nurses are also not satisfied to spend a lot of time in data collection for ALDS. They complain that some patients even become annoyed when being asked to answer "yes" or "no" to a too "easy" item. Therefore, for a further development in ALDS project, to construct a website with an embedded simultaneous statistical tool seems in an urgent necessity.

The information analysis of ALDS website lasts 6 months, from March to September in the year of 2008, organized by the Academic Medical Center of Amsterdam University and the University of Twente. The main purpose for this study is to do the preparation for ALDS website and design the work flows for a website-based computerized adaptive testing for ALDS.

The major tasks include: to collect information on ALDS current implementation procedures (e.g. item selection, questionnaire construction, latent variables estimation, and administration and monitoring of ALDS users), to design workflows of ALDS website, to investigate the feasibility of website-based CAT and to propose further scheme in improving accuracy and efficiency in ALDS application.

1.3 Outline

The main subject of this thesis is the information analysis for website of AMC Linear Disability Score project. In **Chapter 2**, the framework of ALDS is introduced. The introduction focuses on the ALDS item bank calibration, measurement procedures, as well as strong and weak points of ALDS application. In **Chapter 3**, a construction plan for ALDS website is proposed. Besides the website workflows, the data flows, storage and safety and maintenance are also discussed in this part. **Chapter 4** investigates the feasibility of computerized adaptive testing in ALDS website and illustrates how the CAT approaches in the program. For a better guidance for the clinicians and researchers in constructing clinical trials with ALDS, **Chapter 5** proposes a statistical tool for power analysis, which will help the users select optimal items and define sample size within the IRT framework. Finally, **Chapter 6** presents a conclusion on the ALDS website project and makes suggestions for the future research.

In addition, three appendixes are attached in this thesis. **Appendix A** records the recent modified 73-item ALDS item bank. In **Appendix B**, the work flows for ALDS website are exhibited. The work flows are divided into four parts: registration page, patient page, assistant page and researcher page. **Appendix C** is a practice in setting performance standards for brain stroke patients by both ALDS and clinical instruments. The categorized standards of ALDS and mRs[1] parameters are successfully linked to each other in this study, which helps the clinicians who are not familiar with ALDS make a relatively accurate estimate on patients' ability when they try this new IRT instrument.

[1] mRs, modified Rankin scale, is a concise index of global disability, ranging from 0 (no symptom) to 6 (dead).

Chapter 2 AMC Linear Disability Score (ALDS) Project

2.1 Brief Introduction

Taking the advantages in flexibility and accuracy, item response theory (IRT) has been popularly used in diversified fields besides educational tests. The Academic Medical Center (AMC) Linear Disability Score (ALDS) project is an efficient application in clinical measurements with IRT models. ALDS was created at the beginning of 2000's aiming to construct an item bank regarding daily activities to measure the disability status of patients with a broad range of diseases. Nowadays, it has been used as a basis for computerized adaptive and other innovative testing procedures to assess the functional status of patients in a wide variety of clinical studies (Holman, 2005; Weisscher, 2008).

In contrast to numerous multi-item questionnaire constructed by using the classical test theory (CTT) such as SF-36, the ALDS measures at the item level by placing the patients' ability on the same linear scale with the item difficulty. This means that disability status can be assessed in a much more flexible way and that each patient can be presented a smaller selection of items than is possible using sum score based methods. The adaptive testing procedures is implemented as more difficult items (e.g. "bike for two hours") are presented to less disabled patients while the easier items (e.g. "put on an T-shirt") to more severely disabled patients.

Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific functional health instruments and supplemented by diaries of activities performed by healthy groups (Holman, 2005). The item bank has been calibrated by using the responses from over 4,000 patients with a broad range of stable chronic conditions. A total of 196 items were identified as clinically applicable items and then described in detail at the initial stage (Holman, 2005). But only 77 items are remained currently as applicable ones to be commonly used in measurements, with the range of difficulty level from -3.49 to +3.05 (Weisscher, 2008).

Patients in ALDS project are asked whether they can, rather than do, carry out the activities. The ALDS uses dichotomous frame at present, the two response options are "I can carry out the activity" and "I cannot carry out the activity"[2]. If patients had never had the opportunity to experience an activity, the response of "not applicable"[3] is recorded.

The two-parameter logistic IRT model (2PL) was fitted collected data in the calibration phase of the ALDS project.

$$P_{ik}(\theta_{ik}) = \frac{e^{a_i(\theta_k - b_i)}}{1 + e^{a_i(\theta_k - b_i)}} \tag{2.1}$$

[2] ALDS actually has three response options, "I can", "I can but with difficulty" and "I cannot". For a simple statistical calculation, the two positive options "I can" and "I can but with difficulty" are combined as one option "I can".

[3] Responses in the category "not applicable" are regarded as missing data, which are statistically treated as if the items had not been presented to the individual respondent.

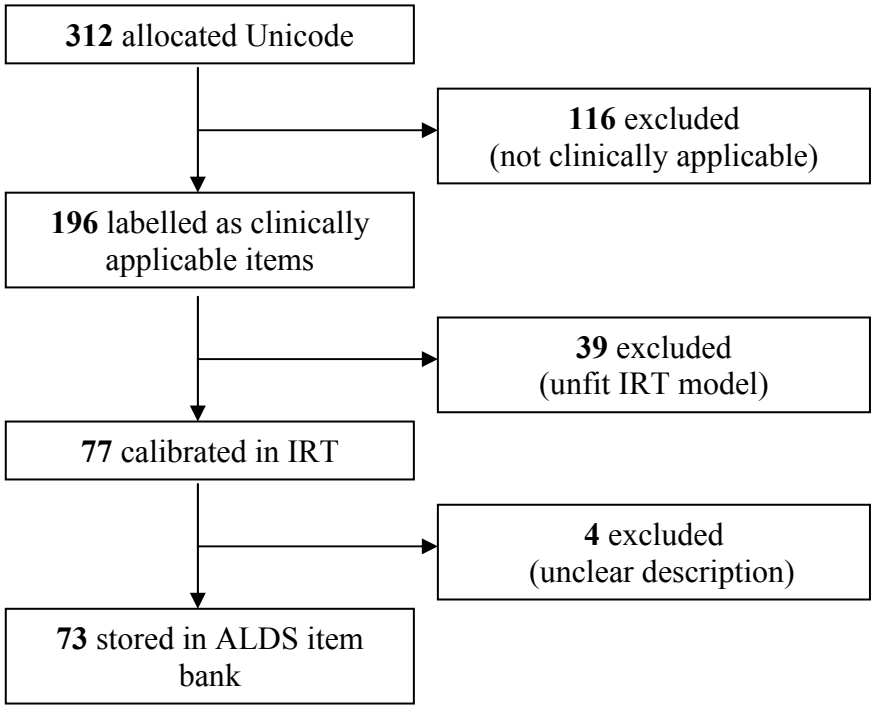
This model was chosen because it allows a more realistic model for the data to be built than the more restrictive one-parameter logistic model. In 2PL model, both of the item difficulty and discrimination degree are required to take into consideration in order to get the "best selected" items for different groups of patients. To make the results easier to interpret, the logit scores are linearly transformed into values between 0 (bottom value) and 100 (ceiling value) after the sum of probabilities that the patient (theta) can give correct answer to each item in the bank is derived (Weisscher, 2008).

2.2 ALDS Item Bank

Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific functional health instruments and supplemented by diaries of activities performed by healthy adults. A total of 312 items were listed in the initial item bank with Unicode allocated to each one, among which 196 items were identified as clinically applicable ones and then described in detail. For example, the item "shopping" was expanded to "travelling to the shopping center on foot, by car, by bike or by other public transport, walking around the shopping center, getting into a number of shops, trying on clothes or shoes, buying a number of articles, paying for the bill, and returning home". However, only 77 items from 196 were fit in two-parameter logistics model (2PL) and calibrated to have difficulty and discrimination item parameters.

It is interesting to find that 25 items are often selected by doctors in clinical studies. A few items are complained because of confusion or unclear explanation. For example, the item "can you wash your face" can be understood in two ways: go to the sink by the patient himself and wash face; or cannot go to the sink by the patient himself but can wash face if taken to the sink. Thus, four items with unclear definitions and low discrimination parameters were eliminated from the latest version of ALDS item bank. (The latest version of ALDS 73-item bank is attached in Appendix A.)

Figure 2.1 Summary of Calibration Process of ALDS Item Bank



The 77-item bank (item description and explanation) has been translated into English, Dutch, French and Chinese.

2.3 Application Scope of ALDS

ALDS is firstly developed from neurology studies with the aim to use the daily activity to describe symptom. Nowadays, ALDS has been widely spread to diversified clinical studies. According to Weisscher's report (2008), a total of 36 studies have been implemented via ALDS, among which 8 studies have been completed. Thousands of patients have responded to ALDS questionnaires so far. In addition, the application of ALDS is not only within the Netherlands, but also to other countries around Europe.

ALDS is usually used to trace the progress of patients by comparing their previous and present status. On account that the clinical measurement instruments such as Rankin and mRs are often used to categorize patients, to have a linkage between the ALDS score and clinical measurement results seems also important for the further development of ALDS.

2.4 Current Measurement Procedures of ALDS

The current measurement procedures of ALDS can be divided into four steps: item selection, data collection, data analysis and results output and storage.

2.4.1 Item Selection

Item selection is the initial preparation for the adaptive measurement procedure. Due to the limited application of computerized adaptive testing in clinical trials, the ALDS item selection depends on the group's average functional ability level, instead of individual level. For example, in the brain stroke trials, typically three booklets with various difficulty levels are in demand to testing patients in three functional statuses, less disabled, moderately disabled and severely disabled. The easiest booklet will be used in the severely disabled group while the most difficult booklet for the less disabled group.

The booklets are normally constructed by researchers and clinical staffs together. Two questions are always focused on: "which items should be included" and "in which level of booklets should the selected items be placed". The clinical staff (especially nurses/doctors in the related discipline) usually gives advices on item inclusion according to their clinical experience, while the researcher is responsible for considering the IRT statistical issues on the base of item parameters. For instance, items with the equivalent difficulty parameters around the patients' group level have to be selected more than those located far away. And the researcher also has to consider that if only one booklet is in need in certain disease, the difficulty parameters of selected items should be spread out.

Generally speaking, 20 to 25 items are included in each booklet. 5 to 10 items are usually contained in each booklet of one study as the common items. Different sets of booklets are used for different diseases. The booklets in various diseases are seldom copied from each other. However, if one set of booklets is well constructed in the same

disease studies, it will be kept for repetitive implementation or readjusted by a small scale.

Items with the following characteristics are favored in selection:

(1) Items described in short term but with precise meaning;
(2) Items highly correlated with the diseases of the study;
(3) Items with little potential bias (differential item functioning, DIF) in gender, age, living conditions and etc.

2.4.2 Data Collection

Data collection is usually implemented by trained nurses via phone, mails or face-to-face interviews. During the interviews, nurses are required to ask items to patients and record answers by ticking the corresponding box. (e.g. “I can”, “I can but with help”, “I cannot”, and “not applicable”)

Although the items listed on booklets rank in ascending difficulty order, the nurses prefer to reorganize items by their contents to make the communication with patients more reasonable. For example, the items about actions, such as walking, jogging and cycling, are categorized in one group, while items about washing, such as showering and washing face are categorized in another group. The item difficulty parameters are not taken into account in the reorganization procedure.

Nurses often use different approaches to begin interviews when confronting patients with “zero” experience or “nonzero” experience in ALDS studies. As for the patients who are interviewed for the first time, nurses usually begin with greeting questions, e.g. “how are you these days” or basic functional items, e.g. “can you walk around or have to use a wheelchair”. Nurses are able to get a rough impression on the patients’ current disability status to a large extent according to their clinical experience and then decide which level of the booklet should be used for this specific patient. For instance, suppose there are 11 booklets constructed for one study, ranking in ascending difficulty order. When interviewing a patient for the first time, nurses typically start from the booklet 5 or 6. After asking 3 to 4 items in this booklet, nurses can have a further judgment on the patient’s status, and hence decide to change a harder or easier booklet in necessity.

As for the patients who have been interviewed before, nurses have known some information about them, so the question as comparison with previous performance, for instance, “do you feel better these days?” or “how are you going these days, better or worse?” will be used as the beginning. After getting patients’ responses, nurses will decide to use more difficult or easier items than those they had used in the previous time.

Most nurses usually keep on asking all the items to patients although they have known the answers. Depending on nurses’ clinical experience, if a patient can give positive answer to a more difficult answer, he can surely positively respond to an easier item in the same item group. Hence, some nurses sometimes may skip the easier items in this situation.

During the phone interview, nurses always ask to talk with the patients themselves. The patients' relatives or helpers are interviewed on behalf of patients only when patients are not able to answer the phone (e.g. deaf) by themselves. If the phone is answered by the patient himself, nurses will skip asking the item "can you answer the phone" if it is listed in the booklet. If the phone is not directly answered by the patient, the nurse with ask to change the phone-picker to the patient.

The options "can (with difficulty)" and "cannot" are well distinguished. When the patient is able to perform the activity independently, without any help from anybody else, but aids or devices are allowed, the response "can" is recorded. If a person is physically not able to perform an activity, needs help from somebody else or if the symptoms would be badly increased without others' help, the response should be "cannot", because the patient cannot finish the activity independently.

Nurses would like to control the answer "I don't know" as low as possible because "not applicable" option will be calculated as missing data. In order to reduce the missing data rate in interview, nurses often add explanations on the items that patients give responses as "I don't know". For example, when patient says "I don't know" to the item "vacuum a flight of stairs" (maybe the patient do not use vacuum or do not have stairs at home), the nurse will let the patients imagine the situation and give a relative accurate response.

Different studies have different follow-up plans on patients. Typically, one to three times of follow-up interviews are implemented within one year. The same booklets are usually used in the follow-up studies. To avoid psychological impact on patients' emotions, all the ALDS scores are kept blind and confidential to the patients. Only the patient PIN code and date of birth are shown on booklets and recognized by the computer.

A majority of patients are glad to answer ALDS items because they would like to feel concerned by hospital and have a good opportunity to communicate with nurses. But sometimes patients may boast or overreport their performance as if they were better than before. On the contrary, a few patients may regard too easy items as an "insult", hence they would like to underreport, refuse answering items or even give converse answers rather than telling the truth.

Different approaches in data collection may generate measurement bias, which can be analyzed in four aspects:

(1) Patients will get more information in face-to-face and phone interview than mail questionnaire. When patients are not clear about the items, the interview in the first two modes can make some further explanations, while the mail questionnaire cannot get such help.
(2) During the interviews via face-to-face and phone, patients usually do not give direct response "yes" or "no" on the items, but make lots of comments, which needs the nurses to deduce the final answer by their subjective judgment. However, the patients have to give direct answers in the mail questionnaire.
(3) During the phone interview and mail questionnaire, it is hard for nurses to judge whether the patient tells the true current disability status; while the face-to-face is easier to observe whether the patient is "lying".

(4) Interviews via phone and face-to-face may be influenced by the nurses' tones or expressions, while the mail questionnaire avoid this problem.

In addition, different explanations provided by various interviewers may also generate bias. For example, in the description of item "can you get in/out of a car" should include the action of fastening seat belt, but some nurses forget to emphasize this point. Another example is that some nurses even replaced the item "go shopping" by the item "go to post office" when they get the "not applicable" responses from male patients in "shopping" item. But actually these two items have big difference in item parameters.

Therefore, the bias in data collection process should arouse the researchers' attention. Although a periodic item review has been implemented, the problem is still deserved more efforts in further research.

2.4.3 Data Analysis

Data analysis is implemented with SPSS and BILOG[4]. The procedures can be divided into five steps: recode Unicode, recode category labels, theta estimation in BILOG, multiple imputations from posterior distribution of theta-hat with standard error and linear transfer to ALDS score.

In the first step, raw data are input to computer by typing or scanning via computer recognition software. On account that only 77 from 312 items in the bank have been calibrated in IRT models, the Unicode has to be recoded to avoid the confusion in the further analysis. The recode process is implemented in SPSS, that is, to find the corresponding Unicode for items selected in booklets in the item bank. Items selected in the booklets will show the patients' responses "0" or "1", representing "cannot" and "can" respectively, the items not selected in the booklets are systematically labeled "9". The recoded Unicode is saved into two separate files with the extension filename *.sav* and *.dat* to be recognized by SPSS and BILOG respectively.

Secondly, category labels recoding is a big problem in previous ALDS analysis. Because the category label scoring was not unique, different studies adopted various labels for scales, which caused lots of confusion and mixture. For example, in the neurology studies, the categories were labeled as 1, 2, 3 and 4 for "I can" "I can but with difficulty", "I cannot" and "not applicable"; while in other studies, the categories were labeled as 0, 1, 2 and 3 for "I cannot", "I can", "I can with difficulty" and "not applicable" respectively. Thus, the researchers had to recode category labels every time to keep a consistent standard in theta estimation. At present, the category scoring has been unified for all of the studies as 1, 2, 3 and 4 for "I can" "I can but with difficulty", "I cannot" and "not applicable". Consequently, this step can be omitted in the future studies.

Thirdly, BILOG read in the newly generated *.dat* file in the first step and estimate theta by using formula (2.1) on the base of fixed item parameters stored in BILOG. The estimated theta-hat and its standard error are calculated by maximum likelihood

[4] BILOG implements an extension of item response theory. It has diversified functions in multistage analysis, especially in deriving theta and item calibration. Maximum likelihood method is majorly used in this program.

method and the results are saved into an ASCII file with the extension filename *.sco*, which can be directly recognized by SPSS.

Fourthly, SPSS reads in the estimated theta-hat ($\hat{\theta}$) and measurement standard error from *.sco* file generated in the third step. Because the theta-hat ($\hat{\theta}$) is an estimator of the true value, theta, the measurement error is unavoidable. In order to minimize the measurement error, multiple imputations drawn from the posterior distributions of the latent variable, theta-hat ($\hat{\theta}$) is randomly implemented by the "SEED". But one problem in SPSS is that the "SEED" changes every time in execution. The estimated theta value for each patient in the same trial cannot be fixed, that is, the estimated theta-hat in the first execution is totally different from the second one, all the analysis on the theta-hat generated from the first execution will be in vain if a second execution has to be made. This problem has aroused high attention, which need to be solved in an urgent way.

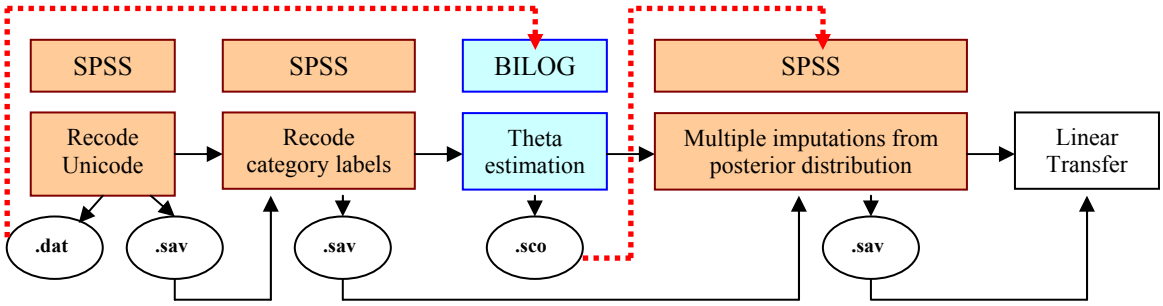
Finally, to make the estimation results easier to interpret, the logit scores are linearly transformed into values between 0 (bottom value) and 100 (ceiling value) after the sum of probabilities that the patient (theta) can give correct answer to each item in the bank is derived (Weisscher, 2008).

$$\hat{T} = \sum_{n=1}^{77} P_n = \frac{e^{a_1(\theta - b_1)}}{1 + e^{a_1(\theta - b_1)}} + \frac{e^{a_2(\theta - b_2)}}{1 + e^{a_2(\theta - b_2)}} + \ldots + \frac{e^{a_n(\theta - b_n)}}{1 + e^{a_n(\theta - b_n)}} \tag{2.2}$$

$$ALDS = 10 + \frac{80}{77} \cdot \hat{T} \tag{2.3}$$

where $\hat{T}$ is the best estimate of the sum of probabilities that a certain patient can give positive response to each item in the bank. The ALDS score in the 77-item bank ranges from 11 to 89.

Figure 2.2 Current Data Analysis Procedures of ALDS



2.4.4 Data Output and Storage

The output results are saved in SPSS, listing only three columns: patient ID, theta-hat and transferred ALDS score. The researcher is responsible to send the results back to doctors and nurses.

It is a pity that no database has been constructed yet to store the data collected from various studies. Datasets are spread in the hands of ALDS users, only part of which are stored in AMC.

2.5 Comments on ALDS

On the base of information collected from patients, nurses, physicians and researchers, the ALDS project is concluded as "a highly welcomed scoring system but with a couple of inconvenience". Its strong points and weak points can be summarized as following several aspects.

Its strong points consist of:

(1) ALDS is very practical and easy to make communications with patients.
(2) ALDS is not very technical in implementation, thus it is easily accepted and welcomed by patients.
(3) In contrast to medical measurement, ALDS focuses on patients' daily activities, helping patients to describe symptoms.
(4) ALDS stands on a new orientation to make clinical analysis, providing supplementary analysis for clinical information and patient-related outcomes.

