RASCH MODEL ACCOUNTING FOR MEASUREMENT ERRORS

by

Weining Y. Volinn

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

Department of Family and Preventive Medicine

The University of Utah

August 2012

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Weining Y. Volinn**

has been approved by the following supervisory committee members:

Christina A. Porucznik , Chair **May 21, 2012**
Date Approved

Stephen C. Alder , Member **May 21, 2012**
Date Approved

Jaewhan Kim , Member **May 21, 2012**
Date Approved

Richard Holubkov , Member **May 21, 2012**
Date Approved

Xiaoming Sheng , Member **May 21, 2012**
Date Approved

and by **Stephen C. Alder** , Chair of

the Department of **Family and Preventive Medicine**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

This dissertation focuses on two prominent issues encountered in analyzing questionnaire data. First, the summed score of all individual questions in a questionnaire is characteristically used as a measure of disease severity, even though it often does not have interval properties. Second, measurement errors exist whenever there is measurement. For example, questionnaires, composed of questions with predefined response categories, force patients to make choices.

The data for the dissertation were binary response data from Simple Shoulder Test questionnaire (SST). A Rasch model was used to estimate Rasch scores. The minimum clinical important difference (MCID) in Rasch scores was then compared with the MCID in summed scores. MCID was defined as the statistically significant difference in change from baseline between patient groups (No Change and Minimal Improvement). Two anchored questions were used to delineate patient groups. In Rasch scores, conclusions about the MCID reached through both questions supported MCID in summed scores.

To address issues of nonlinearity and measurement errors, I constructed a Rasch model accounting for measurement errors and mapped out Markov Chain Monte Carlo (MCMC) steps to estimate model parameters. The optimal setting of factors affecting MCMC implementation was identified.

To evaluate the effect of measurement errors, I applied Rasch model accounting for measurement errors to SST data and obtained Rasch scores accounting for measurement errors (i.e., MCMC Rasch scores). MCID analysis was performed on MCMC Rasch scores. I found that MCID is unascertainable through MCMC Rasch scores. Inconsistent MCID findings in these two types of Rasch scores may be due to bias of estimates of Rasch SST when measurement errors are left unconsidered.

In sum, Rasch scores accounting for measurement errors: 1) provide more accurate estimates for person abilities indicated by mean square errors; 2) provide unequal spaces between scores compared with summed scores, which may more accurately describe patients' experiences; 3) provide estimates corresponding to extreme summed scores with reasonable variances, which remain inestimable in the classical Rasch model; 4) may be used as a continuous variable, unlike summed and classical Rasch scores, because the measurement errors were treated as random effects.

To my beloved husband Ernest P. Volinn

and my parents Zhijian Liu and Xianjie Yang

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

ACKNOWLEDGMENTS

I would first like to thank and acknowledge Dr. Xiaoming Sheng, a mentor and expert on the topic of my dissertation. As a member of my committee he started me down this road, and continued to walk alongside of me every step of the way. I would not have succeeded without him. I would also like to thank Dr. Christina Porucznik, my committee chair, for giving me the opportunity to pursue a topic truly important to me and for supporting me during the time of my dissertation in countless ways. She guided me to a dataset that opened up the dissertation and proved to be invaluable to me. Originally, this was Dr. Robert Tashjian's dataset, and I would like to acknowledge him for allowing me to use it. I would like to acknowledge the dedication and contribution of my other committee members: Dr. Stephen Alder, Dr. Richard Holubkov, and Dr. Jaewhan Kim. I cannot express how much I appreciate their mentoring and guidance. I am grateful for the support from my company, Watson Laboratories, in terms of flexibility and financial support. In particular, I would like to thank Heather Thomas and Gary Hoel of Watson.

CHAPTER 1

INTRODUCTION

The convergence of a number of trends in society has made the evaluation of the health care treatment outcomes a pressing issue. The costs of health care have risen rapidly, and the need to make judicious cuts in health care has become apparent. There is also a growing awareness among patients as well as physicians that health care is not always efficacious. Increasingly, the question is the following: Given a particular treatment, what are the outcomes?