Meanwhile, besides the advantages, ALDS project has also some inconvenience and problems in application. Here come its weak points:

(1) The booklets contain too many items (approximately 20 to 30 items). It is time consuming especially when the item is not selected in an adaptive way. For instance, very easy items have to be asked to the less disabled patients. Actually, these relatively easy items can provide little information on the less disabled patients' functional status.
(2) Some items are not described in a clear way, which causes unnecessary misunderstandings in interviews and generates bias in data collection process.
(3) Some items have different senses in various situations, e.g. in a telephone interview, nurses will never ask the item "can you pick up the telephone" although it is listed in the booklet. But this item is sensible in the face-to-face interview.
(4) Item selection process depends on researchers' judgment and clinical staffs' medical experience to a large extent, instead of patients' performance. The arbitrariness may cause the errors in measurement.
(5) Researchers complain that lots of repetitive calculations have to do in data analysis process. They would like to focus on their interesting studies instead of doing all the calculations.
(6) Complicated software problem. Most of the ALDS users, except special researchers are not familiar with BILOG. The uncommonly used program prevents more users from implementing ALDS project.
(7) Few researches have been done in linking the ALDS scoring system with the clinical measurement. It is hard to interpret the ALDS result comparing with the widely used clinical scales such as Rankin and mRs.

Chapter 3 ALDS Website Construction

As stated earlier, ALDS plays an important role in many aspects, but still needs improvement to be perfect from many aspects. With the development of network, it would be a good idea to solve the problems of ALDS by the intervention of computer programs. Thus, constructing a specific website for ALDS appears necessary and urgent. In this chapter, the proposal of an ALDS website construction will be discussed.

3.1 Objectives and Feasibility

The construction of ALDS website aims to help the researchers and doctors use the ALDS instrument in a more accurate, flexible and efficient way. Firstly, unlike the current measurement procedures, the ALDS website will provide a statistical tool for researchers following an automatic computerized item selection and store all the databases on ALDS in the server. Secondly, the website is flexible to the specialists to create their own clinical trials with the on-line statistical guidance. In addition, for the convenience of users, this website can offer both computerized adaptive testing (CAT) and pencil-and-paper (PP) versions of ALDS booklets. Thirdly, this website is built as a platform for ALDS users to exchange their ideas and propose problems they have met during the application of ALDS. ALDS instrument will be promoted through the website and attract more researchers to have a try. It will relax researchers and clinical staff from repetitive calculations. And with the integration of CAT module, the website may save 50 percent of the time and items compared with the traditional way. To avoid the complexity of statistical programs, the BILOG will not be used in the website, but the function of theta estimation will be embedded as a module into the network framework.

The website construction is also feasible in the electronic era. Network and computer is very popular nowadays. Although it is hard for every senior people to possess a computer at home, this website will be welcomed by the young and middle-aged patients. For the convenience of patients, a computer can be installed at the waiting room. And the patients who are not able to answer the items at home can be informed to the hospital 10 minutes earlier and finish the item booklet on the computer with the help of nurses. Besides, because the pencil-and-paper is also available in the website, the doctor can choose to use the PP version for the senior patients.

3.2 Website Framework

To build up the ALDS website is a complex task. In order to clarify the functions of different modules, an ID Code identification system is designed to distinguish the users as patients, assistants and researchers. (The workflow of the whole website framework is attached in Appendix B.)

3.2.1 Registration Page

The layout of the website is divided into three blocks by different identity properties. ALDS users are asked to fill in personal information and email address for the first time. The system then judges the identity of the user, patient, assistant or

researcher, and save the user's information on the server. Meanwhile, the system will allocate an ID code for the user according to his identity. For example, P12345 is identified as a patient, A12345 as an assistant and R12345 as a researcher. After the registration, the website ID code and password will be sent to the user by email immediately.

After the registration for the first time, the user may change his password when logging into ALDS website. The old users who have already got the ID code and password can directly go to the LOGIN page. After inputting ID code and password, the user will get into the page special for his identity, that is, different ID code will lead into different webpage for the user. For example, the patient with ID code P12345 will directly get into the patient page, ID code A12345 will directly get into the assistant page and ID code R12345 will automatically get into the Researcher page.

3.2.2 Patient Page

The patient page has the simplest layout, which provides a platform for patients to practice the computerized adaptive testing in ALDS. The patient can get into the patient page after inputting his ID code and password. Then the system asks the patient to type in the code number for required test that has been informed by the doctor assistant via email. According to the patients' gender and age, the items will be preselected, that is, the items with DIF for the specific patient will not be shown. For example, the item favored by the female, such as items regarding shopping will not be shown if the patient is a male.

After patient getting into the CAT page, the system starts the computerized adaptive testing program immediately. Generally speaking, the first item displayed on the screen is often selected by collateral information, that is, to pick out an item that best fits the patient's background with medium difficulty to start the computerized adaptive testing. From the second item, the computer will automatically select items according to the patient's response to the previous one. Following the instruction on screen, the patient is required to click on the answer and then continue to the next item. If the patient wants to quit the process of CAT, he can click the button of "cancel". No results will be saved in this situation. CAT will stop when the standard error becomes below a certain threshold or the maximum of items has been reached. (The stopping rule of CAT program is explained in details in Chapter 4.) The system will save the patient's results into database under the code number of sub-item bank and email the data to the specified nurse or researcher automatically. But the patients cannot see their results to avoid unnecessary psychological impacts.

3.2.3 Assistant Page

Assistant page is more complicated than the patient one. As a helper for the doctor, the assistant can see the items in the sub-item bank and the summary report under the code number. After getting into the assistant page, the assistant is required to type in the code number for required test. Then three options will be displayed on the screen: to get into CAT test or PP test, show items in the sub-item bank under the code number, and show the summary report in *.xls* file stored on the website under the code number.

Sometimes the assistant or nurse has to help the patients who are not able to implement the CAT test. (For example, the senior people may have problems in using

computer or websites.) The assistant can ask the patient's ID code and get into the patient page as if he/she were the patient. Then the assistant can read the items on screen and click on the options responded by the patient during the interviews. This function can also be used when the patient has to do the CAT test in the waiting room.

Furthermore, the assistant can also choose the pencil-and-paper version in necessity. The assistant can print out the sub-item bank in difficulty order or random order, collect data from patients via interviews and input the data into database under the code number on the website.

3.2.4 Researcher Page

Researcher page is the most complicated page that includes all the functions of ALDS website. After inputting the ID code and password, the researcher loges into the researcher page where four options are shown: full item selection process, severity disability category, function category and existed studies.

In the full item selection approach, a thermometer-shaped score range with typical item description every 10 items will help users locate the level of target population. The researcher is asked to input the range of their patients, suppose from 30 to 50 as ALDS score, then the system will default the normal distribution and calculate the mean value and standard deviation for the group, and transfer it into IRT terminology, theta. (The statistical tool for item selection process is introduced in details in 3.3)

The severity disability category and function category are both based on the previous studies. According to the distribution of patients in previous studies, we can set them as the default values, and offer researchers the categories. For example, in severity disability category, full range, high severity, medium severity and low severity will be listed. The researcher can make a peer item selection on the base of the selected items by the system. And some restrictions will be set, for instance, at least 25 items have to be included in the booklet, some items with the essential accuracy cannot be deleted. The full range will be directly linked to the automatic item selection part.

Like the severity disability category, the function category with the sub-item banks also needs to be constructed in advance. Then the researcher can select which function he would like. For instance, finger function, lower body function and etc.

After the item selection process, a new sub-item will be generated and system will allocate a systematic code number for this new sub-item bank. Afterwards, the researcher can also select to use the CAT or pencil-and-paper version. If CAT is selected, the code number of the test will be directly sent to the patient. If the researcher prefers the PP version, then the sub-item bank will be printed in difficulty order or in random order. The researcher can also select both of CAT and PP versions. After data collection from patients, the raw scores can be input into the database under the sub-item bank code number. If the CAT version is selected, the system will directly get the results and save it into the database. But if the PP version is selected, the procedure will be a little bit more complicated. The input data will be saved in *.dat* file, including the raw data and patient' pin code. The category scores are labeled as cannot-0, 1-can, 2-can with difficulty, 9-not applicable, 8-systematic missing data (item does not selected). Then the options "cannot" and "not applicable" will be

recoded into 0, “can” and “can but with difficulty” into 1, the systematic missing data will be remained as 8.

A program of theta estimation is recommended to embed in website by using computer language, such as FORTRAN. Similar as BILOG-MG 3, the program reads in the response file *.dat* (data collected from patients) and the item parameter estimate file. The Maximum Likelihood (ML) method is adopted. ML estimates are computed by the Newton-Raphson method starting from a linear transformation of the logit of the percent-correct score for the subject. In those rare cases where the Newton iterations diverge, an interval-bisection method is substituted. Estimates for respondents with all correct or all incorrect responses are attributed by the half-item rule. That is, respondents who score all incorrect are assigned one-half a correct response to the easiest item; respondents who score all correct are assigned one-half a correct response to the hardest item. The estimate is then computed from this modified response pattern. Standard errors are computed as the square root of the negative reciprocal of the expected second derivative of the log likelihood at the estimate, i.e., square root of the reciprocal Fisher information (Zimowski, Muraki, Mislevy & Bock, 2003).

After the theta is estimated, random draw from the posterior distribution of the measurement error should be done. The linear transform is made to get ALDS score at the last step. Afterwards, all the results will be saved as a summary report in *.xls* on the website under the code number. The newly generated file can be downloaded by the researcher.

3.3 Transforming ALDS Scores into IRT Terminology

In common sense, the most important transformation is from θ scale into true-score scale. Let X, the number-right score, be defined as

$$X = \sum_{j=1}^{n} U_j \tag{3.1}$$

where U_j is the 1 or 0 response of an examinee to item j. If the true score is denoted as τ, then

$$\tau = E(X) = E(\sum_{j=1}^{n} U_j) = \sum_{j=1}^{n} E(U_j) \tag{3.2}$$

Since U_j takes on value 1 with probability $P_j(\theta)$ and value 0 with probability $Q_j(\theta) = 1 - P_j(\theta)$, it follows that

$$E(U_j) = 1 \cdot P_j(\theta) + 0 \cdot Q_j(\theta) = P_j(\theta) \tag{3.3}$$

Thus,

$$\tau = \sum_{j=1}^{n} E(U_j) = \sum_{j=1}^{n} P_j(\theta) \tag{3.4}$$

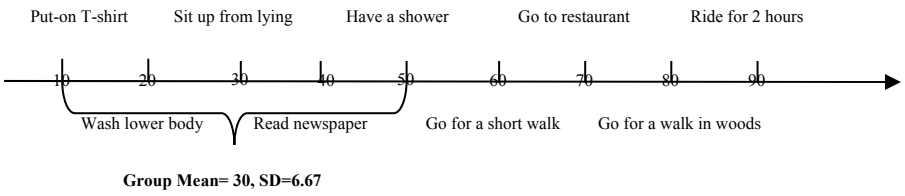
that is the true score of an examinee with ability θ is the sum of the item characteristic curves. The true score, τ called the test characteristic curve because it is the sum of the item characteristic curves (Hambleton, 1991).

The true score τ and θ are monotonically related; that is, the true score may be considered to be a nonlinear transformation of θ. Since $P_j(\theta)$ is between 0 and 1, τ is between 0 and n. Hence, τ is on the same scale as the number-right score, except that τ can assume non-integer as well as integer values. The transformation from θ to τ is useful in reporting the ability values; instead of the θ values, τ values that lie in the range 0 to n are reported (Hambleton, 1991).

However, in the ALDS project, doctors may not be familiar with statistics, especially IRT models, thus we have to converse the above process, that is to make a transformation from the ALDS score (linearized true score) into θ scale.

As Figure 3.1 exhibits, the 1-100 true score scale is easier for doctors to locate patients' ability status. The data of reference population previously collected in various branches of diseases are expected to set as the default values for statistical tool. For instance, in brain stroke studies, we may set the default values of group mean ranging from 30 to 50 in ALDS scores. Doctors can either choose this default value or adjust as his requirement, for example, to enlarge the range as 10 to 50. Then the tool will automatically calculate the ALDS mean value as 30 with standard deviation as 6.67, defaulting the population following in a normal distribution.

Figure 3.1 Input ALDS Score into Statistical Tool



Because of the linear association between ALDS score and the sum of probabilities that the patient can give positive responses to each item, we can use the algorithms in (2.2) and (2.3) to calculate the theta.

Through Maximum Likelihood Method and Newton Raphson's Method (van der Linden & Glas, 2000), the estimated ability of patient, theta-hat ($\hat{\theta}$), in IRT models, positioning at the same scale as item difficulty can be derived.

$$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)} \tag{3.5}$$

$$f(\theta_n) = \hat{T} - \sum_{i=1}^{k} \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} \tag{3.6}$$

It is worthy to notice that the starting value of Newton Raphon's method is one of the most essential steps in estimation of theta. Typically, all the items are supposed to be the same with the unique difficulty value as 0 and discrimination as 1. For the model in (3.5), the initial value θ_0 is expected to:

$$T = \sum_{i}^{k} \frac{e^{\theta}}{1+e^{\theta}} = \frac{K \cdot e^{\theta}}{1+e^{\theta}} = KP(\theta) \tag{3.7}$$

$$K - T = K - \frac{K \cdot e^{\theta}}{1+e^{\theta}} = K \cdot \frac{1}{1+e^{\theta}} = \frac{K}{1+e^{\theta}} = KQ(\theta) \tag{3.8}$$

$$\frac{T}{K-T} = \frac{KP(\theta)}{KQ(\theta)} = e^{\theta} \tag{3.9}$$

$$\log(\frac{T}{K-T}) = \log e^{\theta} = \theta_0 \tag{3.10}$$

where K indicates the total number of items in the bank; T is the sum of probabilities that the patient gives positive responses to each item with the same difficulty value 0 and discrimination value 1.

3.4 Data Flows

The response raw scores can be automatically saved as a *.dat* file in the CAT program as the input data for the website. However, if the pencil-and-paper version is adopted, the raw scores have to be typed or scanned into the computer first and then generate the *.dat* file. The *.dat* file is easily recognized as ASCII. After the calculations through the website, the output file will be produced also in an ASCII file, which can be read by the widely-used software MICROSOFT EXCEL. The contents of output file include: pin code of patient, theta (after random draw from posterior distribution), raw data for each item, total number of correctness and ALDS score.

3.5 Data Storage

It is wise to temporarily store all the datasets (e.g. at least six months after the researcher finish his project with ALDS) on the server. If the external researchers agree that their datasets can be used by the ALDS project team, then the dataset can be permanently stored in the server; otherwise, the dataset will be cleared after the temporary storage period. According to the time that datasets begin to store on the server, we can grasp the process of researcher's studies with ALDS. For their convenience, it seems also good to remind the researchers keep on follow-up study by email.

3.6 Safety and Maintenance

It is very technical issue to keep the security of the network. Firewall should be taken into consideration. All the users need to fill in the authentic email address, which

is the unique channel to get the website ID code and download the data results and sub-item bank.

It is proposed to have certain person keep the website updated. A platform will be built up for users to ask questions and get instructions from the ALDS experts.

Meanwhile, some comparisons with other health-care websites have been made. The SF-36 is a multi-purpose, short-form health survey with only 36 questions (Brazier et al., 2002). It yields an 8-scale profile of functional health and well-being scores as well as psychometrically-based physical and mental health summary measures and a preference-based health utility index (Nichol, Sengupta & Globe, 2001). The website, Sf-36 (http://www.sf-36.org/) with a very simple and clear style is unfortunately not based on IRT techniques whose statistical tool is totally different from ALDS.

Patient Reported Outcomes Measurement Information System (PROMIS) in the United States shares the similar goals as ALDS, but its scope is much larger and has been constructing a global network currently. Its objectives are to develop and test a large bank of items measuring patient-reported outcomes, create a computerized adaptive testing system that allows for efficient, psychometrically robust assessment of patient-reported outcomes in clinical trial research involving a wide range of chronic diseases, and create a publicly available system that can be added to and modified periodically and that allows clinical researchers to access a common repository of items and computerized adaptive tests. The website of PROMIS (http://www.nihpromis.org/default.aspx) offers many an instrument based on IRT methods. It features in simultaneous statistical tools design and Q&A forum construction, which could be learned for ALDS website.

Chapter 4 Computerized Adaptive Tests in ALDS

4.1 Computerized Adaptive Tests

In principle, tests have always been constructed to meet the requirements of the test-givers and the expected performance-levels of the examinees as a group. It has always been recognized that giving a test that is much too easy for the examinees is likely to be a waste of time. On the other hand, questions that are much too hard, also produce generally uninformative test results, because examinees cease to seriously attempt to answer the questions, resorting to guessing, response sets and other forms of unwanted behavior (Linacre, 2000).

It is a really tough problem in the classical test theory. As Hambleton said, any single test administered to a group of examinees could not provide the same precision of measurement for every examinee. Thus, the ideal testing situation would be to give every examinee a test that is "tailored", or adapted, to the examinee's ability level (Hambleton, Swaminathan, & Rogers, 1991). That is the reason that the "adaptive testing" aroused people's attention.

Weiss (1985) defined "an adaptive test is one in which different sets of test questions (items) are administered to different individuals depending on each individual's status on the trait being measured". The earliest application of tailored or adaptive testing was in the work of Binet on intelligence testing in 1908 (Weiss, 1985). However, little additional work on the adaptive testing took place until the advent of computers.

As stated in chapter 1, item response models are particularly suitable for adaptive testing because it is possible to obtain ability estimates that are independent of the particular set of test items administered. In fact, adaptive testing would not be feasible without item response theory. Even though each examinee receives a different set of items, differing in difficulty, item response theory provides a framework for comparing the ability estimates of different examinees (Hambleton et al., 1991).

In computerized adaptive testing (CAT), the sequence of items administered to an examinee depends on the examinee's performance on earlier items in the test. The values of item probability and item information indicate how well an item differentiates between contiguous ability levels or how precisely the item measures at that point on the ability scale (Weiss, 1985). Based on the examinee's prior performance, items that are maximally informative about the examinee's ability level are administered. In this way, tests may be shortened without any loss of measurement precision. High-ability examinees do not need to be administered relatively easy items, and low-ability examinees do not need to be administered the most difficult items, because such items provide little or no information about the examinee's ability (Hambleton et al., 1991).

Two procedures are used currently for item selection in an adaptive mode (Kingsbury & Zara, 1989). The first, maximum information (Weiss, 1982), involves the selection of an item that provides maximum information (i.e., minimized the standard error) at the examinee's ability level. The second method, Bayesian item

selection (Owen, 1975), involves the selection of the test items that minimizes the variance of the posterior distribution of the examinee's ability. Bayesian methods require specification of a prior belief about the examinee's ability; hence, the success of the method depends in part on the appropriateness of the prior distribution. The impact of the prior distribution diminishes as more items are administered. The most apparent difference between these two methods is that: maximum likelihood estimation poses problems when the number of test items is small. Bayesian procedures overcome the problems encountered with maximum likelihood procedures but may produce biased estimates of ability if inappropriate prior distributions are chosen (Hambleton et al., 1991).

The stopping rule of CAT was summarized by Linacre in 2000. There are three possibilities to stop the CAT besides the standard error method. First, the item bank is exhausted. This occurs, generally with small item banks, when every item has been administered to the examinee, the test has to be stopped. Secondly, the maximum test length is reached. There are a pre-set maximum number of items that are allowed to be administered to the examinees. This is usually the same number of items as on the equivalent paper-and-pencil test. Thirdly, as Weiss's mentioned, the ability measure is estimated with sufficient precision. Each response provides more statistical information about the ability measure, increasing its precision by decreasing its standard error of measurement. When the measure is precise enough, test stops automatically. The procedure can be stopped when the standard error of the examinee's ability estimate stops decreasing by a specified amount or reached below a certain threshold. Actually, we use a mixture of these stopping rules nowadays. For example, both of the precision measurement and maximum test length can be set before the test. The CAT will stop automatically when either of these criteria is reached. Meanwhile, a larger item pool is recommended to build up to avoid the first stopping rule used.

The typical stopping rule for standard error threshold is .32, assuming the test reliability is around 90% (Linacre, 2000). This concept comes from classical test theory, in which the standard error is the square root of 1 minus reliability.

$$SE = \sqrt{1-\rho} = \sqrt{1-.90} \approx .32 \tag{4.1}$$

In the IRT theory, the amount of information provided by a test at θ is inversely related to the precision with which ability is estimated at point. The test information equals to the sum of item information.

$$SE = \frac{1}{\sqrt{I}} \tag{4.2}$$

$$I = \sum_{i=1}^{n} I_i \tag{4.3}$$

Thus, the standard error .32, i.e. reliability 90% can be interpreted as approximate 10 items each with 1.0 item information in the Rasch model.

In 2PL model, the item information defined as

$$I = a^2 P(\theta)Q(\theta) \tag{4.4}$$

Then the standard error .32 can be interpreted as 40 items have to be included in the test if the discrimination parameter equals to 1 and patient ability equals to the difficulty parameter in the 2PL-model. If we denote the item numbers as K,

$$\because a = 1, \theta = b$$
$$\therefore I = a^2 P(\theta)Q(\theta) = K \cdot 1 \cdot .50 \cdot .50 = 10$$
$$\therefore K = 40$$

Obviously, if we increase the discrimination parameter ($a > 1$), then less items are required to use in order to get the reliability 90%; however, if the discrimination parameter is even lower than 1, then more items have to be used to keep at the same level of standard error.

In addition, Stoop added an idea in 2001 regarding mixing fixed and variable length test. He reckoned that all kind of mixtures between fixed-length and variable-length can be used in CAT. An advantage of the variable-length rule is that the measurement precision is guaranteed for each examinee, which may not always be the case when the fixed-length rule is used (Stoop, 2001).