Outcomes measurement, however, relies upon the choice of a particular outcome measure or measures. A key consideration in selecting them is what matters to patients. From the patient’s point of view, outcomes that mainly pertain to laboratory or clinical findings may be of lesser importance than outcomes they themselves may report. For many diseases and conditions, most relevant to them is that their pain improves and they are capable of carrying on with daily activities and return to work.[1] For this reason, especially since the 1980s, outcome evaluators have increasingly adopted measures of patient reported outcomes.[2]

Patient reported measures have been developed to evaluate treatment of diverse conditions. For example, the incontinence impact questionnaire (IIQ) is used to evaluate severity of urinary incontinence and posttreatment relief of symptoms.[3] Another example of measure based on patients' reports, and the one primarily to be examined here, is the Simple Shoulder Test (SST) used to evaluate shoulder functions.[4]

Questionnaires, however, inevitably lead to measurement errors, which potentially undermine an analysis or distort results. They may result in attenuation to the null, loss of power, and bias of estimates. Especially, in questionnaires, we necessarily narrow the choices with which subjects may respond. For example, in binary response questions, there are only two categories for them to choose (e.g., 1=yes or 2=no), which is exemplified in the SST. In other words, questionnaires, particularly those consisting of binary response categories, force subjects to make a choice among predefined response categories, even though their true choice may fall in between the available values of categories. This suggests that measurement errors (with respect to individual items as well as their overall summed score) are much more than simple recording or instrument error. They encompass many different sources of variability.[5] In this dissertation, I focused on the measurement errors in questionnaires with binary response categories.

In questionnaires, the summed score is conventionally used as a measurement of disease severity. Problems that may ensue from this are the following: (1) there are no natural numerical upper or lower limits to health status (one conceivable exception with regard to lower limits is "worse than dead"), and for this reason, a zero value does not have an inherent meaning; (2) furthermore, psychometric instruments (i.e., questionnaires) do not necessarily have interval characteristics (i.e., are not scaled in a linear fashion). In other words, one cannot assume that the two point difference between scores of 10 and 12 is equal (in a patient's perception or other respects) to the same two point difference between scores of 1 and 3. This leads to difficulty in interpreting changes in the score at follow-up visits. Fortunately, to deal with the nonlinear property of summed scores for SST, the Rasch model converts ordinal measures into linear measures through logit transformation.[6] Generally speaking, the Rasch model is a transformation from SST summed scores into linear space, provided that the fit to the model and several assumptions about the model are appropriate.[7] Through the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ is given by:

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)}$$

where

X_{ni} is the response (1=Yes, 0= No) for Person n to Item i;

β_n is the ability of Person n (*n = 1, …, N),* which hereafter will be referred as the "Rasch score";

δ_i is the difficulty of Item i (*i = 1, …, I);*

P(•) is the probability that Person n has a true or observed response to Item i.

In other words, the person ability β_n $(n = 1, \ldots, N)$ indicates how able n^{th}person can perform items in SST*;* the item difficulty δ_i $(i = 1, \ldots, I)$ indicates the difficulty of i^{th}item. The person ability estimates are used for parametric analysis.

There is substantial evidence from the literature to support the use of the Rasch model to compare the outcomes either among patients, within patients (change scores), or between treatments. In 1996, Stucki et al. published an article with an informative title "Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts". [8] Stucki and colleagues illustrated difficulties in interpreting change scores of a health status measure (the physical ability scale of the SF-36) for clinical research or practice. Briefly, depending on baseline health status, seemingly equal gains of ordinal health status measures may actually imply different meanings. Such findings have several important implications for interpreting health status instruments for clinical sensibility. Consequently, as several studies have found, the relative precision (RP) (sometimes called the relative validity [RV]) is greater for Rasch scores than conventional summed scores. RP is defined as the ratio of pair-wise F statistics, which is the F statistic for the comparisons between groups based on the Rasch scores divided by the F statistic based on the summed scores.[9-12]

Additionally, White[13] revealed that the conclusions based on Rasch scores are not necessarily consistent with those based on Likert scaling. Rather, the two types of scores may lead to essentially differing conclusions, with disability categories based on Likert and Rasch score not equivalent.

Importantly, Tennant[7] recently compiled a primer on Rasch analysis that may serve as a basis for recommendations. In doing so, Tennant explained what “Rasch analysis” is, why it should be used, and when to apply it. According to this article, Rasch analysis should be applied whenever change scores need to be calculated from ordinal scales. The data must be shown to meet model expectations so that an interval (logit-based) estimate may be derived and logit person estimates may be exported for parametric analysis.