4.2 Rationale of CAT in ALDS

As stated earlier, the ALDS project follows the IRT framework, which satisfies the premise of computerized adaptive testing. In addition to shortening tests without loss of measurement precision, the benefits of computerized adaptive testing in ALDS are numerous.

First, following the CAT procedures, the item selection will no longer be arbitrarily undertaken by researchers or clinical staffs but by the patients' responses to the items. What the researchers and physicians need to do is to build up a big item bank for various diseases. The CAT program successively selects questions so as to maximize the precision of the exam based on what is known about the patient from previous questions. From the patient's perspective, the difficulty of the exam seems to tailor itself to their level of function status. For example, if the patient performs well on an item of intermediate difficulty, he will then be presented with a more difficult question. Or, if he performed poorly, he would be presented with a simpler question. Compared to the traditional questionnaire with a fixed set of items, computer-adaptive tests require fewer test items to arrive at equally accurate scores.

Furthermore, the "cheating" problem is also an ignorable issue in interviews with patients. CAT can help doctors find the "authentic" results though some patients are willing to boast themselves sometimes. For example, the patients who are cheating may have to answer more questions than those who do not cheat, in order to get a relatively precise evaluation and minimize the standard errors.

Thirdly, on the base of item response theory, the CAT can further relax doctors and nurses from the repetitive work. To be honest, although the patients have been categorized into different levels in ALDS, e.g. less disabled, medium disabled and

severely disabled, the booklets that the project is now using still depends on the group ability mean instead of the patient individual ability. Thus, to a large extent, it is not the authentic adaptive testing. When following the CAT, doctors and researchers do not need to repetitively design tests, input patients' data or calculate the ALDS score by themselves. The program will simultaneously do these all. Because of this advantage, the CAT is easily to trace the patients' recovery progress, which will be convenient for doctors to build the profile for each patient.

Fourthly, on account that power analysis on clinical trials mostly depends on the doctors' prediction on patients' ability, the wrong prior estimation is an inevitable problem in RCTs' construction process. We can imagine if doctor has a wrong prior estimation on effect size, the mean value of alternative mean group will be predictable incorrect. As a result, the item selection for the alternative group will not be as accurate as expectation. However, the doctors' clinical experience and their estimation are less vital factors in the CAT. The reason is that the computerized adaptive test calculates the individual's ability based on the patient personal responses to each optimal item. Even though the doctors' estimates deviate a bit, the result from CAT will still produce a relatively accurate estimate on the patient's ability.

Finally, as the computer era has been coming, the computerized adaptive testing is well feasible for patients to take at home, at office or at the waiting rooms of hospital. The module of CAT program is also not difficult to embed into the ALDS website if it is written into the standard web language, e.g. PHP.

4.3 CAT Approach

The CAT program is installed as a statistical tool in the ALDS website. As introduced earlier in chapter 3, the CAT program starts up after the patient or assistant type the test code number into the system.

The program will read in two input files: the parameters file and response file. The former will be fixed in the program because of the stable item parameters of the ALDS at present. The response file records the patient's responses (raw scores) to the items in his own computerized adaptive testing.

CAT data is input by the patient to answer each item automatically selected by the system. The item selection follows maximum information procedure because of a small item bank consisting of 77 items as a total. The ALDS CAT program offers three options of stopping rule for the users: stopping when the maximum number of items is reached (default at 25), or when the gap between the current standard error compared with the previous stage is within a certain value (default at 0.01), or reach below a specified threshold of standard error (default at .32). The users can change the default value as they like, and can adopt the three stopping rules at the same time. The computer will automatically stop when any of the criteria is reached. The output file is directly saved as *.xls* and saved into the database under the code number. The contents in output file can be defined as the same as the pencil-and-paper version, including the patient ID code, theta (after random draw from posterior distribution), ALDS score, raw score, number of correctness and etc. A pre-selection of items based on patients' gender and age will be made before the CAT starts. This process helps reduce the DIF error.

For highlighting the features of CAT ability estimation and item selection, an example coming from brain stroke sub-item bank is used for demonstration. For the purpose of the example, the item bank listed here only contains of 10 items (Table 4.1). (The complete sub-item bank (26 items)[5] for brain stroke studies has been attached in Appendix C.)

Table 4.1 Item Parameters of Sub-Item Bank of ALDS for Brain Stroke Studies

Item Content	*b*	*a*	ALDS
1. ride a bike for at least 2 hours	3.05	2.45	89
2. walk up a hill or high bridge	0.78	1.99	73
3. change the sheets on a bed	-0.21	1.56	58
4. walk up a flight of stairs	0.19	2.19	65
5. put long trousers on	-2.38	2.74	24
6. cut your toe nails	0.66	1.63	72
7. travel by local bus or tram	1.23	2.86	78
8. walk for more than 15 minutes	0.82	2.13	74
9. have a shower and wash your	-0.66	1.95	50
10. prepare breakfast or lunch	-1.52	2.27	36

Source: Weisscher, N. (2008). The AMC Linear Disability Score (ALDS): Measuring Disability in Clinical Studies. Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Suppose a female patient with the age of 60 takes this brain stroke exam, a sequence of events that might occur in computerized adaptive testing is as follows:

(1) Item 5 "put long trousers on" will be eliminated immediately in the pre-selection. The reason is that the old ladies in the Netherlands seldom wear long trousers, but skirts or dresses. Considering the potential problem in DIF, for the female patient, the program will skip out Item 5. Hence, only 9 items remain in the item bank.

(2) Item 4 is selected; this item is of average difficulty (0.19) and high discrimination (2.19). Suppose the patient answers Item 4 correctly. A maximum likelihood estimate of ability cannot be obtained until the examinee has answered at least one item correctly and one item incorrectly. (Zero or perfect scores correspond to $-\infty$ and $+\infty$ ability estimates, respectively.)

(3) Another item is selected. Item 8 is chosen because it is more difficult ($b = 0.82$) than the previously administered item ($b = 0.19$). Suppose the patient correctly answers Item 8. Again, a maximum likelihood estimate of ability cannot be obtained.

[5] In practice, an item bank would consist of hundreds, and possibly thousands, of test items. The ALDS project only has 77 items calibrated. As for different diseases, doctors define the sub-item bank by selecting the items highly correlated with the disease discipline of their research. That is the reason why not many items are included in each sub-item bank.

(4) Item 7 is chosen next; it is more difficult than Item 4 and 8. Suppose the patient answers this item incorrectly. The patient's item response vector for the three items may be represented as (1, 1, 0). Through use of the maximum likelihood procedure for estimating ability with known item parameters, an ability estimate can be obtained ($\hat{\theta} = 1.07$). The test information for the three items at this ability level is $I(\hat{\theta} = 1.07) = 3.43$, and the corresponding standard error is $SE(\hat{\theta} = 0.54)$. These values appear in Table 4.2.

(5) Next, the information provided by each of the remaining items in the bank is computed at $\theta = 1.07$. These values are reported in Table 4.3. Item 6 is selected next because it provides the most information ($I = 2.64$) at $\theta = 1.07$. Suppose that Item 2 is administered and then is still answered incorrectly by the examinee. A new ability estimate is obtained for the response pattern (1, 1, 0, 0). The new ability estimate is $\hat{\theta} = 0.77$.

(6) Then the item information at $\theta = 0.77$ for the remaining items is computed. The process described above for administering an item, estimating ability, determining the information provided by unadministered items, and choosing an item to be administered next based on the information it provides is continued. To continue this procedure, Item 6 is chosen next, following by Item 3, and then finally Item 9. The procedure stops when the standard error of the patient's ability estimate stops decreasing by 0.01. As can be seen from Table 4.2, the decrease in the standard error when Item 9 is administered in stage 7 compared with the standard error at stage 6 is 0.01. The procedure stops at this point. The estimate of the patient's ability is taken as $\hat{\theta} = 0.62$.

Figure 4.1 shows the trend of maximum likelihood ability estimates and standard error for this patient. It is apparently to see that the theta starts from a very sharp decrease from the first point to the third point and keep slight waves around 0.6. Similar status could also be found in the standard error curve. The standard errors keep decreasing from 0.54 to 0.43.

Table 4.2 Maximum Likelihood Ability Estimates and Standard Error for One Patient at the End of Each Stage in ALDS CAT Program

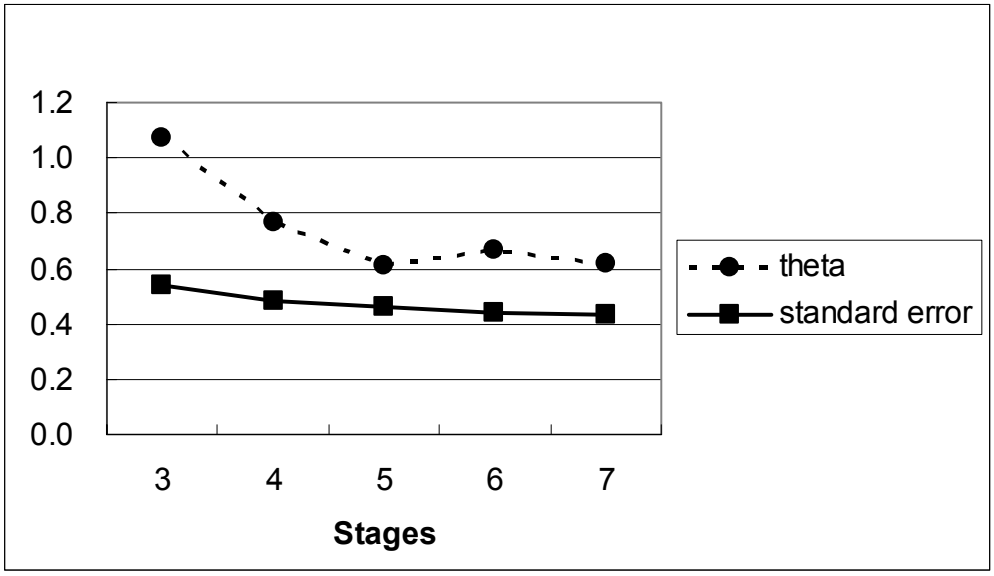
Stage	Item Number	Item Response	$\hat{\theta}$	$I(\hat{\theta})$	$SE(\hat{\theta})$
1	4	1	-	-	-
2	8	1	-	-	-
3	7	0	1.07	3.43	0.54
4	2	0	0.77	4.34	0.48
5	6	0	0.61	4.73	0.46
6	3	1	0.67	5.17	0.44
7	9	1	0.62	5.41	0.43

Note: The ability estimate calculation based on the IRT program: Baker F. B. (1998) The Basics of Item Response Theory Windows Version 1.0

Table 4.3 Information Provided by Unadministered Items at Each CAT Stage

Stage	$\hat{\theta}$	Information provided by Item									
		1	2	3	4	5	6	7	8	9	10
4	1.07	0.13	**2.64**	0.74	-	-	1.72	-	-	0.35	0.04
5	0.77	0.06	-	1.03	-	-	**1.90**	-	-	0.60	0.08
6	0.61	0.04	-	**1.20**	-	-	-	-	-	0.79	0.12
7	0.67	0.05	-	-	-	-	-	-	-	**0.71**	0.10
8	0.62	0.04	-	-	-	-	-	-	-		0.11

Figure 4.1 Maximum Likelihood Ability Estimates and Standard Error for One Patient in ALDS CAT Program



Source: Data in Table 4.2 Column 4 and 6.

4.4 Problems of CAT Application in ALDS

Although the computerized adaptive testing can function well in improving the efficiency and accuracy, it still has some problems in application. For example, some patients think the phone call interviews appear much friendlier than the CAT. Patients would like to communicate with nurses and doctors to get the feeling that they are well concerned, but the computerized testing is really not good at this point. Some patients even think the CAT is just as "teacher check pupils' performance", thus they don't think it is an appropriate way to express their needs to the doctors.

Furthermore, the CAT program will bring the financial problem. In order to implement the CAT, the Health Department, hospital, insurance company and maybe patients have to input money for the computers, website construction, program design and maintenance.

In addition, because of the practical problems, the CAT and pencil-and-paper versions have to be parallel used in the transition period, which may generate the measurement bias. Because the CAT version is designed on individual level, not all the

items written in PP version could be administered for an individual patient. It implies that the items that are not administered for the certain patient should be recorded as missing data corresponding to the PP version. However, unlike the CAT procedure, the PP version follows the classical testing theory that is based on a group level. This essential difference between these two versions makes their results incomparable. But fortunately the measurement bias will not be substantial if the test is made to check the mean value of a group instead of an individual person. And group-based approach is commonly used in clinical trials.

Chapter 5 A System of Power Analysis for Constructing Clinical Trials in ALDS

5.1 Background

Typically, in clinical trials, it is important to ensure that enough patients are included to have a reasonable power of detecting the effect of interests. The higher statistical power, the more certainty holds on effect size. In order to avoid a "bad case", doctors usually estimate the potential power before the trial gets into practice. Four ways are generally used to increase the power: 1) Increasing significance value α . 2) Enlarging the distance between two group means, namely expand the expectation on effect size. 3) Increasing the sample size. More data will provide more information at the group mean thus values of μ is possibly better distinguishable. 4) Decreasing the standard deviation of samples (Moore & McCabe, 2003). It has the same effect as the third.

However, if we put the classical power analysis into IRT models, another important factor, the number of items could not be neglected. Since IRT offers a framework, in which the number of items used to assess patients can be easily varied for different disability status, the power analyses need to consider not only the number of patients, but also the number of items used.

Holman, Glas and de Haan (2003) once investigated sample size calculations in RCTs and the relationship between statistical power and item selection procedures. They examined the power in a two-legged trial with the two-parameter logistics model (2PL) as measurement model and concluded that the number of patients in each arm of a randomized trial required to detect effect sizes varies with the number of items used. Furthermore, Holman (2004) also found that if a selection of items suitable for the population was used, fewer patients are required than if a selection spanning the whole item bank. Glas, Geerlings, van de Laar and Taal (2008) made an extension work of longitudinal RCTs in IRT after Holman. They found that on the base of two-step maximum likelihood estimation, using multiple imputations drawn from the posterior distributions of the latent variables could solve the problem of estimation error. (The outcome variables in MML2 are not direct observations but estimates with an estimation error.) Their research showed that in applications where the number of respondents was relatively small (which is usually the case in clinical trials) the power of hypothesis testing using plausible value imputation was larger than the power of MML2.

These conclusions certainly laid good foundations for further studies. But it is a pity that none of them has provided an accurate and simultaneous calculation method for the number of items, sample size, power and effect size. And even less attention has been paid to the impact of group's standard deviation on the number of items and the measurement in the ceiling and bottom groups. To design a simultaneous statistical tool for power analysis in ALDS is a real challenge, but it is worth trying not only for facilitating doctors using this good instrument but also for helping patients enjoy the benefits from the advanced techniques in psychometrics.

5.2 Objectives

The idea to design a simultaneous statistical tool in power analysis for ALDS was generated in the interviews with doctors and nurses who used ALDS in their daily work. The power analysis system aims to:

(1) Construct a complete framework in power analysis to guide the ALDS users construct single or successive clinical trials.
(2) Design a simultaneous statistical tool for ALDS, providing concurrent data analysis regarding power, number of items, sample size, effect size and etc.
(3) Find the potential rules between the number of items and sample size by using the methods of item information in IRT.
(4) Realize automatic optimal item selection procedure in ALDS project instead of arbitrary selection in a manual way.
(5) Find a solution to measure the functional status of patients with ceiling or bottom ability level.

5.3 Methodology and Approach

5.3.1 Effect Size for Hypothesized and Alternative Group

In a straightforward randomized clinical trial (RCT), the patient sample is randomly divided into two equally sized groups. The members of each group receive a different treatment regime and all outcomes are assessed at the end of the study. Clinically, the two groups are said to be different if the ratio of the difference between the mean health status in the two groups and the standard deviation of health status in the population under consideration is larger than a given treatment effect size (Holman, 2000). Typically, the treatment sizes are arbitrarily defined as 0.2, 0.5, and 0.8, indicating minimal, moderate and substantial effect respectively (Cohen, 1988).

In each RCT, the functional status of the patients in the control (hypothesized) group is usually simulated using the mean and the standard deviation observed in the original data. The functional status of patients in the treatment (alternative) group is simulated using the same standard deviation but with the mean equal to the original mean plus a treatment effect determined by the effect size under investigation.

$$\text{Hypothesized group distribution: } (\mu_0, \sigma)$$
$$\text{Alternative group distribution: } (\mu_A, \sigma)$$

where μ_0 is the hypothesized group mean, and μ_A is the alternative group mean. If the effect size is denoted as $\hat{\delta}$, then $\mu_A = \mu_0 + \hat{\delta}$.

The special case occurs when there is no effect size ($\hat{\delta} = 0$), namely, $\mu_A = \mu_0$. The hypothesized and alternative groups can be assumed as totally overlapped. Power analysis in the two groups transfers to item information analysis in this case, which will be further discussed in the section 5.3.7.

5.3.2 Power Analysis and Item Information Analysis

In classical statistics, the number of patients required to detect a given effect size with a particular power depends on the values of the effect size, the standard errors of the estimates of mean health in the two groups and the significance level used (Cohen, 1988). However, in IRT models, the number of items is also an essential factor to take into consideration. The reason is that in IRT models, the patients' functional status is on the same scale with item difficulty. Easy items will be adaptively allocated to more disabled patients, while the hard ones will be allocated to less disabled patients. The optimal item selection not only reduces the number of items but also gathers more information around the patients' ability. Thus, in IRT, it is possible to detect the effect size with the same power as classical statistics but with smaller sample size and less number of items.

The results suggested in Holman's investigation (2005) are too conservative and inapplicable, quite large sample size has to be used if the number of items is reduced. The reasons are probably from two aspects: firstly, in her simulation studies, item selection procedure had many limitations instead of following adaptive way according to patients' ability data. Secondly, less attention was paid to the item information analysis, which is a unique concept in IRT to evaluate how well the item can explains the targeted ability.

For a more precise measurement on power analysis in IRT tests, the item information analysis method is to be integrated into power analysis, which is expected as a new approach in IRT research.

Item information is an essential indicator in IRT to check the "power" of an item. The higher information, the better job this item can do, and the more power can be guaranteed to detect the patient's functional status. $I_i(\theta)$ is the information provided by item i at θ. In the maximum likelihood estimation, the test information is the simple sum of information of all items.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{5.1}$$

Item information functions can play an important role in test development and item evaluation in that they display the contribution items make to ability estimation at points along the ability continuum. This contribution depends to a great extent on an item's discriminating power, the higher a, the steeper the slope of P_i, and the location at which this contribution will be realized is dependent on the item's difficulty.

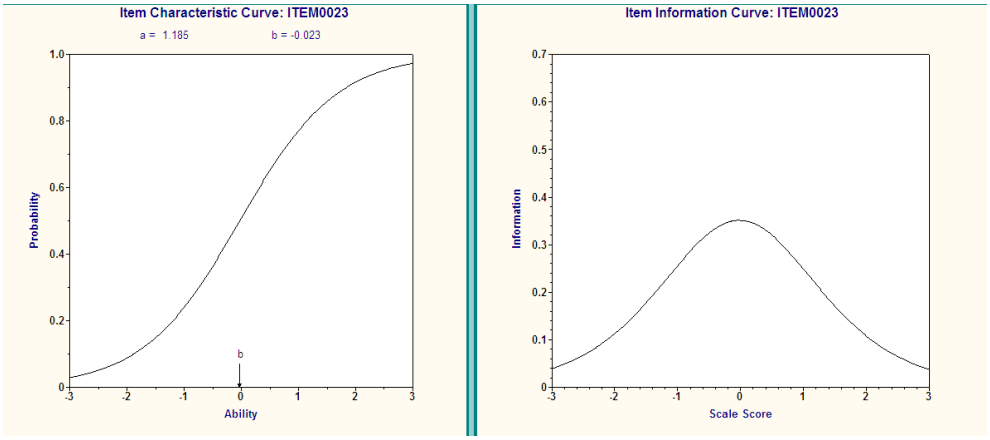
Let's take Figure 5.1 as an example. It is obvious to find that the first item plays a better role in measuring the patient with ability close to 0 because of its high information (0.38) around the scale score of 0. On the contrary, the second item has a relatively flat item characteristic curve due to less discrimination value. As a result, its information curve is as low as 0.1, suggesting that this item cannot provide sufficient information for the patient whose ability is expected around 0.

The amount of information provided by a test at θ is inversely related to the precision with which ability is estimated at that point:

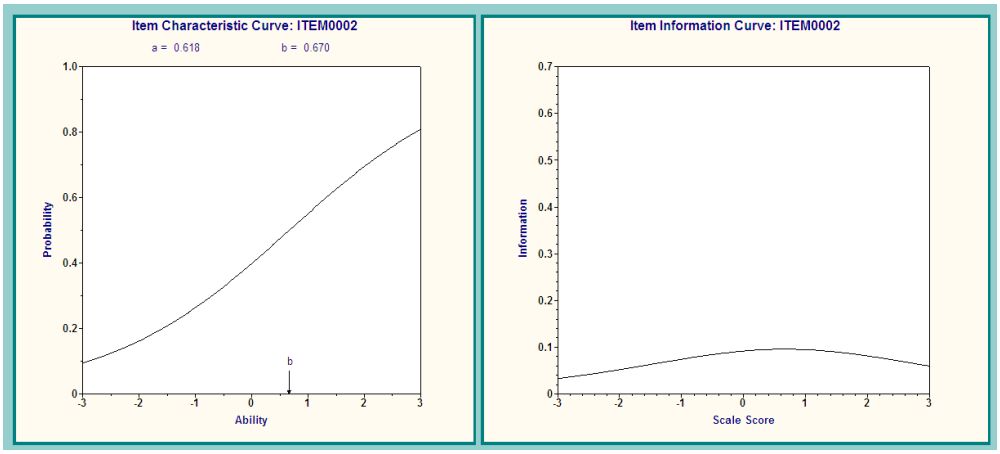
$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \tag{5.2}$$

where $SE(\hat{\theta})$ is called the standard error of estimation. This result holds whenever maximum likelihood estimates of θ are obtained. In the framework of IRT, $SE(\hat{\theta})$ serves the same role as the standard error of measurement in classical measurement theory. It is important to note, however, that the value of $SE(\hat{\theta})$ varies with ability level, whereas the classical standard error of measurement does not (Hambleton, Swaminathan and Rogers, 1991).

Figure 5.1 Item Characteristic Curves and Item Information Curves



(a) Item with high information



(b) Item with low information

Relying on the features of item information, the IRT item information and classical power analysis are combined in this study. Suppose a two-legged clinical trial is constructed, Group 1 (hypothesized) and Group 2 (alternative). The group mean of the first group can be expressed as:

$$\mu_0 = \frac{1}{N}\sum_{n=1}^{N}\theta_n \tag{5.3}$$

where N indicates the number of patients in one branch. If the item number is denoted as K, according to formula (5.1) and (5.2), the variance of estimation at μ_0 can be calculated as:

$$SE^2(\hat{\mu}_0) = VAR(\hat{\mu}_0) = VAR(\frac{1}{N}\sum_{n=1}^{N}\theta_n) = \frac{1}{N^2}\sum_{n=1}^{N}VAR(\theta_n) = \frac{1}{N^2}\sum_{n=1}^{N}\frac{1}{I(\theta_n)}$$
$$= \frac{1}{N^2}\left[\frac{1}{I(\theta_1)} + \frac{1}{I(\theta_2)} + \frac{1}{I(\theta_3)} + \ldots + \frac{1}{I(\theta_n)}\right] \quad (5.4)$$
$$= \frac{1}{N^2}\left[\frac{1}{\sum_{i}^{k} x_i I_i(\theta_1)} + \frac{1}{\sum_{i}^{k} x_i I_i(\theta_2)} + \frac{1}{\sum_{i}^{k} x_i I_i(\theta_3)} + \ldots + \frac{1}{\sum_{i}^{k} x_i I_i(\theta_n)}\right]$$

where x_i equals to 1 when the item is selected or 0 when the item is not selected.

In order to simplify the calculation at the initial stage, the standard deviation of the population is not taken into consideration temporarily. Thus the algorithms (5.4) for variance of measurement at group mean can be followed as:

$$SE^2(\hat{\mu}_0) = VAR(\hat{\mu}_0) = \frac{1}{N^2}\cdot\frac{N}{\sum_{i}^{k} x_i I_i(\mu_0)} = \frac{1}{N\sum_{i}^{k} x_i I_i(\mu_0)} \quad (5.5)$$

Although this expression looks simple, it plays an important role in linking the IRT item information and the classical measurement in power analysis. Furthermore, from formula (5.5), we can easily find that the variance of measurement at the group mean depends on the sample size and the number of items, or item information, to be exact. We are able to minimize the variance of measurement by either increasing the number of patients or adding more items, or even using both of these two methods. However, if we keep variance of measurement constant at a certain level, the sample size and the number of items will show a trade-off relationship: the more sample size, the fewer items are required, and vice versa.

5.3.3 Three Approaches to Detect Power, Sample Size and Item Number

With the help of item information method, we can find that the number of items maintains the least when the item selection procedure follows an optimal way. Thus, if we keep a certain level of variance in measurement, how big the sample size required in the IRT clinical trials is actually dependent on the test information, rather than the number of items, in my point of view. To get the same test information, more items have to be used if their difficulty values are far from the targeted group mean or discrimination parameter is low; conversely fewer items are needed if their difficulty values are close to the population ability or items have high discrimination value. To minimize the number of items but maximize the test information is always the ideal target in designing an IRT test, but hard to do so.

Under the premise that all the items selected in an optimal way, the relationship between number of items, sample size, power and effect size will be investigated by three approaches (Table 5.1). The research could be implemented on simulation studies with data accumulated from previous clinical trials.

Table 5.1 Three Approaches to Detect Power, Sample Size and the Number of Items

Approach	Effect size ($\mu_A - \mu_0$)	Power (z)	Item Number (K)	Sample size (N)	Expected Relationship
1	✓	✓	?	?	$K\uparrow$, $N\downarrow$
2	✓	?	?	✓	$K\uparrow$, Power $\uparrow$
3	✓	?	✓	?	$N\uparrow$, Power $\uparrow$

Note. The tick symbolizes the data required to input by the doctors. The question mark indicates the output data calculated by the statistical tool.

5.3.3.1 Approach 1: ($K, N \mid z, \mu_A - \mu_0$)

The first approach is the most useful but also the most complex option in the statistical tool. Suppose we are to create a two-legged clinical trial, hypothesized group (control group) and alternative group (treatment group). Given the presumed mean values of the two groups and the power expected to reach, the tool will figure out the sample size corresponding to the number of items ranking from 1 to 100. Doctors can decide which pair is the most suitable in his clinical trials according to their own situation, for example, how many patients they can get and how long the test they want to have. Adopting item information method, the trade-off association between these two factors is expected more accurate than before. Suppose the hypothesized group mean is μ_0 and the alternative group mean is μ_A, $\hat{\delta} = \mu_A - \mu_0$ and $\xi = \frac{\sigma_\delta}{\sqrt{n}}$, we can get the power at standard normal distribution as:

$$z = \frac{\mu_A - \mu_0}{\sigma_\delta / \sqrt{n}} = \frac{\hat{\delta}}{\xi} \tag{5.6}$$

When z and $\mu_A - \mu_0$ has known, according to formula (5.5) we can derive ξ as:

$$\xi = \frac{\hat{\delta}}{z} = \frac{\sigma_{\delta}}{\sqrt{n}} = \frac{\sqrt{\mathrm{var}(\mu_A - \mu_0)}}{\sqrt{n}} = \frac{\sqrt{\mathrm{var}(\mu_A) + \mathrm{var}(\mu_0)}}{\sqrt{n}}$$
$$= \frac{\sqrt{\frac{1}{n}\left(\frac{1}{\sum_{i}^{k} x_i I_i(\mu_A)} + \frac{1}{\sum_{i}^{k} x_i I_i(\mu_0)}\right)}}{\sqrt{n}} \qquad (5.7)$$
$$= \frac{1}{n}\sqrt{\frac{1}{\sum_{i}^{k} x_i I_i(\mu_A)} + \frac{1}{\sum_{i}^{k} x_i I_i(\mu_0)}}$$

If the number of items K is set to 1, 2, 3…until 100, we can derive the values of sample size N corresponding to each K by the formula (5.7). Because the doctors prefer use the same length of test for both groups, the item numbers in different branch are kept equal.

However, although some numerical algorithms have been found to calculate the number of items and sample size in clinical trials by using item information method, there is still a lot of work to do in the future. For example, whether or not a fixed coefficient exists in the relationship between sample size and the number of items and whether the association between these two factors varies in different diseases. Such questions are very interesting in further studies.

5.3.3.2 Approach 2: ($K, z \mid N, \mu_A - \mu_0$)

The second approach is easier than the previous one. It is applicable for doctors who have already known the quantity of patients and want to grasp how much power the trial can reach by using various numbers of items. It is expected to find that the more items in inclusion, the more power the trial has. Like the first approach, the results will calculate the power corresponding to the number of items, ranking from 1 to 100, provided the effect size and the sample size are given. The calculation is relatively simple with the transformation of formula (5.6) and (5.7), setting K from 1 to 100.

5.3.3.3 Approach 3: ($N, z \mid K, \mu_A - \mu_0$)

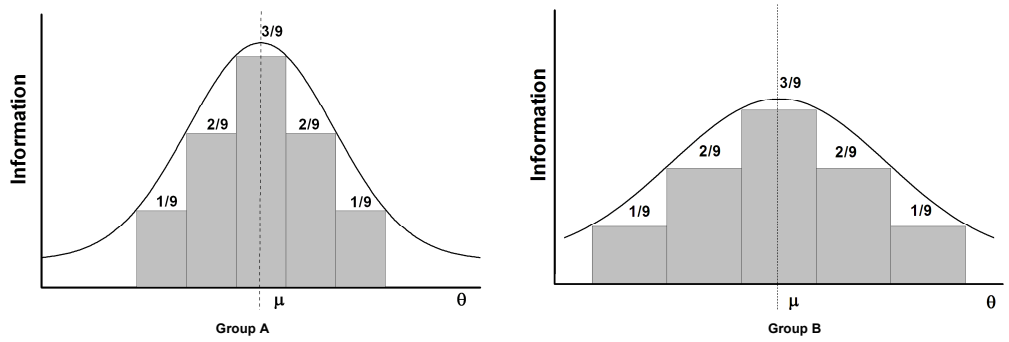
The third approach is expected to help doctors decide the sample size and corresponding power when the effect size and the number of items have been known. Considering that too many items may not be welcomed by patients and doctors, we can set a maximum on the number of items, such as 15 or only 10 items. The same as the previous two approaches, with the formula (5.6) and (5.7), we can derive ξ, but this time we need to set the sample size as a constant from 1 to 5000, for example, instead of setting the number of items. The final results will display the sample size and its corresponding power of the trial. It is reasonable to find that the larger sample size, the more power the trial can have.

5.3.4 Standard Deviation and Item Information Analysis

For a simple calculation in power analysis, the discussions above did not take the standard deviation of population into account. But the standard deviation does exist in fact, which makes the item information and power analysis more complicated.

Suppose the population follows the normal distribution. Group A and B shown in Figure 5.2 have the same mean value but with different standard deviation. If we just follow the item information analysis mentioned above (formula 5.5), the test information of Group A and B should be the same. But actually Group A has obviously more information than B. The reason is that A has a smaller standard deviation, thus the items whose difficulty value is close to group mean can provide more information than B. Compared to Group A, Group B spreads broadly. The items close to group mean value fail to show high information, which means that the item and the ability of patients does not match well. Therefore, if Group B wants to have the same power as Group A, B has to add more items if the sample size keeps unchanged or add more patients if the same number of items has to be used.

Figure 5.2 Comparison on Test Information of Two Groups with the Same Mean Value but Different Standard Deviation



To solve the problem of standard deviation, the test information is divided averagely into L parts, each of which takes the weight equaling to its proportion to the total. Thus the test information can be rewritten as:

$$I_{total} = \sum_{g}\sum_{i}^{k} x_{gi}(\sum_{l} P_l I_i(\theta_l)) \tag{5.8}$$

where g is the parameter for group, P_l indicates the weighting proportion of each part to the total. Based on formula (5.8), if standard deviation has to be taken into account in necessity, the formula (5.5) can be transformed as:

$$SE^2(\hat{\mu}_0) = VAR(\hat{\mu}_0) = \frac{1}{I_{total}} = \frac{1}{\sum_{g}\sum_{i}^{k} x_{gi}(\sum_{l} P_l I_i(\theta_l))} \tag{5.9}$$

In the above example (Figure 5.2), suppose $L = 5$, that is, both Group A and B follow in normal distribution with averagely 5 columns. The total information in these two groups can be written as:

$$I_{total} = x_{Ai} \cdot \frac{1}{9} \cdot I_i(\theta_1) + x_{Ai} \cdot \frac{2}{9} \cdot I_i(\theta_2) + x_{Ai} \cdot \frac{3}{9} \cdot I_i(\mu) + x_{Ai} \cdot \frac{2}{9} \cdot I_i(\theta_4) + x_{Ai} \cdot \frac{1}{9} \cdot I_i(\theta_5) \quad (5.10)$$

Because the normal distribution is symmetrical to the mean value, so the algorithm (5.10) can also be transferred as:

$$I_{total} = 2\left[x_{Ai} \cdot \frac{1}{9} \cdot I_i(\theta_1) + x_{Ai} \cdot \frac{2}{9} \cdot I_i(\theta_2) \right] + x_{Ai} \cdot \frac{3}{9} \cdot I_i(\mu) \quad (5.11)$$

On the base of calculation in (5.10) and (5.11), a general rule of weights in each column of normal distribution could be explored. If we divide the cumulative distribution into infinitely small parts, suppose (2m+1) parts in total can be derived. So the total information is calculated as:

$$I_{total} = 2 \cdot \left[x_i \cdot \frac{1}{(m+1)^2} \cdot I_i(\theta_1) + x_i \cdot \frac{2}{(m+1)^2} \cdot I_i(\theta_2) \ldots + x_i \cdot \frac{m}{(m+1)^2} \cdot I_i(\theta_m) \right] + x_i \cdot \frac{(m+1)}{(m+1)^2} \cdot I_i(\mu) \quad (5.12)$$

Therefore the one-side distribution (left or right part of the symmetrical axis) has the weight to each column as $\frac{1}{(m+1)^2}, \frac{2}{(m+1)^2}, \frac{3}{(m+1)^2} \ldots \frac{m}{(m+1)^2}, \frac{m+1}{(m+1)^2}$, ranking in an ascending order. The biggest weight is $\frac{m+1}{(m+1)^2}$, belonging to the column containing the group mean.

5.3.5 Ceiling and Bottom Measurement

Ceiling and bottom ability measurement is very challenging in clinical trials because of few information can be derived around the extreme point. The ceiling (very healthy) and bottom (extremely disabled) groups are no longer distributed in a normal way, but could be regarded as skew-right and skew-left distribution approximately. The idea is to follow the formula (5.8) and (5.9) to divide the abnormal cumulative distribution averagely into small parts, similar to the discussion above regarding the standard deviation of population and item information analysis. One point worthy emphasis is that the group mean in the normal distribution should be changed to median value in skew-right or skew-left distribution, because the data is not symmetrical to the mean in this situation.

Figure 5.3 Test Information for Bottom Group in Clinical Trials

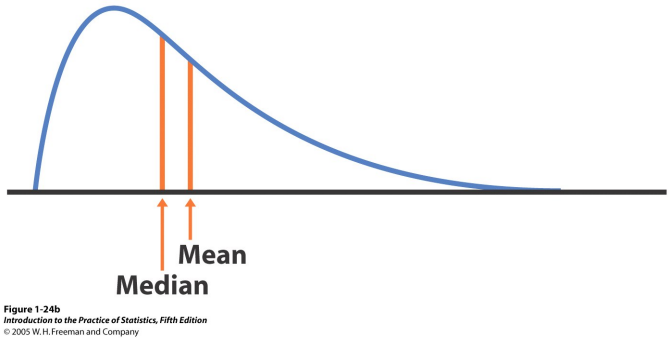


Figure 1-24b
Introduction to the Practice of Statistics, Fifth Edition

The right-skew distribution in Figure 3 shows the test information for the bottom group. Like the method with inclusion of standard deviation, the bottom group can be averagely divided into L parts. The proportion of each part to the total is the weight on the test information. To use the three options in the statistical tool, the variance of the median value must be acquired. Thus the formula (5.9) transfers as

$$SE^2(median) = VAR(median) = \frac{1}{I_{total}} = \frac{1}{\sum_g \sum_i^k x_{gi} (\sum_l P_l I_i(\theta_l))} \tag{5.13}$$

for the purpose of ceiling and bottom group in measurement. With the results generated from formula (5.13), the three approaches in the statistical tool can also be applied to ceiling and bottom measurement, in which the number of items, sample size and power can be detected.

However, the feasibility of this proposed method in measurement of ceiling and bottom group needs to be investigated further. How to set the median value and how to set the general rule of weighting in each part are suspending questions in this method. Since the skew-right and skew-left distribution is more complex than the normal distribution, more knowledge in statistics is required to learn.

5.3.6 Reliability of the Test

The concept of reliability comes from classical test theory, in which the standard error is the square root of 1 minus reliability. This indicator is quite familiar to the researchers and doctors when checking whether the ALDS test is good enough. In IRT test, the calculation of reliability is transferred by the standard error measurement. The smaller standard error of theta-hat, the more reliable the test is.

$$SE(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})} = \sqrt{\frac{1}{\sum x_i I(\theta)}} = \sqrt{1-\rho} \tag{5.14}$$

$$\rho = 1 - \frac{1}{\sum x_i I(\theta)} \tag{5.15}$$

where x_i equals to 1 when the item is selected or 0 when the item is not selected and ρ indicates the reliability.

For example, the typical rule for setting standard error is 0.32, assuming the test reliability is around 90% (Linacre, 2000).

$$SE = \sqrt{1-\rho} = \sqrt{1-.90} \approx .32 \tag{5.16}$$

Meanwhile, in the IRT theory, the amount of information provided by a test at θ is inversely related to the precision with which ability is estimated at point.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum I(\theta)}} \tag{5.17}$$

Thus, when the standard error of estimation is around .32, the information value is approximate 10.

$$\begin{aligned} &\because .32 \approx \frac{1}{\sqrt{\sum I(\theta)}} \\ &\therefore \sum I(\theta) \approx 10 \end{aligned} \tag{5.18}$$

Suppose the items are following Rasch Model, that is, only one parameter b, the discrimination parameter $a = 1$ and the guessing parameter $c = 0$, then the test needs 40 items when the ability θ equals b.

$$I(\theta) = a^2 P(\theta) Q(\theta) = 40 \times .50 \times .50 = 10 \tag{5.19}$$

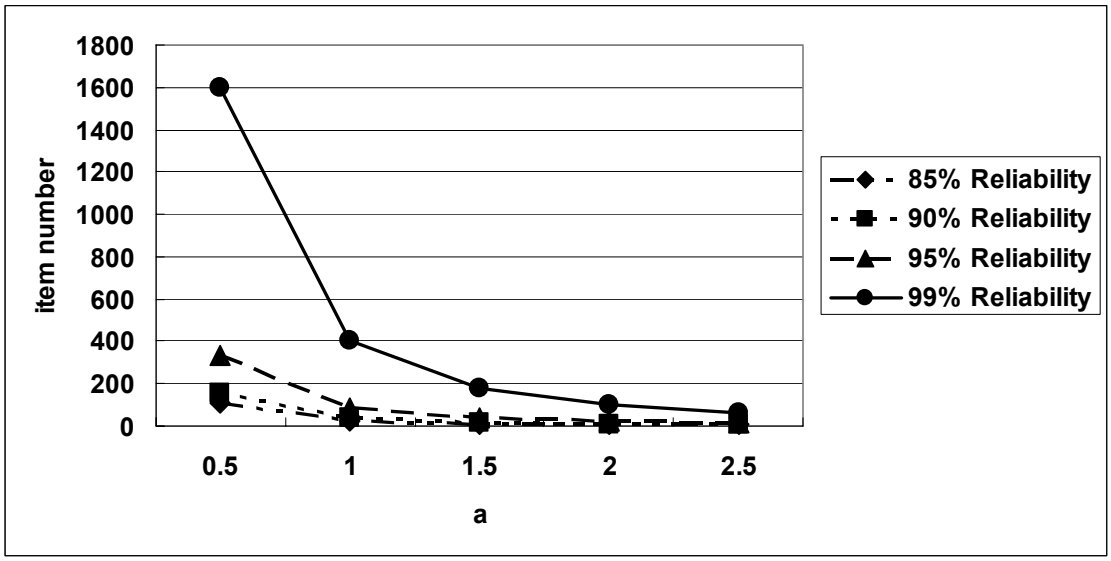
Thus, if we increase the discrimination parameter ($a > 1$), then fewer items (less than 40) need to be used in the test to get the same information. But if the discrimination drops down ($a < 1$), then more items (more than 40 items) are in requirement. Table 5.2 shows the variation of items demanded in the test to reach the amount of information (reliability) when changing the discrimination parameter.

Table 5.2 The Relationship Between Reliability, Item Numbers and Discrimination Parameter

	a				
	0.5	1	1.5	2	2.5
Reliability=85%, SE=.39, I=6.57					
Item Number	105	26	11	6	4
Reliability=90%, SE=.32, I=9.77					
Item Number	156	40	17	10	6
Reliability=95%, SE=.22, I=20.66					
Item Number	330	82	36	20	13
Reliability=99%, SE=.10, I=100					
Item Number	1600	400	178	100	64

From the above analysis, two conclusions can be drawn: the more reliability, the less standard error, and the more information of the test in IRT context. To obtain more information, we need more items. To increase the discrimination parameter of items can help reduce the item number. But it does not imply that the higher discrimination, the better effectiveness we can get. From Figure 5.4 we can see no matter what degree of reliability the curve represents, all of them decrease sharply at the beginning but gradually slow down the changing speed. It suggests that even if we put great efforts in increasing discrimination, especially when $a > 2.5$, the effects could not be changed by a large scale.

Figure 5.4 The Relationship Between Item Numbers and Discrimination Parameter under Different Reliability Value



5.3.7 DIF in Power Analysis

Differential Item Function (DIF) is an unavoidable problem in the item bank. Due to the differences of age, gender, nationality, living areas among the respondents, items with special orientation may generate bias between different patients. Generally, DIF items can be labeled with different item parameters when they are used by different groups (Holman, 2005). For example, the item "go shopping" uses different item parameters when answered by male and female respondents. However, this method is more applicable in computerized adaptive test (CAT), in which the DIF element, such as age and gender can be filtered before the patient answers to adaptive tests. Unlike the CAT, patients are randomly divided into two groups in clinical trials. Thus, the different sets of item parameters for a single item are hard to realize in normal IRT test.