Nevertheless, previous studies on the Rasch Model have not dealt with measurement errors, which, as discussed, are virtually ubiquitous. The key purpose of this dissertation, then, is to address measurement errors in Rasch model. To do so, I proposed to incorporate them into the Rasch model under an established framework of measurement errors. The proposed model addressed linearity issues of the summed scores from questionnaires and more general issues related to measurement errors. Typically, the measurement errors in self-reported questionnaires may be modeled as classical measurement error. In the classical measurement error model for an item response data,[14] I introduce the following terms: an unobserved true response to Item i by Person n (X_{ni}) is measured by some individual-specific random error. The observed response W_{ni} may be different from X_{ni} because it is mainly caused by inaccurate information obtained in the self-report questionnaires. Accordingly, Rasch model accounting for the classical measurement error structure may be modeled in the following way: the logit transformed latent variables L_W and L_X linked the true response **X** and the observed response **W**.

$$\boldsymbol{L_W = L_X + U}$$

$$L_{X_{ni}} = \beta_n - \delta_i$$

where each component U_{ni} of **U** are i.i.d. random variables of measurement error.

Due to the difficulty in modeling the likelihood of the Rasch model along with measurement errors directly, I adopted the Markov Chain Monte Carlo (MCMC) approach to find the estimates of Rasch model parameters. From simulation exploration, I discovered optimal setting of the factors that affected the MCMC implementation.

Using the optimal setting of factors that affected the MCMC implementation, I applied the Rasch model accounting for measurement errors to SST data. From this model, I obtained the estimates of Rasch model parameter estimates ($\boldsymbol{\beta}$), denoted as MCMC Rasch SST scores. With MCMC Rasch SST scores, I determined the minimum clinically important difference (MCID) for patients with rotator cuff tendonitis or tearing. The SST, a questionnaire designed to assess patient reported shoulder function, consists of 12 yes/no questions.[4] For each question in the SST, answer "Yes" is coded as one (1) and answer "No" is coded as zero (0). The sum of answers (1 vs. 0) over all SST questions for each subject is used to characterize the patient's shoulder function, denoted as the SST summed score. The higher the SST summed score, the better is the patient-reported shoulder function. The range of SST summed score is from 0 to 12.[4]

The SST data were obtained from a larger dataset collected for a study whose Principal Investigator was Robert Z. Tashjian, MD, Department of Orthopedics, University of Utah, School of Medicine, Salt Lake City, UT.[15, 16] Permission to use these data was obtained from Dr. Tashjian (i.e., PI of the study for which data were collected).

In this dataset, 81 patients with rotator cuff tendonitis or tearing provided SST data both at Screening and Week 6 follow-up after the intervention with nonoperative modalities. All of them responded to two anchored questions at the week 6 follow-up. Through Rasch model accounting for measurement errors, I obtained the estimates of Rasch model parameters ($\boldsymbol{\beta}$) at both Weeks 0 and 6, called MCMC Rasch scores. If the difference of change from baseline in MCMC Rasch scores between two patient groups (i.e., No Change vs. Minimal Improvement) is statistically significant at 0.05 level, then this difference was defined as MCID.[17, 18] The two patient groups were classified by either15-item or four-item anchored questions (see Appendix C of Chapter 4 for the details of the two anchored questions). To evaluate the role of measurement errors on conclusions about MCID, I calculated the difference in change from baseline between the two patient groups delineated above (reflecting the two anchored questions) in three ways: SST summed scores, Rasch SST scores, and MCMC Rasch SST scores. Results were then compared. The anchored questions were (1) “Since your last clinic visit, has there been any change in the function of your treated shoulder?” (with 15 response categories ranging from “A very great deal worse” to “A very great deal better”); (2) “Since your last clinic visit, please rate your response to treatment.” (with 4 response categories ranging from “None” to “Excellent”).

Purpose of Study

In dealing with nonlinearity and measurement errors in self-reported responses to questionnaires, this dissertation proposes Rasch model accounting for measurement errors. Our particular purpose was to explore the construction of the Rasch model accounting for measurement errors in questionnaires with binary responses. Because of complexity of this model (i.e., the direct model of distribution for measurement errors is not available), there was no maximum likelihood solution. For this reason, I used Markov Chain Monte Carlo (MCMC) algorithm to obtain the numerical estimates for the model parameters. Next, via the Metropolis-Hastings algorithm with Gibbs sampler, I explored a spectrum of factors that may affect the MCMC implementation. In doing so, I found optimal setting for factors in the MCMC implementation. Using the optimal setting of factors that affect the MCMC implementation, I then applied Rasch model accounting for measurement errors to SST data in patients with rotator cuff tendonitis or tearing. The scores obtained through Rasch model accounting for measurement errors are referred to as MCMC Rasch scores. Finally, in order to evaluate the effect of measurement errors on the MCID determination, I compared the conclusions about the MCID obtained through SST summed scores, Rasch SST scores, and MCMC Rasch SST scores.