Some restrictions for DIF in the process of construction clinical trials could be set in advance. For example, the percentage of the gender-favored items in the test should keep the same as the gender ratio in the population of each group. Suppose there are 100 patients in each branch, and 40 of each are female, the female-favored items should keep as 40% in the total IRT test. A new variable for DIF can be added in the information expression, for instance:

$$y_i = \begin{cases} -1 & (\text{male-favored}) \\ 0 & (\text{nobias}) \\ 1 & (\text{female-favored}) \end{cases} \tag{5.20}$$

Setting restrictions on item selection procedure is a sector similar to "shadow test". Only those that are satisfied with all restrictions can get through the "filter" into the sub-item bank for further adaptive optimal item selection. The mathematical restrictions on items will be set at the initial stage of the statistical tool before the qualified items entry into automatic item selection procedure.

5.3.8 Automatic Item Selection

Automatic item selection procedure can also be regarded as a special case when the hypothesized group has the same mean value as the alternative group $\mu_0 = \mu_A$, that is, the effect size between the two groups equals to 0.0. When the two groups are totally overlapped, the power analysis between two groups would be simplified as item information analysis on the targeted theta.

The statistical tool in ALDS website plans to design three options to help doctors select the optimal items in their tests. First, the doctor is asked to input the targeted theta (or ALDS score) and the number of items he expects in the test. Then the tool will calculate information for each item on the base of theta and rank the information of each item by a descending order. The selection process will start from the top item with highest information and stop at the number that the doctor has input. Finally the system will output these most optimal items as well as the total test information. This option ensures that the doctors can get the most powerful items for the targeted theta.

Secondly, the doctor is asked to input the targeted theta (or ALDS score) and the expected reliability of the test. According to formula (5.14), the test information can be easily derived. Similar as the first option introduced above, the tool will calculate information for each item and rank them by descending order. The cumulative test information will be stopped until the expected standard is reached. The test containing the items above the cumulative test information standard can also be regarded as the shortest test as the doctor's expectation.

Thirdly, the doctor is asked to input the targeted theta and select items of his preference. Then the tool will feedback information for each selected item on the base of targeted theta, and sum the item information as total to get the test information. Meanwhile, this tool can also advise the number of items that the doctor needs to add if he sets a certain power to be reached. Followed the same procedure as the second option, we can get the shortest test as expectation. Suppose the shortest test is denoted as Test A, and the test made by the doctor himself set as Test B. The comparison of information functions is done by computing the relative efficiency of one test, compared with the other, as an estimator of ability at theta.

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)} \tag{5.21}$$

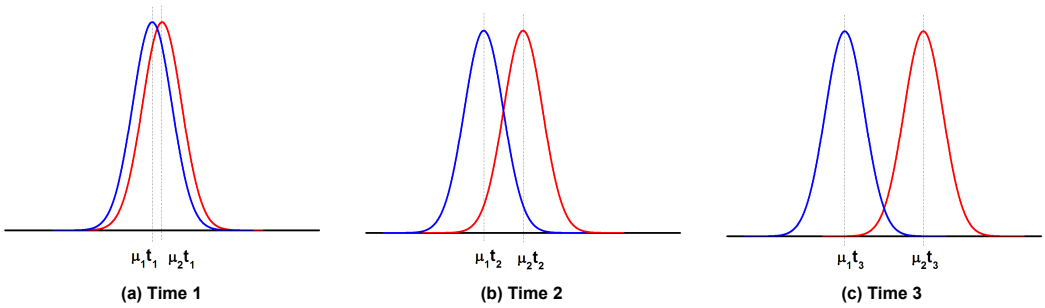
where $RE(\theta)$ denotes relative efficiency and $I_A(\theta)$ and $I_B(\theta)$ are the information functions for Test A and Test B, respectively, defined over a common ability scale, θ. If, for example, $I_A(\theta) = 25.0$ and $I_B(\theta) = 20.0$, then $RE(\theta) = 1.25$, and it is said that at θ, Test A is functioning as if it were 25% longer than Test B. Then, Test B would need to be lengthened by 25% (by adding comparable items to those items already in the test) to yield the same precision of measures as Test A at θ.

5.3.9 Power Analysis in Pretest and Posttest

Pretest and posttest are popular methods to track the treatment effects on patients. Based on construction of single clinical trials, the statistical tool is also capable of providing guidance for doctors in successive trials, which is assumed as an extension of the single trial in statistical aspect. Time as a new parameter in power analysis makes a two-dimension matrix.

Firstly, if we take the time as the predominant factor, the comparison of treatment effects on two groups would be detected at different time points. As Figure 5.5 shows, the effect size between the two groups enlarges with the time extending. If the significance level keeps the same as 95%, and the population in each group maintain unchanged, the power in (a), the first examine time point has the lowest power because it has the smallest effect size. The measurement at the last time point has the highest power to detect the effect size between these two groups. The power analysis can be made as the same as a single trial at each time point. And doctors can use the time-group dimension to track the changes between groups at different time point.

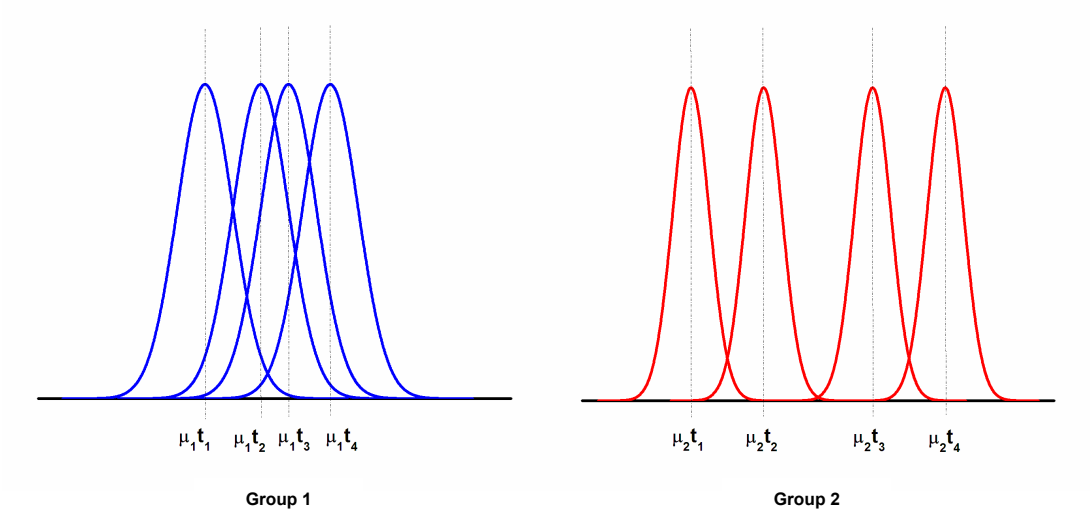
Figure 5.5 Comparison of Power Analysis on Two Groups at Three Time Points by Time Dimension



Secondly, if we take the group as predominant factor, the changes of each group will be detected separately during the same time periods. Figure 5.6 illustrates the different treatment effects in Group 1 and Group 2 at four time points. The trial in each group can be regarded as a single one to make power analysis. For example, in Group 1, $\mu_1 t_1$ and $\mu_1 t_2$ can be regarded as hypothesized group and alternative group respectively. Following the framework of power analysis and item information analysis introduced above, it is not difficult to make the comparisons between the pairs. Suppose Group 1 is the control group and Group 2 is the treatment group, we can obviously track that the second group has better effects than the first one at four various points.

A special case needs to note that when there is only one group in the trial, for example, a group of patients in brain stroke disease needs to examine periodically by quarter, we can transfer this case into a single trial directly by setting $\mu_0 = \mu_{t1}$ (pretest), $\mu_A = \mu_{t2}$ (posttest).

Figure 5.6 Comparison of Power Analysis at Four Time Points on Two Groups by Group Dimension



In addition, the doctors can also choose whether they want to include the same items or overlap some items in the tests for two groups or in pretest and posttest. The algorithms of the item selection restrictions may be written as:

(1) Inclusion of totally same items in two tests: $-\delta \le x_{1i} - x_{2i} \le \delta$, for all i

(2) Inclusion of totally different items in two tests: $x_{1i} + x_{2i} \le 1$, for all i

(3) Inclusion of overlapping items in two tests:

$$\left\{ \begin{array}{l} x_{1i} + x_{2i} \ge 1 \\ \sum (x_{1i} + x_{2i}) < 2 \sum x_{1i} \cdot overlap\% \end{array} \right\}$$

Chapter 6 Conclusions and Recommendations

This thesis focuses on the information analysis of website construction for AMC Linear Disability Score project. As stated earlier, the ALDS project is based on a modern psychometric method, the item response theory (IRT), which is adapted from educational measurement to determine the cognitive ability of schoolchildren.

6.1 Conclusions

Chapter 2 of this thesis describes the current status of ALDS system and its application. The ALDS item bank has been calibrated by using the responses from over 4,000 patients with a broad range of stable chronic conditions. A total of 190 items were identified and then described in detail at the initial stage. But only 77 items are remained currently as applicable ones to be commonly used in measurements, with the range of difficulty level from -3.49 to +3.05. As the recent modification, the item bank remains only 73 items. The current measurement procedures of ALDS are divided into four steps: item selection, data collection, data analysis and data output and storage. Nowadays, the item selection is done by the clinicians and researchers together to pick out the preferred items and combine as booklets. The booklets are varied with difficulty levels. The data analysis is based on two computer programs: SPSS and BILOG, and then linear transferred to the ALDS score. No all the datasets have been saved in a common place so far. And a lot of repetitive work has to be done for ALDS calculation.

Chapter 3 proposes a website framework for ALDS project. In order to clarify the functions of different modules, an ID Code identification system is designed to distinguish the users as patients, assistants and researchers. The researchers and clinicians can ask the website to automatically select the optimal items after they input the group mean level and standard errors. The selected items are organized as a sub-item bank and are allocated as a code number. The patient page is linked with the CAT program. When the code number of test is input, the website will automatically start the CAT program and save the patients' results on the server and send to the specified doctor. As a helper for the doctor, the assistant can see the items in the sub-item bank and the summary report under the code number besides the CAT program. All the dataset will be stored in the server temporarily or permanently.

Chapter 4 discusses the rationale of CAT in ALDS project and illustrates the CAT procedures with an example from brain stroke study. Because of the limitations in practice, the CAT version and pencil-and-paper version will be parallel used in the website. This may bring measurement bias between these two instruments.

Chapter 5 describes a system of power analysis for constructing clinical trials in ALDS. Item information is an essential indicator in IRT to check the "power" of an item. With the help of item information method, we can find that the number of items maintains the least when the item selection procedure follows an optimal way. This statistical tool is suggested to add into the ALDS website, thus the power, sample size, and item numbers can be calculated simultaneously when the researcher input preferred effect size and the value of group mean. In addition, the ceiling and bottom

measurement, DIF problems and pretest and posttest in power analysis are also discussed on the base of item information analysis.

6.2 Recommendations

For a better measurement on different group of functional status, new items are highly recommended to add in. There are four reasons to support this view:

(1) The item bank was constructed a couple of years ago. Some items are apparently outdated, ambiguous and potentially biased, which have to be substituted by the fresh ones.
(2) The 73-item bank ranging from 11 to 89 in ALDS score is lack of items with the difficulty at the in ceiling and bottom level. The "holes" need to be filled in as soon as possible.
(3) The item bank was calibrated in a too strict and complex way; as a result, over two-thirds items were skipped out because of unfitness in the IRT model. If the calibration criteria were less restricted, some items in the non-calibrated group would be included into the item bank.
(4) The Unicode of 312 items always makes confusion in interpretation. To label a new code for the items included in the bank will highly improve the efficiency in data analysis. Nowadays, in order to solve the problem in Unicode, researchers in AMC are trying to build up a database by ACCESS to locate items in requirement by activating systematic code.

Item exposure issue is also a good point to taken into account in the future research. Considering patients will be unlikely to give responses every time for the repeated items, new items with similar difficulty level and discrimination would be substituted some highly repetitive items, especially in the longitudinal studies. Since ALDS project only has 77 items so far, it is not that necessary to take the item exposure rate into consideration at this moment. However, with the development of ALDS, when more items are included in the item bank, the item exposure needs to be added in.

The power analysis discussed in Chapter 5 on the base of theoretical analysis. No simulation studies or pilot studies have been practiced yet. For the effectiveness of this new method –integration of item information method with power analysis –should be further checked with real data or simulation studies.

ALDS project adopts dichotomous options, that is, only yes or no answers will be recognized. In spite of simplicity, this method loses a lot of information in estimates. For example, each item pencil-and-pen version will have four options: a) "I can", b) "I can but with difficulty", c) "I cannot" and d) "not applicable" while the CAT version only leaves two options "I can" (combine option a) and b) in the PP version) and "I cannot" (combine option c) and d) in the PP version). Half of the information has been lost in the option combination process. Therefore, a polytomous options or psychometric scale for ALDS is recommended. For example, the patient could be asked the degree of his disability as the range from 0 (no symptom) to 10 (severely). More attentions could be paid on such aspect in the future.

References

Bergner M., Bobbitt R. A., Carter W. B., & Gilson B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Med Care*. 1981, 19, 787-805.

Bonita, R., & Beaglehole, R. (1988). Modification of Rankin Scale: Recovery of Motor Function after Stroke. *Stroke, 19*, 1497-1500.

Brazier, J., Roberts, J., & Deverill M. (2002) The estimation of a preference-based measure of health from the SF-36, *Journal of Health Economics*, 21 (2002) 271–292.

Cizek, G. J. (2001). *Setting performance standards: concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif.: Sage Publications.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Creswell, J. W. (2003). *Research design: qualitative, quantitative, and mixed method approaches* (2nd ed.). Thousand Oaks, Calif.: Sage Publications.

Glas, Geerlings, van de Laar and Taal, "Analysis of Longitudinal Randomized Clinical Trials Using Item Response Models" (2008).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif.: Sage Publications.

Holman, R. (2005). *Item Response Theory in Clinical Outcome Measurement.* Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2 (4), 359-375

Linacre, J. M. (2000). *Computer-Adaptive Testing: A Methodology Whose Time Has Come.* University of Chicago, Chicago, USA.

Lindeboom R., Vermeulen M., Holman R., & de Haan R., (2003). Activities of daily living instruments: optimizing scales for neurologic assessments. *Neurology* 2003, 60, 738-742.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). New York, NY: Chapman & Hall.

Mills, C. N. (Ed.). (2002). *Computer-based testing : building the foundation for future assessments*. Mahwah, N.J.: L. Erlbaum Associates.

Mills, M., Bunt, G. G. v. d., & Bruijn, J. d. (2006). Comparative Research: Persistent Problems and Promising Solutions. *International Sociology, 21*, 14.

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: W.H. Freeman and Co.

Nichol M. B., Sengupta N., Globe. (2001) Evaluating Quality-Adjusted Life Years: Estimation of the Health Utility Index (HUI2) from the SF-36, *Med Decis Making*, 21, 105-112.

Nunnally, J. C. (1967). *Psychometric theory*. New York,: McGraw-Hill.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Øyen, E. (Ed.). (1990). *Comparative methodology : theory and practice in international social research*. London ; Newbury Park, Calif.: Sage.

Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry*. New York,: Wiley-Interscience.

Sireci, S. G., Patelis, T., Saba, R., Dillingham, A. M., & Rodriguez, G. (2000). *Setting Standards on a Computerized Adaptive Placement Examination*. Paper presented at the USA Annual Meeting of the National Coluncil on Measurement in Education.

Scheerens, J., Glas, C. A. W., & Thomas, S. (2003). *Educational evaluation, assessment, and monitoring : a systemic approach*. Lisse [Netherlands] ; Exton, PA: Swets & Zeitlinger.

Stoop, E. M. L. A. (2001). *Detection of Misfitting Item-Score Patterns in Computerized Adaptive Testing.* The University of Twente, Enschede, the Netherlands.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology : combining qualitative and quantitative approaches*. Thousand Oaks, Calif.: Sage.

van der Linden, W. J. (2005). *Linear models of optimal test design*. New York, NY: Springer.

van der Linden, W. J. (2003). *Computerized Test Construction.* the University of Twente, Enschede, the Netherlands.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing : theory and practice*. Dordrecht ; Boston: Kluwer Academic.

van der Linden, W. J., & Zwarts, M. A. (1987). *Some Procedures for Computerized Ability Testing.* the University of Twente, Enschede, the Netherlands.

VanSwieten, J., Koudstall, P., Visser, M., Schouten, H., & VanGijn, J. (1988). Interobserver Agreement for the Assessment of Handicap in Stroke Patients. *Stroke, 19*, 604-607.

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing : a primer* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473-492.

Weiss, D. J. (1985). Adaptive Testing by Computer. *Journal of Consulting and Clinical Psychology, 53-6*(American Psychological Association, Inc.), 15.

Weisscher, N. (2008). *The AMC Linear Disability Score (ALDS): Measuring Disability in Clinical Studies.* Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands.

World Health Organization. (2007). *Disability and Rehabilitation WHO Action Plan 2006-2011*.

Zeller, R. A., & Carmines, E. G. (1980). *Measurement in the social sciences: the link between theory and data*. Cambridge ; New York: Cambridge University Press.

Zimowski, M F., Muraki, E., Mislevy, R.J., & Bock, R.D., (2003). *BILOG-MG for Windows*, Scientific Software International, Inc. (Version: 3.0.2327.2)

Appendix A: Modified ALDS 73-Item Bank

No.	Item (Are you able to...)	b	a	Descriptions	Remarks
1	… ride a bike for at least 2 hours?	3.50	2.45	a nice weather, a long cycling	DIF may exist in different countries.
2	… vacuum a flight of stairs (10 to 15 stairs)?	2.65	3.23	getting the vacuum cleaner out, carrying it up and down one flight of stairs and putting it back in the cupboard, the stairs may be set as in patients' house or some public settings	DIF may exist in different countries.
3	… carry a bag of shopping upstairs?	2.13	2.70	walking up a flight of stairs with a full bag of shopping	
4	…clean the whole bathroom?	1.95	3.07	getting the cleaning materials, cleaning the floor, walls, shower, bath, sink plug holes and taps, and then putting everything back	
5	...vacuum a room and move light furniture?	1.87	2.46	getting the vacuum cleaner, vacuuming a whole living room or bedroom, moving light furniture such as chairs, vacuuming under a dining table or bed, put everything back.	
6	...fetch groceries for 3-4 days?	1.63	2.44	fetch groceries for a number of days, paying for them, carrying the shopping home	DIF may exist in different gender, female-favored.
7	...go for a walk in the woods?	1.50	2.56	walking for a while (usually more than 15 minutes) on an uneven and unpaved path without getting lost	
8	...travel by local bus or tram?	1.23	2.86	going to the bus stop, getting on, finding a seat, sitting down and getting down	
9	...walk for more than 15 minutes?	0.81	2.13	walking for more than 15 minutes on a well paved path without getting lost	
10	...carry a full tray?	0.80	1.62	carrying a full tray (full tea or coffee pot, cups and biscuits) from the kitchen to the dining room safely	
11	...walk up a hill or high bridge?	0.78	1.99	walking 100 to 150 meters long up to a hill	DIF may exist in different areas (urban and rural).
12	...go shopping for clothes?	0.72	3.40	going to a shopping centre (walking, cycling, by car or by public transport), walking around, getting into a number of shops, having to try clothes or shoes on, buying a number of articles, paying for them and going home, the whole process is in rush hour	DIF may exist in different gender, female-favored.
13	...cut your toe nails?	0.66	1.63	sitting in bed or on a chair with scissors or nail clippers within reach	
14	...go to a party?	0.55	1.41	going to a birthday or wedding party in the evening (walking, cycling, by car or by public transportation), taking an active role in a social event with more than 10 people, having a drink or something to eat and going home	
15	...stand for 10 minutes in the row?	0.52	1.83	patients are required to stand in a row for about 10 minutes	
16	...go to a restaurant?	0.48	1.98	going to a restaurant (walking, cycling, by car or public transport), ordering food and drinks, going to the toilet, paying for the meal, going home	
17	...sweep the floor?	0.45	2.87	sweeping one room with a long-handled broom, and putting everything back to original position away	
18	...hang out and take in a load of washing?	0.44	2.26	hanging a full load of wet washings outside on a washing line or inside on a clothes horse, take everything down again after clothes become dry, do not use the automatic drier	
19	...vacuum without moving any furniture?	0.35	2.47	getting vacuum machines, vacuum living room or bedroom without moving furniture, but the floor under the table or bed should be included, and putting back the vacuum	

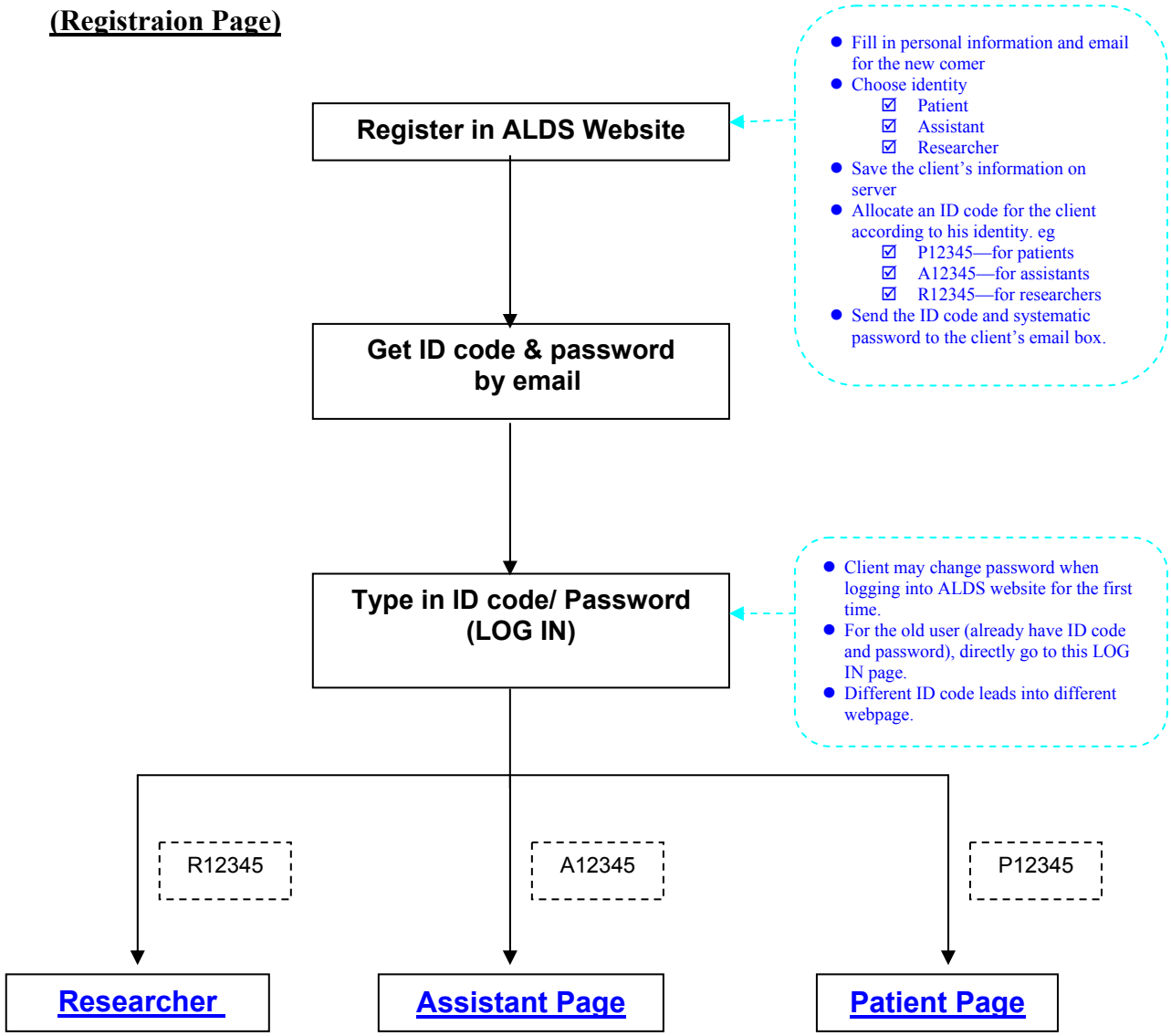
20	...move a bed or table?	0.30	1.34	a bit heavy table, desk or bed, for example during vacuuming patients have to mop under the furniture	
21	...use a washing machine?	0.23	2.07	putting clothes in washing machine, using washing machine, taking out clothes and putting into laundry basket after washing, hanging up is not included	
22	...get something out of a top kitchen cupboard above your head?	0.23	1.53	the cupboard is above head, the patients must reach by hand to open the cupboard door, get out of objects, such as cup and package of coffee, long-handle sticks are not required to use; standing on toes is not necessary	
23	...walk up a flight of stairs (10 to 15 stairs)?	0.19	2.19	a flight of stairs in patients' own house or some house with stairs, 10-15 stairs	
24	...go to the bank or post office?	0.13	3.12	going to post office or bank (walking, cycling, by car or by public transport), handling necessary administrative affairs, and going back home.	Internet may be taken into consideration in new calibration
25	...walk down a flight of stairs (10 to 15 stairs)?	0.02	2.62	a flight of stairs in patients' own house or some house with stairs, 10-15 stairs	
26	...make an appointment and go to the general practitioner?	-0.02	3.29	making an appointment with general practitioner and remembering the appointment, going to general practitioner (walking, cycling, by car or by public transport), and going back home	
27	...use a dustpan and brush?	-0.08	2.50	sweeping a small quantity of dust with a brush by stooping or kneeling on the ground, gathering dust in dustbin, throwing dust away, and putting everything back	
28	...go for a short walk less than 15 minutes?	-0.07	2.06	walking for less than 15 minutes on a well paved path without getting lost	
29	...change the sheets on a bed?	-0.20	1.56	taking away old sheets and pillow cases into laundry basket, putting on clean sheets and pillow cases	keep the item until a better one to replace
30	...cross the busy road in the city?	-0.22	2.91	crossing the busy road, pavement edge, within cultivated circle, safely b) the road is specified in the city, at a busy crossing	
31	...open and close a window?	-0.23	1.42	window above head, patients have to use crutch or chair to help open or close	complex item, vagueness in CAT &PP
32	...fetch a few things from the shops?	-0.29	2.53	settling up messages (bread, milk, cheese and etc.), carrying them back home, help of stickers is allowed	
33	...polish shoes?	-0.34	1.90	getting cleaning materials and shoes, polishing shoes, sitting is permitted, and putting everything back	
34	...have a shower and wash your hair?	-0.65	1.95	going to bathroom, taking soap and towel along, sitting is permitted in shower, using the tap, washing hair, drying whole body, leaving bathroom, but dressing up is not included. Patients can be taken into the washroom by nurses or other people instead of walking to washroom by themselves	
35	...fold up the washing?	-0.70	1.56	a full laundry basket with dried laundries, folding clothes, and putting them back into wardrobe	DIF may exist in different gender, female-favored
36	...dust?	-0.70	2.39	cleaning table, furniture, window bank and easily reached place) with dry or wet cloth	
37	...put on and take off socks and lace-up shoes?	-0.76	1.58	putting on and taking off socks and shoes with laces by squatting or sitting on a chair	
38	...clean a toilet?	-0.77	2.10	getting cleaning materials and toilet brush, cleaning inside and edge, putting everything back	
39	...make a bed?	-0.84	1.52	making a bed to prepare sleeping, positioning quilts, pillows, sheets and etc.	keep the item until a better one to replace
40	...cut your finger nails?	-0.90	1.73	sitting in bed or on a chair with scissors or nail clippers within reach	

41	...bent or kneel to reach under the table?	-0.91	1.44	picking up objects (cups, magazines and etc.) under the table by stooping or kneeling on the ground, but brooms and sticks with long handles are not allowed to use	
42	...make egg or beans on toast?	-1.02	3.08	getting bread, butter, eggs and etc., putting beans on bread, all operations should keep safe, and putting everything back	
43	...bent to reach into a low cupboard?	-1.09	1.51	opening the low cupboard door by stooping or kneeling to get out objects (e.g. shoes) and putting everything back	
44	...move between two low chairs?	-1.14	1.38	standing up from a low chair (armchair) walking to another low chair in the same room, sitting on the other low chair	check the English version with native speaker
45	...pick something up from the floor (not under the table or bed)?	-1.15	2.02	stooping or kneeling to pick up objects (cup, magazine, clothing and etc.) from the ground, not under table or bed	
46	...clean a bathroom sink?	-1.18	2.78	going to bathroom, getting the detergent and cleaning materials along, cleaning the wash-hand basin, taps and pocks, putting everything back	
47	...put the washing up away?	-1.26	2.00	turning off bowl-washing machine and putting all the dishes and bowls back to cupboard	check the English version with native speaker
48	...read a newspaper?	-1.27	0.90	sitting beside a table or on a chair, understanding what the patients are reading	Keep the item until a better one to replace
49	...get in and out of a car?	-1.34	2.17	getting into a normal private car or taxi, and getting out of the car	
50	...make porridge?	-1.36	2.44	getting milk from fridge, pouring milk in pan, warming up (possibly in microwave oven), stirring the porridge, putting everything back, and all operations should keep safe	
51	...clear the table after a meal?	-1.47	2.56	putting bowls and dishes back to kitchen, putting food back into fridge or cupboard, putting rubbish in dustbin	
52	...peel and core an apple?	-1.49	1.20	getting apples and knife, peeling an apple and cutting it into pieces	
53	...prepare breakfast or lunch?	-1.52	2.27	getting bread, butter and glass, putting on tablecloth, placing bowls, dishes, forks, knives and glasses and pouring milk in a glass	
54	...clean the kitchen surfaces?	-1.76	2.96	getting cleaning materials and detergent, cleaning work surface and hob (e.g. after cooking a warm meal), putting everything back	
55	...put an empty chair up to a table?	-1.77	2.06	putting an empty chair back to a table (after a meal)	
56	...eat a meal at the table?	-1.78	1.35	cutting, mashing and eating food by patients themselves instead of being feed, the food has been ready on the table, the patients can be taken to the table	check the English version with native speaker
57	...wash up?	-1.86	2.24	washing cups, glasses, bowls and etc., filling sink with water, using detergent and brush, putting the wash-up in rack, drying is not required	
58	...put on and take off socks and slip on shoes?	-1.93	1.90	taking on and off socks and shoes without laces (e.g. slippers) by squatting or sitting on a chair	
59	...get a book off the shelf at your eye-level?	-2.11	1.67	getting a book from shelf at eye level, and putting it back	
60	...answer the telephone?	-2.14	1.16	hearing, seeing, feeling the indicator, picking up the telephone, making a short and understandable conversation	keep the item until a better one to replace
61	...hang clothes up in a wardrobe?	-2.19	2.65	opening the wardrobe, hanging on clothing (jacket, undershirt and etc)	
62	...make coffee or tea?	-2.35	2.32	making coffee or tea by electric case or kettle and pouring hot water in a pot or cup, all operations should keep safe	
63	...put long trousers or skirts on?	-2.36	2.74	putting on trousers or skirts, and clasping zipper or buttons	

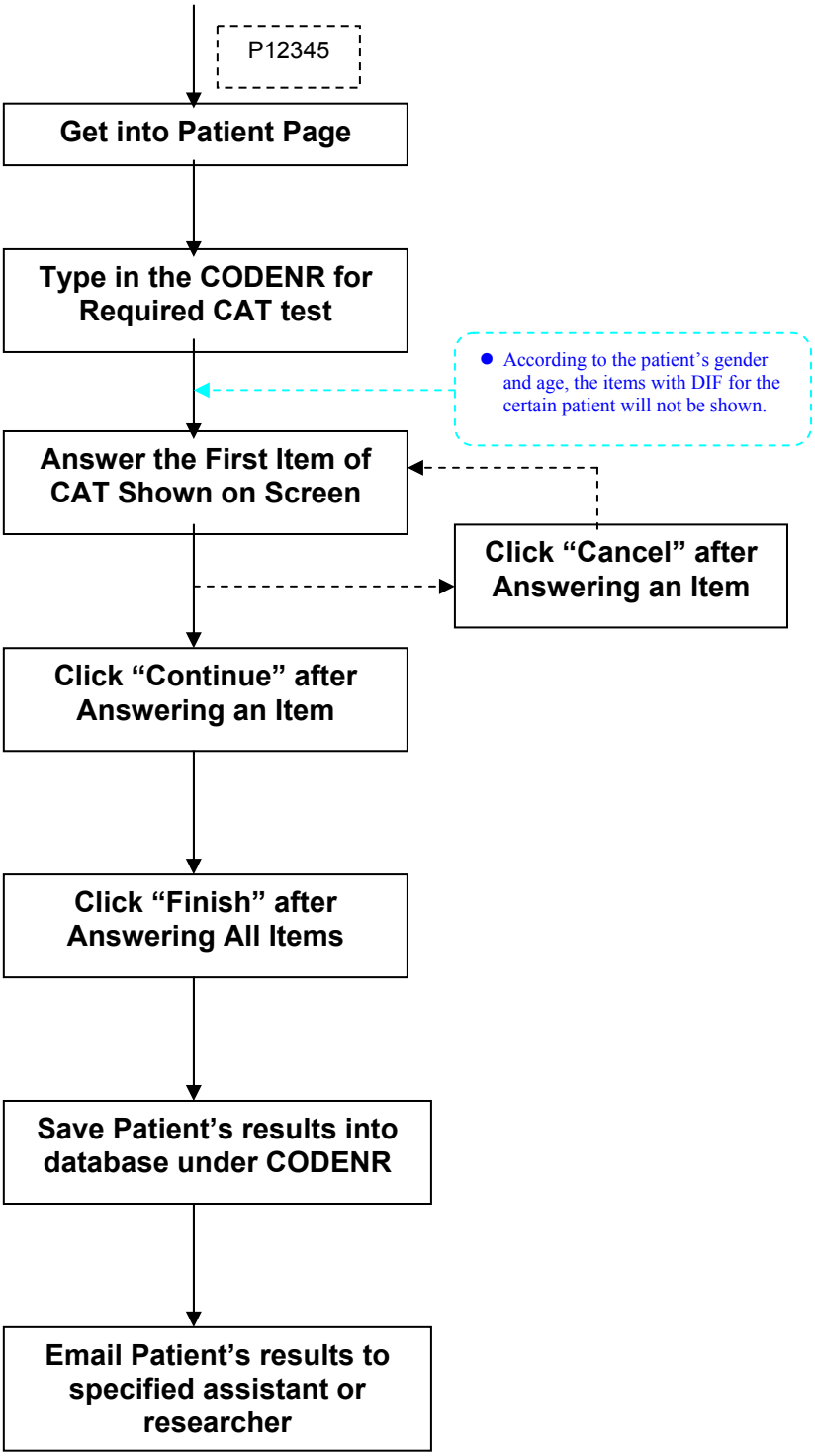
64	...make a bowl of cereal?	-2.28	2.29	getting milk or yoghourt packages from fridge, putting cereal into milk or yoghourt, stirring and mixing, and putting everything back	
65	...sit on the edge of the bed from lying down?	-2.67	1.45	sitting at the edge of bed (with legs out of bed) from lying position in the middle of the bed	
66	...move between two dining chairs?	-2.72	2.35	standing up from a dining chair, walking to another dining chair in the same room, sitting on the other dining chair	check the English version with native speaker
67	...wash and dry your lower body (including feet and legs)?	-2.77	3.03	going to wash-hand basin (on foot or by wheelchair), taking the towel and soap along, washing and drying bottom, legs and feet, sitting is permitted during washing activity, and leaving bathroom	
68	...put on and take off a coat?	-2.85	2.39	clasping and unclasping zipper or buttons, and taking off coats, but hanging clothes on is not required	
69	...wash and dry your face and hands?	-2.96	2.07	going to wash-hand basin (on foot or by wheelchair), taking the towel and soap along, washing and drying face and hands, sitting is permitted during washing activity, and leaving bathroom	
70	...get out of bed into a chair and vice versa?	-2.98	2.26	a chair (wheelchair) is placed beside the bed, getting into the chair from the middle of bed, and vice versa	
71	...go to the toilet?	-3.07	2.95	going to toilet in patient's house, doing/undoing clothing arrangement, standing up and sitting on closestool in patient's own house, using toilet paper	
72	...wash your lower body when taken to the sink?	-3.23	3.14	being taken to wash-hand basin (on foot or by wheelchair), taking the towel and soap along, washing and drying bottom, legs and feet, sitting is permitted during washing activity, and leaving bathroom	
73	...put on and take off a T-shirt?	-3.49	2.69	putting on and taking off T-shirt (undershirt without buttons), clothing is within hand reach	

Appendix B: Workflow of Website for ALDS

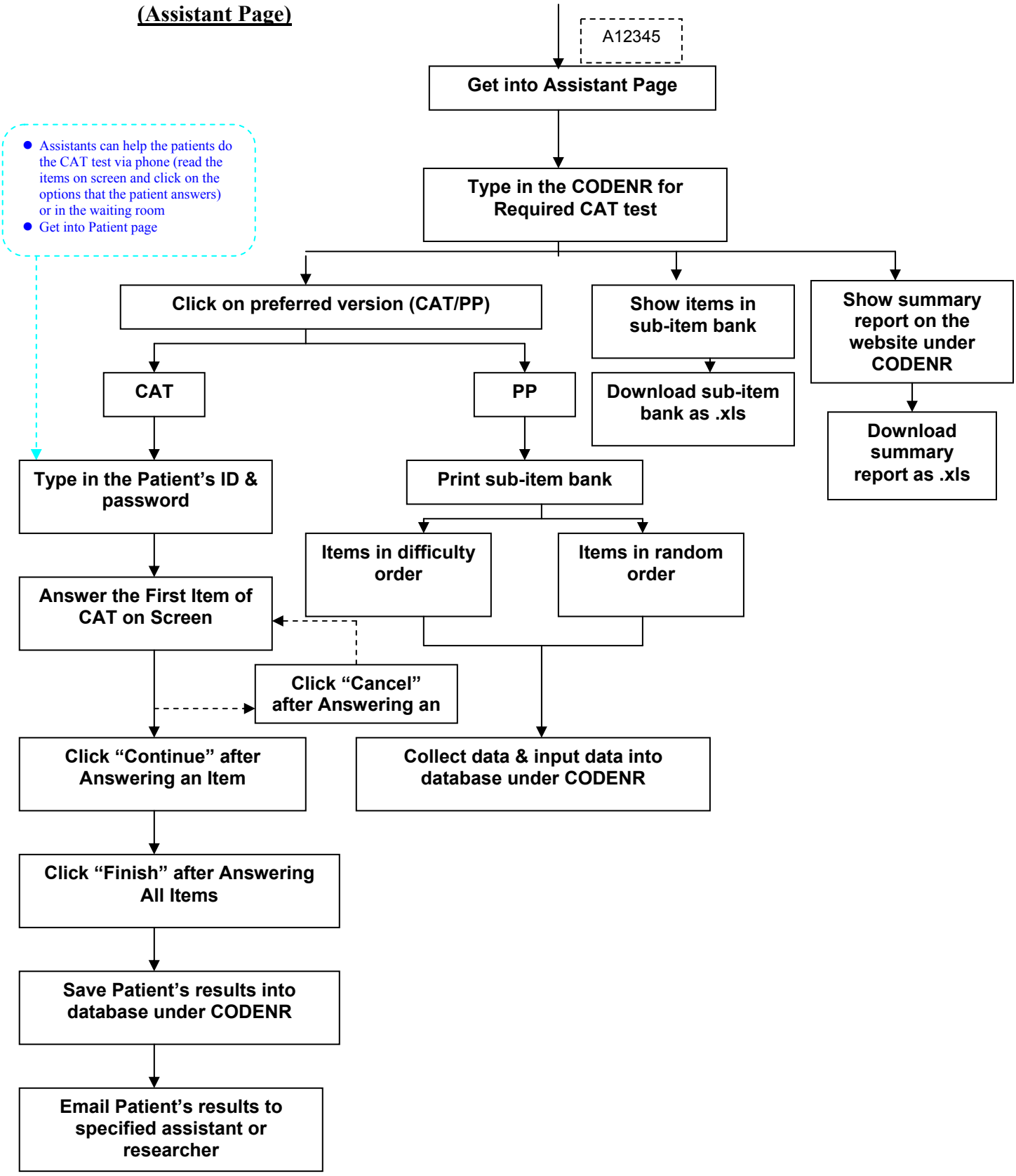
(Registraion Page)