References

1. Deyo RA. Clinical research methods in low back pain. *Physical Medicine and Rehabilitation: State of the Art Reviews* 1991; **5**: 209-222.

2. Deyo RA. The quality of life, research, and care. *Ann Intern Med* 1991; **114**: 695-697.

3. Shumaker SA, Wyman JF, Uebersax JS, McClish D, Fantl JA. Health-related quality of life measures for women with urinary incontinence: the Incontinence Impact Questionnaire and the Urogenital Distress Inventory. *Quality of Life Research* 1994; **3**: 291-306.

4. University of Washington. Simple Shoulder Test, Available at http://www.orthop.washington.edu/uw/simpleshoulder/tabID__3376/ItemID__186/PageID__356/qview__true/Articles/Default.aspx, Accessed Jan 7, 2012

5. Carroll RJ, Ruppert D, Stefanski LA. Chapter 2 Important Concept. *Measurement Error in Nonlinear Models A Modern Perspective* (Second ed). Chapman & Hall/CRC, 1995.

6. Sheng X, Carrière KC. An improved CML estimation procedure for the Rasch model with item response data. *Statistics in Medicine* 2002; **21**: 407-416.

7. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research* 2007; **57**: 1358-1362.

8. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *J Clin Epidemiol* 1996; **49**: 711-717.

9. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-40): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997; **50**: 451-461.

10. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998; **51**: 1203-1214.

11. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs. raw scores in measuring change in health. *Medical care* 2004; **42**: I25-I36.

12. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, Gregg PJ. A comparison of Rasch with Likert scoring to discriminate between

patients' evaluations of total hip replacement surgery. *Quality of Life Research* 2004; **13**: 331-338.

13. White LJ, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. *Archives of Physical Medicine and Rehabilitation* 2002; **83**: 822-831.

14. Sheng X. A Bayesian model for measurement errors in diagnosis of rheumatoid arthritis. *Communications in Statistics - Theory and Methods* 2009; **38**: 3419 - 3431.

15. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010; **92**: 296-303.

16. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009; **18**: 927-932.

17. Copay AG, Subach BR, Glassman SD, Polly Jr DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal* 2007; **7**: 541-546.

18. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989; **10**: 407-415.

CHAPTER 2

MINIMUM CLINICALLY IMPORTANT DIFFERENCE FOR THE SIMPLE SHOULDER TEST: RASCH SCORE VS. SUMMED SCORE

Abstract

According to the literature, a common way to determine the minimum clinically important difference (MCID) is through an anchored question, and it is defined as the statistically significant difference in change from baseline (CFB) between a No Change group and a Minimal Improvement group. Using this method, summed scores are conventionally used to ascertain the MCID, even though this may be open to criticism. Given the importance of the MCID in clinical and outcomes research, I used another method to view the MCID, i.e., Rasch model transformation. Data, provided by patients treated conservatively for rotator cuff disease, consisted of their responses to the Simple Shoulder Test (SST). I used the Rasch Model to transform SST summed scores and two anchored questions to determine the MCID. According to the 15-item anchored question and transformed SST scores, the difference (95% CI) of the CFB between two groups was 1.97 (-0.17, 4.10). According to the four-item anchored question and transformed SST scores, the difference (95% CI) of the CFB between two groups was 2.38 (1.03, 3.74). Each result ascertained through Rasch transformation supported the other, and, furthermore, both of them are consistent with MCID’s ascertained by summed

scores. [1] This consistency in results, even though they were obtained from different analyses, may be reassuring to those who employ the MCID of the SST.