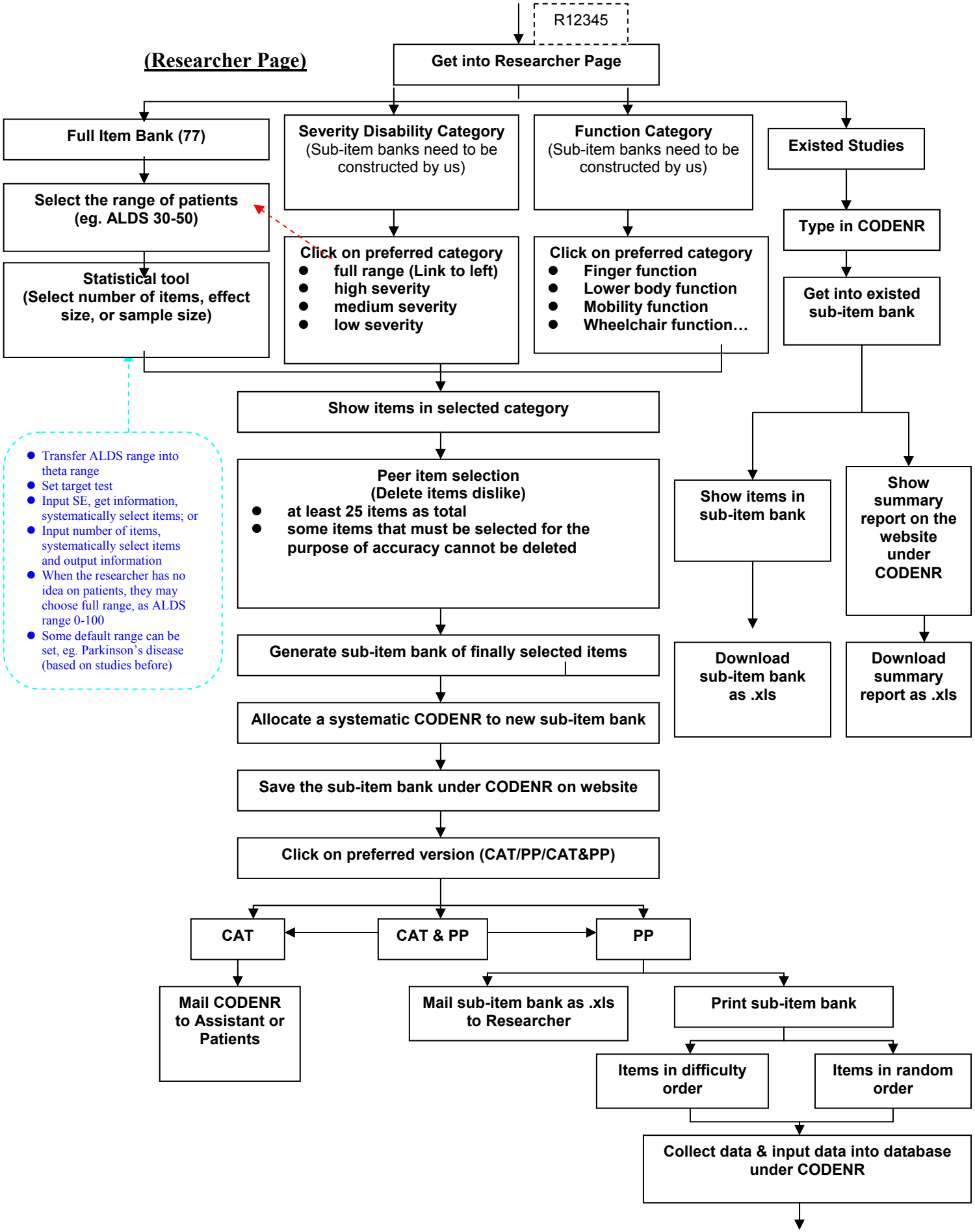


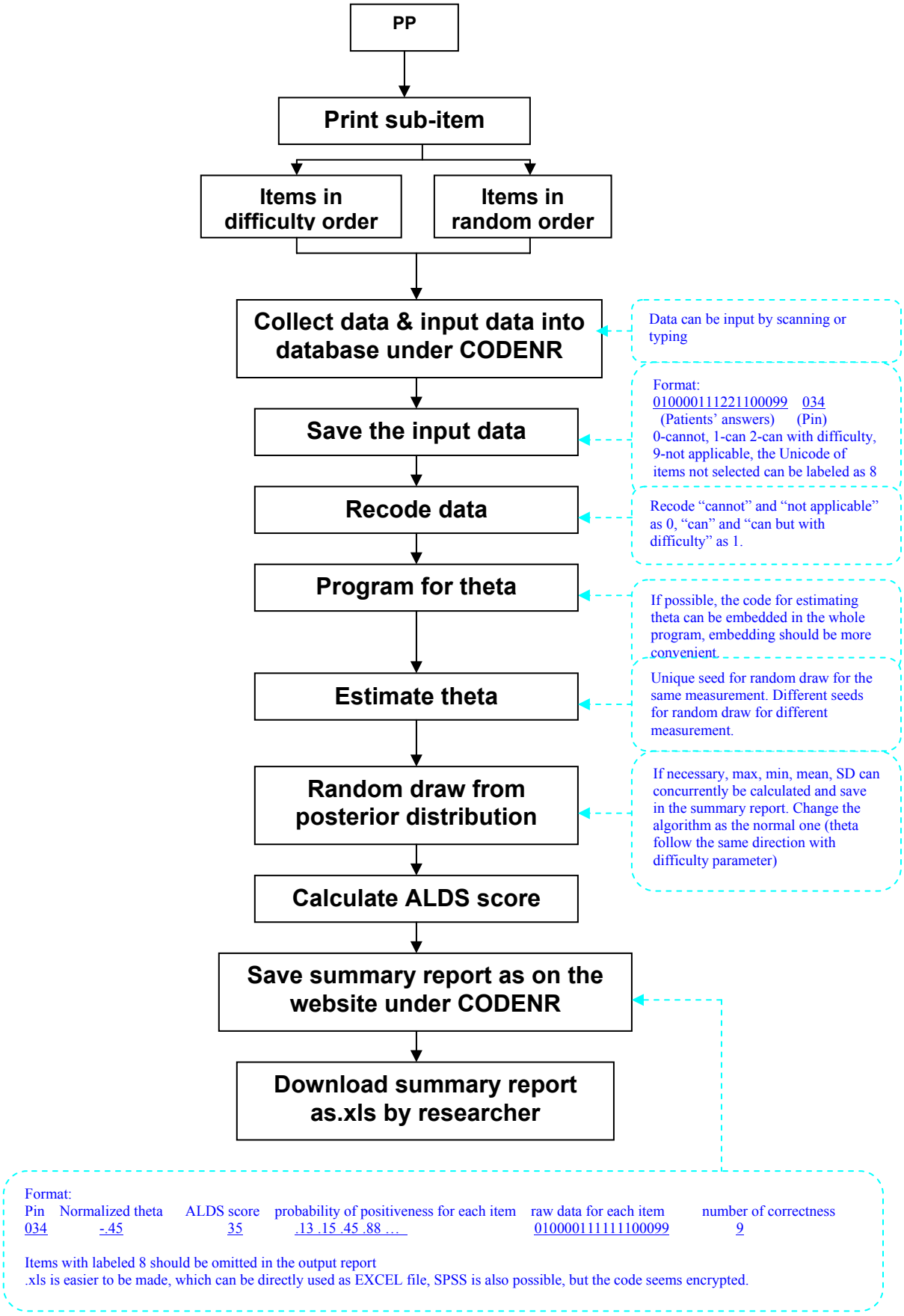
(Patient Page)



(Assistant Page)




R12345
(Researcher Page)
Get into Researcher Page
Full Item Bank (77)
Severity Disability Category
(Sub-item banks need to be constructed by us)
Function Category
(Sub-item banks need to be constructed by us)
Existed Studies
Select the range of patients
(eg. ALDS 30-50)
Statistical tool
(Select number of items, effect size, or sample size)
Click on preferred category
full range (Link to left)
high severity
medium severity
low severity
Click on preferred category
Finger function
Lower body function
Mobility function
Wheelchair function…
Type in CODENR
Get into existed sub-item bank
Transfer ALDS range into theta range
Set target test
Input SE, get information, systematically select items; or
Input number of items, systematically select items and output information
When the researcher has no idea on patients, they may choose full range, as ALDS range 0-100
Some default range can be set, eg. Parkinson's disease (based on studies before)
Show items in selected category
Peer item selection
(Delete items dislike)
at least 25 items as total
some items that must be selected for the purpose of accuracy cannot be deleted
Show items in sub-item bank
Show summary report on the website under CODENR
Download sub-item bank as .xls
Download summary report as .xls
Generate sub-item bank of finally selected items
Allocate a systematic CODENR to new sub-item bank
Save the sub-item bank under CODENR on website
Click on preferred version (CAT/PP/CAT&PP)
CAT
CAT & PP
PP
Mail CODENR to Assistant or Patients
Mail sub-item bank as .xls to Researcher
Print sub-item bank
Items in difficulty order
Items in random order
Collect data & input data into database under CODENR


PP
Print sub-item
Items in difficulty order
Items in random order
Collect data & input data into database under CODENR
Data can be input by scanning or typing
Save the input data
Format:
010000111221100099 034
(Patients' answers) (Pin)
0-cannot, 1-can 2-can with difficulty, 9-not applicable, the Unicode of items not selected can be labeled as 8
Recode data
Recode "cannot" and "not applicable" as 0, "can" and "can but with difficulty" as 1.
Program for theta
If possible, the code for estimating theta can be embedded in the whole program, embedding should be more convenient.
Estimate theta
Unique seed for random draw for the same measurement. Different seeds for random draw for different measurement.
Random draw from posterior distribution
If necessary, max, min, mean, SD can concurrently be calculated and save in the summary report. Change the algorithm as the normal one (theta follow the same direction with difficulty parameter)
Calculate ALDS score
Save summary report as on the website under CODENR
Download summary report as.xls by researcher
Format:
Pin Normalized theta ALDS score probability of positiveness for each item raw data for each item number of correctness
034 -.45 35 .13 .15 .45 .88 … 010000111111100099 9
Items with labeled 8 should be omitted in the output report
.xls is easier to be made, which can be directly used as EXCEL file, SPSS is also possible, but the code seems encrypted.

Appendix C: Setting Standards in Computerized Adaptive Testing in Clinical Measurement: A Field Study

"Cut scores – or what are now more commonly called performance standards – have become more necessary and more consequential." (Cizek & Bunch, 2007) Nowadays, numerous choices exist for how to derive these standards; however, little attention has been paid to the process of setting standards on computerized adaptive tests (CATs) based on item response theory (IRT) models. This lack of attention is unfortunate because CATs are becoming more widely used and setting standards on these tests is typically more challenging than on non-adaptive linear tests. (Sireci, Patelis, Saba, Dillingham, & Rodriguez, 2000)

Taking the advantages in flexibility and accuracy, IRT and CATs have been popularly used in diversified fields, for example clinical trials in functional disability measurements for patients. To simplify the process of tracking patients' performance in different level, a guideline in setting standards on CATs is in an urgent necessity.

No simple and applicable method came until "Borderline Method in CAT" was introduced by Sireci and his team in 2000. This new method was firstly aroused attention when Sireci participated in ACCUPLACER Elementary Algebra Test[6]. In the thesis "Setting Standards on a Computerized Adaptive Placement Examination"[7], he detailed the Borderline Method in CAT and recorded the whole process that this method was applied in ACCUPLACER.

The following part of this thesis introduces the "Borderline Method in CAT" and describes how it can be used in clinical measurement by illustrating the results of AMC Linear Disability Score (ALDS) on 152 patients with brain stroke in Academic Medical Center (AMC) of Amsterdam University in the Netherlands (Weisscher, 2008). The third part investigates statistical significance of the setting standards in CATs by checking the consistence of categorization standards generated by "Borderline Method in CAT" with the traditional clinical measurements modified Rankin score (mRs). Finally the implications of this method for future research and practice in clinical trials are discussed.

Introduction on "Borderline Method in CAT"

The Borderline Method in CAT (BMIC) was firstly experimented in ACCUPLACER test in 1998, when the exams, developed and coordinated by the US College Board, were administered to over two million students. Thirteen mathematics experts (7 males, 6 females) from two- and four- year colleges across the US were recruited to participate in the study as panelists (Sireci et al., 2000).

[6] ACCUPLACER is a series of computerized placement exams used throughout the United States for placing students into three post-secondary courses: introductory algebra, intermediate algebra, or college algebra.

[7] Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA, USA. April 25, 2000.

The purpose of this study was to derive two separate cut-scores on the ACCUPLACER score scale. One cut-score was needed for placement into introductory algebra, while the other cut-score was needed for placement into intermediate algebra. The definition of "borderline" has been discussed by the panelists first. After a trial on first five items, the 13 panelists were asked to finish all 112 items in three categories, labeled as the borderline students "very likely to pass this item", "very likely to fail this item" and "not sure" (Sireci et al., 2000).

The item sorting tasks were actually conducted twice. After the first sorting, every panelist got a feedback, on which the percentage of each item regarded as borderline was recorded. On the base of this feedback, the panelists were asked to make readjust on their sorting. Thus the second sorting was set as the final version.

The items labeled in the median category "not sure" were regarded as the "borderline items" to be focused on. The IRT item difficulty parameters (b-parameters) for these "not sure" items were taken as the cut-score (on the IRT theta score scale) for the borderline student. Thus, this procedure took advantage of the fact that, in an IRT model, the item difficulty parameters and examinee parameters are on the same score scale (Sireci et al., 2000), which implies that the mean value of b-parameter for the "not sure" items could be used as the best estimate of the IRT-score (theta) for the borderline student.

$$\bar{b}_{borderline} = \hat{\theta}_{borderline} = \frac{1}{K}\sum_{k=1}^{K} b_k \qquad \text{(C.1)}$$

where k is the number of "not sure" items and b_k is the difficulty parameter of each item in the "not sure" category.

Based on the products from Formula (C.1), the probability that a borderline student would get correct answer to each item ($P(x=1|\theta)$ is easily derived. To transform the cut-scores from the theta scale to the item score scale, the mean value of b-parameters for the "not sure" items was put into the IRT model equation. For instance, 2PL model (along with a-parameter for item discrimination and b-parameter for item difficulty) was used in the test:

$$P_i(x=1|\theta) = \frac{e^{a_i(\hat{\theta}_{borderline}-b_i)}}{1+e^{a_i(\hat{\theta}_{borderline}-b_i)}} = \frac{e^{a_i(\bar{b}_{borderline}-b_i)}}{1+e^{a_i(\bar{b}_{borderline}-b_i)}} \qquad \text{(C.2)}$$

To sum up these probabilities of correctness across all items in the bank, we can derive the final cut-score as:

$$cutscore = \sum_{i=1}^{n} P_i(x=1|\theta) \qquad \text{(C.3)}$$

where n equals to the total number of items.

For the better understandings among students, the final results are often transferred into a standard score scale in necessity. For example, the sum of probabilities that a

borderline student can get correct answer to each item will be linearly translated into 0-120 in ACCUPLACER score scale (Sireci et al., 2000).

Application of BMIC in Clinical Measurements

Compared to educational tests, IRT and CAT techniques are quite new in clinical trials. With predominance in flexibility and accuracy, IRT aroused great interests in medical studies and rapidly applied in functional disability trials and life quality investigation as well.

In contrast to the sum score methods, IRT measures at the item level. Like the IRT-based educational tests, the ability of patients is expressed on the same linear scale as the item difficulty. It means that each patient can be presented with a smaller selection of items than is possible using sum score based methods. In the implementation of adaptive procedure, more difficult items, (e.g., "cycling for two hours") are presented to less disabled patients and easier items, (e.g., "putting on T-shirt independently") are presented to more severely disabled patients. As a consequence, the application of IRT not only saves the costs, time and energy for both of doctors and patients but also supplements patient-relevant outcomes for doctors' diagnoses.

The AMC Linear Disability Score (ALDS)

The Academic Medical Center (AMC) Linear Disability Score (ALDS) project is a good example in clinical measurements with IRT models. It aimed to construct an item bank to measure the disability status of patients with a broad range of diseases. Once the ALDS item bank has been calibrated in IRT models, it will be used as a basis for computerized adaptive and other innovative testing procedures to assess the functional status of patients in a wide variety of clinical studies (Holman, 2005).

Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific functional health instruments and supplemented by diaries of activities performed by healthy groups (Holman, 2005). The item bank has been calibrated by using the responses from over 4000 patients with a broad range of stable chronic conditions. A total of 190 items were identified and then described in detail at the initial stage (Holman, 2005). But only 77 items are remained currently as applicable ones to be commonly used in measurements, with the range of difficulty level from -3.49 to +3.05 (Weisscher, 2008).

Patients in ALDS project are asked whether they can, rather than do, carry out the activities. The ALDS uses dichotomous frame at present, the two response options are "I can carry out the activity" and "I cannot carry out the activity"[8]. If patients had never had the opportunity to experience an activity, the response of "not applicable"[9] is recorded.

[8] ALDS actually has three response options, "I can", "I can but with difficulty" and "I cannot". For a simple statistical calculation, the two positive options "I can" and "I can but with difficulty" are combined as one option "I can".

[9] Responses in the category "not applicable" are regarded as missing data, which are statistically treated as if the individual items had not been presented to the individual respondent.

The two-parameter logistic IRT model (2PL) was fitted collected data in the calibration phase of the ALDS project. This model was chosen because it allows a more realistic model for the data to be built than the more restrictive one-parameter logistic model. In 2PL model, both of the item difficulty and discrimination degree are required to take into consideration in order to get the "best selected" items for different groups of patients. To make the results easier to interpret, the logit scores are linearly transformed into values between 0 (bottom value) and 100 (ceiling value) after the sum of probabilities that the patient (theta) can give correct answer to each item in the bank is derived (Weisscher, 2008).

$$\hat{T} = \sum_{n=1}^{77} P_n = \frac{e^{a_1(\theta - b_1)}}{1 + e^{a_1(\theta - b_1)}} + \frac{e^{a_2(\theta - b_2)}}{1 + e^{a_2(\theta - b_2)}} + \ldots + \frac{e^{a_n(\theta - b_n)}}{1 + e^{a_n(\theta - b_n)}} \quad \text{(C.4)}$$

$$ALDS = 10 + \frac{80}{77} \cdot \hat{T} \quad \text{(C.5)}$$

where $\hat{T}$ is the best estimate of the sum of probabilities that a certain patient can give positive answer to each item in the bank. The ALDS score in the 77-item bank ranges from 11 to 89.

Brain Stroke Studies with mRs and ALDS

The brain stroke studies were conducted in the neurological department in AMC. A total of 152 patients six months post stroke were consecutively admitted to the stroke unit of AMC between January 2004 and May 2005. They were asked to follow two concurrent studies, modified Rankin Scale (mRs) and the AMC Linear Disability Score (ALDS).

In clinical trials, mRs is a concise index of global disability (Bonita & Beaglehole, 1988) (VanSwieten, Koudstall, Visser, Schouten, & VanGijn, 1988). It is scored as follows:

0= No symptoms;

1= No significant disability despite symptoms, able to carry out all usual duties and activities;

2= Slight disability, unable to carry out all previous activities, but able to look after one's own affairs without assistance;

3= Moderate disability, requiring some help, but able to walk without assistance;

4= Moderately severe disability, unable to walk without assistance and unable to attend to one's own bodily needs without assistance;

5= Severe disability, bedridden, incontinent and requiring constant nursing care and attention;

6= dead.

During the research procedure, three trained nurses evaluated the disability level using first the mRs and next the ALDS by telephone interviews six months after patients were treated at the stroke unit. If subjects were unable to answer the questions due to cognitive problems or severe illness, a relative or caretaker was interviewed on their behalf.

The two instruments have their own features. mRs is very sensitive in stroke studies but dependent on doctors' clinical experience, which implies that there may be some potential risk of inaccuracy when different doctors evaluate the same patient. Unlike mRs, ALDS uses statistical method to analyze patient's responses to items regarding their daily activities, fully independent from doctor's personal assessment. The problem of ALDS is that you cannot prevent patients from "telling lies". Once the patient gives an untrue response, ALDS has to fail in giving out a precise evaluation.

However, mRs and ALDS do need mutually in a completed assessment on patients. The former provides the doctor's clinical experience while the later supplements the patient's responses on daily life. It should be excellent if they could be combined or linked together. How can we realize it? My idea is that since medical trials generally treat patients by categorizing them into different disability levels, as similar as cut-scores in educational tests, setting standards seems to be a possible linkage to study further.

Cut-off Points in mRs

The argument of cut-off points between good and poor outcomes in stroke trials of mRs has been lasting for a couple of years. Dr. Weisscher once made a statistics on the literatures of MEDLINE from January 2005 through June 2007 to investigate which mRs cut-off points were set as primary or secondary endpoint in randomized clinical trials. In the total related 20 articles, 7 studies used mRs 0-1 as favorable outcome and 8 studies used mRs 0-2 (2008:96). This result implies that both of mRs 0-1 and 0-2 are generally accepted as cutoff points in stroke trials.

But as for the question when mRs 0-1 or 0-2 should be used, Weisscher gave her conclusions after comparing mRs with ALDS, that if good outcome is defined as the ability to perform outdoor activities mRs 0-1 should be chosen; if complex ADL (activities of daily life) are considered as good outcome, mRs 0-2 is the outcome measure of choice (2008:94). It can be illustrated as the following table.

Table C.1 reports the post stroke mRs and ALDS of 152 patients. Suppose we firstly take good outcome as mRs 0-1, 33 patients (1+22=33) can be assessed "good", but the gap of ALDS mean value between groups with mRs 1 and 2 is 12.6 (87.5-74.9=12.6), over two times less than the gap between groups with mRs 2 and 3, 27.4 (74.9-47.5=27.4). This result suggests that the cut-off point at mRs 0-1 plays a weaker role than mRs 0-2 in distinguishing the good and poor outcomes in brain stroke case. Meanwhile, since ALDS project focus on responses of patients' capability in activities of daily life, it should be more suitable to use mRs 0-2 as the cut-off point (Weisscher, 2008).

Table C.1 mRs and ALDS Scores of 6 Months Post Stroke (n=152)

Number of patients	mRs	ALDS score mean (SD)
1	0	89
22	1	87.5 (2.1)
33	2	74.9 (8.6)
32	3	47.5 (14.1)
28	4	24.5 (14.9)
3	5	12.0 (1.0)
33	6	0

Setting Standards in ALDS

mRs use two categories, good (mRs 0-2) and poor (mRs 3-6)[10] outcome to classify patients, but how can we set the corresponding standards for ALDS in order to link the two instruments? BMIC seems to be a good choice.

In order to avoid mixing with mRs evaluation, other three trained nurses from stroke units in AMC, who had not involved in the previous studies were asked to divide the 26 items[11] that have been presented to post stroke patients into three groups according to difficulty levels for the stroke patients. It suggests that two separate cut-scores on the ALDS score scale need to be derived, one for placement at basic-moderate level, and the other for moderate-difficult level. The 26 items were not ordered as difficulty level when they were presented to the nurses.

As Table C.2 shows, in the nurses' eyes, item 1 to 6 belong to the difficult part, which implies only less disabled patients can handle with these items; item 7 to 18 are distributed to moderate level; and item 19 to 26 are remained in the basic level. Hence, the cut-off point between difficult and moderate items should be at the mid of item 6 and 7; and that between moderate items and basic ones should be at the mid of item 18 and 19.