Introduction

The use of questionnaires in outcomes research immediately raises the question of how to establish the minimum clinically important difference (MCID). Given a particular outcome as measured by patients' responses on a questionnaire and a specific patient population, what is the MCID? A clinically important difference is distinguishable from a statistically significant difference.[2] A statistically significant difference does not necessarily imply a meaningful difference from the patient's point of view. In contrast, a MCID represents "the smallest difference in score in the domain which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management." [3]

From patients' point of view, results of laboratory and imaging tests as well as statistical improvement have little inherent value. What matters most to them is the degree to which they can function and perform daily activities after treatment.[4, 5] The MCID, based on patients' perception of function, by definition is pivotal to patient centered care. Conventionally, it is determined by the summed score derived from totaling responses from all individual questions.[6, 7] In view of the importance of the MCID in clinical care and outcomes research, we would not accept the MCID with a summed score at face value. Rather, aside from the MCID based on summed scores, I bring another method, i.e., Rasch model transformation, to bear on the MCID. Are results based on one of these methods consistent with results based on the other?

I explore this question using responses on the simple shoulder test (SST) provided by patients treated for rotator cuff disease.[8] The SST represents many advantages of using questionnaires in outcomes research.[4, 5] Because patients usually answered questions in questionnaires without assistance, the data collected through them systemically incorporate patients' perspectives in the evaluation of outcomes, and, relative to invasive procedures, they are less expensive and also patients may find them more acceptable. However, outcome measures based on patients' reports, such as SST, have a certain feature in common: their summed score is used as a measurement of disease severity.[6, 7] Among the problems that arise from this feature are the following: (1) there are no natural numerical upper or lower limits to health status, and for this reason, a zero value does not have an inherent meaning. (2) Furthermore, psychometric instruments such as the SST do not necessarily have interval characteristics (i.e., scaled in a linear fashion). In other words, one cannot assume that the two-point difference between scores of 10 and 12 is the same as the difference between scores of 1 and 3. This leads to difficulty in interpreting changes in the score at follow-up visits.

To deal with the nonlinear property of summed scores for SST, the Rasch model converts ordinal measures into linear measures.[9] Generally speaking, the Rasch model is a transformation from SST summed scores into linear space, provided that the fit to the model and several assumptions about the model are appropriate.[10] After the Rasch transformation of the SST summed score, the interval property of Rasch SST score is established. Then, I can analyze Rasch SST scores through t-tests to determine the MCID.

Materials and Methods

Simple Shoulder Test

The SST questionnaire consists of 12 “yes/no” questions (or coded as 1/0), which assess shoulder function.[8] The summed score of 12 SST questions is used to characterize shoulder function: the possible values of SST summed score are 0, 1, 2, …, to 12; the higher the SST score, the better is the shoulder function. The University of Washington, Department of Orthopedics and Sports Medicine, has compiled a website that provides a full description of the SST, including a listing of its questions, psychometric properties, and data obtained from patients.[8] According to the website, the SST may be administered to all patients presenting to a shoulder clinic.

Rasch Model

The basis of outcomes research frequently consists of patient reports obtained from questionnaires and, more specifically, change scores derived from patients’ reports. Rasch analysis is useful because it transforms ordinal scores into linear, interval-level scores, given the fit of data to Rasch model expectations. [10]

Considering the data structure of two category items, in the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ is given by:

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)} \tag{2.1}$$

where

X_{ni} is response (1=Yes, 0= No) from Person n to Item i;

β_n is the ability of person n (*n = 1, …, N),* which hereafter will be referred as the "Rasch score";

δ_i is the difficulty of item i (*i = 1, …, I).*

Through the logit transformation, the dichotomous Rasch model has the following linear form:

$$L_{X_{ni}} = logit[P(X_{ni} = 1)] = \beta_n - \delta_i$$

The person ability β_n $(n = 1, \dots, N)$ is to measure how able n^{th}person can perform items in SST*;* the item difficulty δ_i $(i = 1, \dots, I)$ is to measure the difficulty of the i^{th}item. L_X is the logit (log odds) of performance by a person to an item, based on the model, equal to $\beta_n - \delta_i$.

According to Rasch, "A person having a greater ability than another should have the greater probability of solving any item of the type in question and similarly one item being more difficult than another one means that for any person, the probability of solving the second item correctly is the greater one." [11]

In other words, the greater the distance between the person and item location (the difference between β *and* δ), the greater certainty we would expect in the person's ability to perform successfully on an item. However, as this distance approaches zero, then the more likely we are to say that there is a 50-50 chance that the person will successfully perform the item.[12]

Data Source

The data used in this analysis were obtained from a larger dataset collected for a study whose Principal Investigator was Robert Z. Tashjian, MD, Department of Orthopedics, University of Utah, School of Medicine, Salt Lake City, UT. [1, 13] Prior to data collection, the Tashjian study received approval from the University of Utah Institutional Review Board.