However, meanwhile, we noticed that the 26 items have been listed in descending difficulty order. The closer the items locate to borderlines, the higher uncertainty they belong to their appointed category. For example, item 1 (ride a bike for at least 2 hours) is distinctly more difficult to be accomplished than item 6 (walk up a hill or high bridge); consequently, item 1 is much more surely placed in the difficult item group than item 6. As a result, although the nurses were not asked to directly sort out the "not sure" items, the items whose difficulty parameters were close to the cut-off points could be regarded as "not sure" to some extent.

[10] In the 158-patient brain stroke study, 33 patients (mRs 6) are dead; and only 3 patients were assessed as mRs 5 and their probabilities of positive responses are lower than 5% to each item, which is not statistically significant. Consequently, the present paper uses mRs 3-4 to make further studies instead of mRs 3-6.

[11] A total of 26 items are presented in two booklets, each of which consists 18 items, used for the less disabled patients and severely disabled patients respectively.

Table C.2 Three Categories of Items on Difficulty Levels in Stroke Trials

Item Content	b	a	ALDS
Difficult Items			
1. ride a bike for at least 2 hours	3.05	2.45	89
2. carry a bag of shopping	2.14	2.70	85
3. go for a walk in the woods	1.50	2.56	81
4. travel by local bus or tram	1.23	2.86	78
5. walk for more than 15 minutes	0.82	2.13	74
6. walk up a hill or high bridge	0.78	1.99	73
Moderate Items			
7. cut your toe nails	0.66	1.63	72
8. stand for 10 minutes	0.53	1.83	70
9. walk up a flight of stairs	0.19	2.19	65
10. walk down a flight of stairs	0.02	2.62	62
11. go for a short walk (15 mins)	-0.07	2.06	60
12. change the sheets on a bed	-0.21	1.56	58
13. fetch a few things from the	-0.29	2.53	56
14. have a shower and wash your	-0.66	1.95	50
15. pick something up from the	-1.15	2.02	42
16. get in and out of a car	-1.34	2.17	39
17. peel and core an apple	-1.49	1.20	37
18. prepare breakfast or lunch	-1.52	2.27	36
Basic Items			
19. eat a meal at the table	-1.79	1.35	32
20. sit up from lying in bed	-1.95	1.25	30
21. put long trousers on	-2.38	2.74	24
22. sit on the edge of a bed from	-2.67	1.45	21
23. put on and take off a coat	-2.86	2.39	19
24. get out of bed into a chair	-2.99	2.26	18
25. go to the toilet	-3.08	2.95	17
26. wash your lower body	-3.24	3.14	15

Suppose we include two items above and two below borderlines to consist of "not sure" group. The moderate-difficult borderline group now includes 4 items, from item 5 to 8. Taking the item parameters in Table C.2 into Formula (C.1), we can get:

$$\bar{b}_{borderline} = \hat{\theta}_{borderline} = \frac{1}{4}(0.82 + 0.78 + 0.66 + 0.53) = 0.6975$$

The mean value of item difficulty in borderline equals to the borderline patient's ability. Following Formula (C.2) and (C.3), we get the sum of probabilities that the borderline patient can give positive responses to each item in the 26-item bank as 19.21. The positiveness percentage can be transferred as 19.21/26=73.9%. It indicates that the patients can be assessed as high ability (less disabled) only when he can give at least 19.21 (73.9%) positive responses to the 26 items.

Following the same way, the basic-moderate borderline group ranges from item 17 to 20. The mean value of borderline items difficulty is reckoned as -1.6875, as the same as the ability of the borderline patient. Taking this borderline value into Formula (C.2) and (C.3), we can get the sum of probabilities of positive responses given by the borderline patient at basic-moderate level as 8.43, implying that the patients who give at least 8.43 positive answers, accounting for 32.43% (8.43/26=32.43%) among the 26 items can be regarded as moderately disabled. Otherwise, those who give less than 8.43 positive responses, below 32.43%, would be remained in basic level, namely severely disabled in medical assessment.

To sum up, BMIC helps us set two separate cut-scores at 19.21 (73.9%) and 8.43 (32.43%) respectively for difficult and moderate item groups, corresponding to slightly and moderately disabled patients.

mRs Cut-off Points Linked with ALDS Cut-Scores

On account that every patient in the studies had been evaluated by both of the two instruments, mRs and ALDS, we can make a simple calculation on the probability that patients with different mRs level can give positive answers to each item. As Table C.3 shown, the probability that patients with good outcome mRs 0-2 and patients with poor outcome mRs 3-4 are able to perform 26 activities were listed in decreasing difficulty order. For example, the patients who are scaled as mRs 0-2 has 19% probability to give positive response to the most difficult item No.1 and the probability rises to over 95% from item 14 (have a shower and wash hair). As for the patients in mRs 3-4, the probability to give positive response to item 1 is less than 1%, and the probability to handle the easiest activity (item 26: wash your lower body) is just above 50%.

If we take the cut-scores of ALDS, just calculated above, into consideration, it is very interesting to find that the probabilities of mRs in Table C.3 are amazingly consistent with ALDS cut-scores. It is noticeable that the probabilities of positive responses given by patients with mRs 0-2 are around 74%-79% in borderline items from item 5 to 8, very close to ALDS cut-score 73.9%. If we calculate the mean value of probabilities of positive responses based on mRs from item 5 to 8, we can find the gap (absolute value) between mRs and ALDS is as small as 2.35%. The same case happens at ALDS basic-moderate level with mRs 3-4. The gap between the cut-off points of these two instruments is even smaller, at 0.18%.

The results explain that the ALDS instrument, absolutely independent from doctor's experience, can provide a close standard setting as mRs. It suggests that patients who can give more than 73.9% positive responses in the 26 item bank could be assessed as mRs 0-2 according to the doctor's experience to a large extent. And

those who can give positive responses between 32.43% and 73.9% could be defined as moderately disabled in ALDS and mRs 3-4 in clinical diagnosis.

Statistical Analysis on Setting Standards of mRs and ALDS

Does the Number of Borderline Items Impact Standards Setting?

Some researcher worried that if the borderline items were not arbitrarily set as four, for example, reduced borderline items to 2 or expanded to 8, can mRs and ALDS still keep the association. Simulation experiment is used as following.

In order to test the impact of borderline item number, 2, 4, 6, 8, 10 and 12 items, symmetrically selected on the borderline, were included into the borderline group in succession.

Table C.4 and C.5 recorded the results of simulation experiment on mRs and ALDS standards. Firstly, focus on Table C.4. The first column in the table indicates the number of items including in the "not sure" group. The inclusion range of items in the bank is shown on the second column. The data in the third to fifth columns are results generated by BMIC. For example, the third line of Table 4 expatiates that the borderline group between difficult and moderate level consists of 6 items, i.e. item 4 to item 9. Based on the BMIC, we derive the mean value of difficulty in borderline group as 0.70 with standard deviation 0.34. Since the borderline b-difficulty parameter is the same as borderline theta, following formula (C.2) and (C.3), we find that patients who can give more than 19.23 (73.98%) positive responses can be placed into slightly disabled group (difficult item level) when 6 items are used in the borderline group. The sixth column indicates the average probabilities that patients with mRs 0-2 may give positive responses to the six borderline items. The next two columns are products of column 5 minus column 6 and their absolute values. The last column records the linearly transformed ALDS scores, which are more convenient for doctors to interpret.

If we take a careful look at column 8, the absolute value of the gap between ALDS and mRs, it is clearly to find that the change is very small. No matter the inclusion in borderline group ranges from 2 items or 12 items, the gap between two instruments almost keeps unchanged, especially in the comparison between ALDS basic-moderate level and mRs 3-4 shown in Table C.5. The absolute value of gap in Table C.5 minimizes at 0.18% and maximizes at 2.81%; the change among the six values extends only 2 percentage points. The results seemingly prove that the consistence of setting standards in two instruments is little impacted by the number of borderline items.

Table C.3 Probabilities that Patients with mRs 0-2 and mRs 3-4 are Able to Perform Activities with Decreasing Difficulty (n=116)

Item Content	mRs 0-2	mRs 3-4	ALDS
Difficult Items			
1. ride a bike for at least 2 hours	19%	…	
2. carry a bag of shopping	42%	…	
3. go for a walk in the woods	58%	1%	
4. travel by local bus or tram	65%	1%	
5. walk for more than 15 minutes	**74%**	3%	
6. walk up a hill or high bridge	**76%**	4%	
Moderate Items			**73.9%**
7. cut your toe nails	**76%**	6%	
8. stand for 10 minutes	**79%**	6%	
9. walk up a flight of stairs	86%	8%	
10. walk down a flight of stairs	91%	9%	
11. go for a short walk (15 mins)	90%	11%	
12. change the sheets on a bed	89%	13%	
13. fetch a few things from the	94%	13%	
14. have a shower and wash your	96%	18%	
15. pick something up from the	98%	25%	
16. get in and out of a car	99%	27%	
17. peel and core an apple	96%	**30%**	
18. prepare breakfast or lunch	99%	**30%**	
Basic Items			**32.43%**
19. eat a meal at the table	98%	**34%**	
20. sit up from lying in bed	98%	**35%**	
21. put long trousers on	…	42%	
22. sit on the edge of a bed from	…	45%	
23. put on and take off a coat	…	48%	
24. get out of bed into a chair	…	49%	
25. go to the toilet	…	50%	
26. wash your lower body	…	52%	

Note. The data exclude the patients with mRs 5-6, because only 3 patients were assessed as mRs 5 and the probabilities of positive responses are lower than 5% to each item, which is not statistically significant in the comparison. The patients with mRs 6 deceased, which cannot provide any information on probabilities of positive responses.

Table C.4 Comparison of Standards Setting between mRs and ALDS on Borderline Items between Difficult and Moderate Level

Borderline Item Number	Item Scope	$\bar{b}_{borderline}$ (SD)	ALDS Positive Responses	ALDS Positive Probabilities	mRs 0-2 Positive Probabilities	ALDS-mRs	|ALDS-mRs|	ALDS Score
2	6-7	0.72 (0.08)	19.31	74.28%	76.00%	-1.72%	1.72%	72.50
4	5-8	0.70 (0.13)	19.21	73.90%	76.25%	-2.35%	2.35%	72.25
6	4-9	0.70 (0.34)	19.23	73.98%	76.00%	-2.02%	2.02%	72.00
8	3-10	0.72 (0.49)	19.30	74.22%	75.63%	-1.41%	1.41%	71.90
10	2-11	0.78 (0.69)	19.58	75.29%	73.70%	1.59%	1.59%	72.00
12	1-12	0.89 (0.97)	20.04	77.06%	70.42%	6.82%	6.82%	72.25

Table C.5 Comparison of Standards Setting between mRs and ALDS on Borderline Items between Moderate and Basic Level

Borderline Item Number	Item Scope	$\bar{b}_{borderline}$ (SD)	ALDS Positive Responses	ALDS Positive Probabilities	mRs 3-4 Positive Probabilities	ALDS-mRs	|ALDS-mRs|	ALDS Score
2	18-19	-1.66 (0.19)	8.55	32.90%	32.00%	0.90%	0.90%	34.00
4	17-20	-1.69 (0.22)	8.43	32.43%	32.25%	0.18%	0.18%	33.75
6	16-21	-1.75 (0.38)	8.21	31.59%	33.00%	-1.41%	1.41%	33.00
8	15-22	-1.79 (0.52)	8.06	30.99%	33.50%	-2.51%	2.51%	32.63
10	14-23	-1.78 (0.69)	8.08	31.06%	33.40%	-2.34%	2.34%	33.00
12	13-24	-1.76 (0.85)	8.16	31.40%	33.00%	-1.60%	1.60%	33.67

Statistical Test on Categorization Standards of ALDS and mRs

Although the two instruments have very small differences in setting standards, we yet cannot assert that they can provide exactly the same results until statistical test is conducted. Because of a small sample size in the simulation experiment, only 6 pairs (2, 4, 6, 8, 10, 12 items) included in borderline group, normal distribution can not be used. As a result, Wilcoxon Signed Rank Test, a non-parameter statistical test, was chosen (Moore & McCabe, 2003).

Let's take the borderline group between difficult and moderate level as an example. In this case, we would like to test the hypotheses:

H_0: Positive probabilities in borderline group have the same distribution for both ALDS and mRs 0-2.

H_a: Positive probabilities in borderline group have not the same distribution for both ALDS and mRs 0-2.

Because this is a matched pairs design, we base our inference on the differences, namely the gap between positive probabilities of ALDS (column 5 in Table C.4) and mRs 0-2 (column 6 in Table C.4). Getting the data from the column 8 in Table C.4, we rearrange the absolute value of differences between these two instruments in increasing order and assign ranks, as Table C.6 shown.

Table C.6 Absolute Value of Differences between ALDS and mRs 0-2 and Assigning Rank in Increasing Order

Absolute Value	1.41%	**1.59%**	1.72%	2.02%	2.35%	**6.82%**
Rank	1	**2**	3	4	5	**6**

Note. The bold-faced data are the values that are originally positive.

Keeping track of the values that were originally positive, 1.59% ranking the second position and 6.82% ranking the top; we can obtain the Wilcoxon signed rank statistics as $W^+ = 2 + 6 = 8$, which has the mean and standard deviation as

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{6(6+1)}{4} = 10.5 \tag{C.6}$$

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{4}} = \sqrt{\frac{6(6+1)(2\times 6+2)}{4}} = 11.68 \tag{C.7}$$

We can see that our observed value 8 is slightly smaller than this mean 10.5, which gives an initial impression that the alternative hypothesis may not be significant. Moreover, the one-sided P-value is calculated as $P(W^+ \le 8) = P(W^+ \ge 13) = 0.4168$, with the continuity correction of 0.5 at $P(W^+ \ge 12.5) = 0.4325$. Due to an insignificant P-value, we have to reject the alternative hypothesis and keep the null hypothesis. It implies that there is no evidence for ALDS standard more restrictive than mRs 0-2,

suggesting that setting standards at ALDS moderate-difficult level and mRs 0-2 has no difference in statistics.

The same test can be conducted again on standards setting at ALDS moderate-basic level and mRs 3-4. Suppose we have the similar hypothesis as the previous one:

H_0: Positive probabilities in borderline group have the same distribution for both ALDS and mRs 3-4.

H_a: Positive probabilities in borderline group have not the same distribution for both ALDS and mRs 3-4.

The absolute values of differences between ALDS and mRs 3-4 are listed in Table 7 as increasing ranking order. Keeping the track of values that were originally positive, we get the Wilcoxon rank sum $W^+ = 1 + 2 = 3$. Although the observed value is a bit far from the mean 10.5, the one-sided P-value, $P(W^+ \le 3) = P(W^+ \ge 18) = 0.2611$ with continuity correction of 0.5 as $P(W^+ \ge 17.5) = 0.2776$ is still not significant to accept the alternative hypothesis.

Table C.7 Absolute Value of Differences between ALDS and mRs 0-2 and Assigning Rank in Increasing Order

Absolute Value	**0.18%**	**0.90%**	1.41%	1.60%	2.34%	2.51%
Rank	**1**	**2**	3	4	5	6

Note. The bold-faced data are the values that are originally positive.

In a word, according to the two statistical tests on ALDS borderline groups and mRs, we have reasons to believe that ALDS standards setting is statistically consistent with medical score mRs. This conclusion is supposed to gain attention because it links standards setting in two instruments: mRs, based on doctor's experience, and ALDS, based on patient's responses to the daily activities. In addition, on account of the consistent standards setting with mRs, ALDS that is mainly constructed on IRT models will be definitely more applicable in medical studies than before.

Discussion

Regarding the simulation experiment, more issues were found during the study process. Besides the consistence of setting standards in two instruments, the major factors that may impact the differences between ALDS and mRs in borderline groups were also investigated.

Firstly, because the borderline items were selected symmetrically around the two borderlines, the standard deviation of item difficulty value in borderline groups is definitely extended when the borderline item number increased. In the third column of Table C.4 and C.5, the value of SD in blankets exactly follows in an ascending order. In the moderate-difficult borderline group, when the inclusion is only two items, item 6 and 7, the standard deviation of item difficulty is the least, 0.08. But it increased to

0.97 when 12 items were included into the borderline group. The correlation coefficient between SD and absolute value of gap is calculated as 0.671 and 0.680 for the two borderline levels respectively, implying these two factors have positive associations.

Secondly, I also find that when 8 items (item 3 to 10) included in borderline group of moderate-difficult level, the gap between the two instruments is the lowest. Unlike the above level, the gap between ALDS and mRS minimizes when 4 items (item 17-20) are included in the basic-moderate borderline group. It means that if we choose item 3 to 10 and item 17 to 20 as the borderline items, the differences between these two instruments of setting standards should be the least. Checking data in the last column in Table C.4 and C.5, we can see that the corresponding ALDS scores for the two borderlines are 71.90 and 33.75 respectively, which should be used in setting standard for ALDS in order to get the precise standard for categorization with smallest difference between two instruments.

The linkage between ALDS and mRs in setting standard now can be simplified as Table C.8.

Table C.8 Antithesis on Standards Setting between ALDS and mRs in Brain Stroke Studies

mRs	ALDS (X)	Diagnosis on Patients
0-2	$71.90 \leq X < 100$	Slightly Disabled
3-4	$33.75 \leq X < 71.90$	Moderately Disabled
5	$0 < X < 33.75$	Severely Disabled
6	0	Dead

However, generally speaking, doctors are used to measuring patients' functional disability by more than three categories, such as mRs having 7 levels as total (0-6). But the setting standards illustrated in the present paper classified patients into only 3 disability levels, far from the doctor's demand. Thus, a further study needs to follow up to find a better way in setting 7 standards in ALDS project corresponding to the 7 levels set in mRs. At that time, ALDS would not only be the additional tool for doctors to get patient-relevant outcomes but also easily interpreted to medical language.

Thirdly, the limitation in ALDS is also obvious. Because BMIC was applied in ALDS standards setting process, a problem that criteria for borderline items could be various by different doctors was hardly prevented. We could not be sure that the borderline items remained the same when the nurses were changed. The instability of item sorting makes inevitable errors. How to minimize the errors and keep the standards setting as precise as possible is another question worthy of further pursuit.

Conclusion

The Borderline Method in CAT, also known as Item Sorting Method (Cizek, 2001), a shortcut for gathering Angoff-type item rating data, is easily operated in CAT

environment. Setting standards in CATs for medical trials is a breakthrough to a large extent, because it links standards setting in two instruments: mRs, based on doctor's experience, and ALDS, based on patient's responses to the daily activities. On the base of above analysis on ALDS project in brain stroke studies, we can draw the following conclusions:

(1) BMIC, explored in educational tests can be applied in medical trials with computerized adaptive test environment, which makes it possible to link the statistical data with doctors' clinical experience.

(2) BMIC, also called item sorting method, required panelists to review all the items in an item bank and sort them into three categories, the "not sure" group is regard as borderline group, attracting most attention to set standards by IRT models.

(3) In the AMC brain stroke case, BMIC was applied to set two separate cut-scores for basic-moderate level and moderate-difficult level in ALDS project in order to link with clinical instrument mRs.

(4) In brain stroke studies, with the inclusion of four items in the borderline group, symmetrically selected around borderline, ALDS cut-scores are 73.90% and 32.43% for difficult and moderate levels, as similar as probabilities that patient with mRs0-2 (76.25%) and mRs 3-4 (32.25%) give positive responses to all the items in the bank.

(5) In brain stroke studies, in the statistical respect, the number of items included in the borderline group does not impact the difference between standards setting of the two instruments, ALDS and mRs. The difference between ALDS and mRs is insignificant in statistical test, implying the standards setting in the two instruments are parallel.

(6) The standard deviation of item difficulty in the borderline group expands when the borderline item number increased. In the brain stroke studies, the standard deviation of borderline item difficulty has a high positive correlation with the absolute value of differences between standards setting in ALDS and mRs.

(7) ALDS score and mRs can be simply translated to each other with the minimal difference between these two instruments when 10 items are included in moderate-difficult borderline group and 4 items in basic-moderate borderline group. In brain stroke disease, the patient who gets ALDS score between 71.90 and 100 could be evaluated as slightly disabled with mRs 0-2. Those whose ALDS score ranks between 33.75 and 71.90 could be treated as moderately disabled with mRs 3-4. The severely disabled patient features in ALDS score below 33.75 and mRs 5; and vice versa.

(8) Further studies still need to do to realize a complete parallel standards setting between ALDS and mRs.

References

Bonita, R., & Beaglehole, R. (1988). Modification of Rankin Scale: Recovery of Motor Function after Stroke. *Stroke, 19*, 1497-1500.

Cizek, G. J. (2001). *Setting performance standards : concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting : a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif.: Sage Publications.

Holman, R. (2005). *Item Response Theory in Clinical Outcome Measurement.* Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the practice of statistics* (4th ed.). New York: W.H. Freeman and Co.

Sireci, S. G., Patelis, T., Saba, R., Dillingham, A. M., & Rodriguez, G. (2000). *Setting Standards on a Computerized Adaptive Placement Examination*. Paper presented at the USA Annual Meeting of the National Coluncil on Measurement in Education.

VanSwieten, J., Koudstall, P., Visser, M., Schouten, H., & VanGijn, J. (1988). Interobserver Agreement for the Assesssment of Handicap in Stroke Patients. *Stroke, 19*, 604-607.

Weisscher, N. (2008). *The AMC Linear Disability Score (ALDS): Measuring Disability in Clinical Studies.* Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.