The following briefly describes inclusion and exclusion criteria for the study population. Patients were included if they:

1. were over the age of 18;
2. had clinical diagnosis of rotator cuff tear or rotator cuff tendonitis;
3. agreed to participate in nonoperative treatment.

Patients were excluded if they:

1. were treated with early surgical repair (for acute full-thickness tears or chronic tears in patients under the age of 60), glenohumeral arthritis, adhesive capsulitis;
2. were not willing to participate in nonoperative treatment;
3. had any signs of glenohumeral arthritis present on radiographs, including humeral or glenoid osteophytes, joint space narrowing, subchodral sclerosis, or subchondral cysts;
4. had a diagnosis of rotator cuff disease and stiffness with a global loss of motion in all planes compared to the opposite shoulder
5. had less than 50% external rotation at the side compared to the opposite shoulder.

All patients meeting the inclusion and exclusion criteria were asked to participate in the study. All participating patients went through initial screening activities, which included a history, physical examination, and shoulder radiographs. Additionally, they all had magnetic resonance imaging of the shoulder performed either prior to the initial evaluation (if one was obtained) or during the screening. Their final diagnosis was based on the initial physical examination, and the initial imaging studies (radiographs and magnetic resonance imaging scans).

Eighty-one patients with rotator cuff tendonitis or tearing were enrolled and treated with nonoperative modalities. Patient characteristics are described in the previous publication.[1] All of them provided SST data both at Screening and Week 6 follow-up.

Statistical Analysis

First, through the Rasch model, the summed scores from the SST questionnaire data were converted to Rasch scores for each subject. Then, the minimum clinically important difference (MCID) of Rasch scores was determined using anchor-based approaches [2] as described below. Anchor-based approaches rely on the relationship of the outcome instrument being evaluated and an independent measure of improvement question (i.e., anchor question).[2, 3, 14] Table 2.1 provides the two anchored questions.

Two anchored questions were chosen to determine MCID in SST Rasch scores because these anchored questions were used in the Robert Z. Tashjian's articles. According to the 15-item function question in Table 2.1,

- the patients will be classified as no change group [1] if their answers to this question are one of the following:
  - almost the same, hardly any worse at all;
  - no change;
  - almost the same, hardly any better at all.

Table 2.1: Anchored Questions

15-Item Function Question
Since your last clinic visit, has there been any change in the function of your treated shoulder?
A very great deal worse
A great deal worse
A good deal worse
Moderately worse
Somewhat worse
A little worse
Almost the same, hardly any worse at all
No change
Almost the same, hardly any better at all
A little better
Somewhat better
Moderately better
A good deal better
A great deal better
A very great deal better
four-Item Question
Since your last clinic visit, please rate your response to treatment.
None – no good at all, ineffective treatment
Poor – some effect but unsatisfactory
Good – satisfactory effect with occasional episodes of pain or stiffness
Excellent – ideal response, virtually pain free

- the patients will be classified as minimal improvement group [1] if their answers to this question are one of the following:
  - a little better;
  - somewhat better.

According to the four-item question in Table 2.1,

- the patients will be classified as no change group [1] if their answers to this question are one of the following:
  - none;
  - poor.

- the patients will be classified as minimal improvement group [1] if their answers to this question are one of the following:
    - good.

With the classification of patient groups (No change vs. Minimal Improvement), the mean change from baseline with respect to SST were calculated for each of patient groups. The statistically significant difference of changes from baseline between patient groups was defined as the MCID. The p value and 95% confidence interval (CI) from a t-test was provided on the difference of change from baseline between “No Change” and “Minimal Improvement” in the section of Results.

Results

Table 2.2 presents the SST Rasch score and its change from baseline (CFB) by anchored questions. The estimated difference (95% CI) of the CFB between two groups is 1.10 (-0.05, 2.25) for the 15-item question while the estimated difference (95% CI) of CFB between the two groups is 1.33 (0.59, 2.06) for the four-item question.

I found that SST Rasch score is between -3 and 3 with standard deviation of 1.5 while the SST summed score is between 0 and 12 with standard deviation of 3.2. In order to obtain a fair comparison with the MCID derived from the SST summed score,[1] the SST Rasch scores were rescaled by the factor of the ratio of standard deviations of SST summed score over SST Rasch scores. In other words, I rescaled the SST Rasch score so that it had a similar scale of standard deviation as the SST summed score. The results are shown in Table 2.3.

Table 2.2: SST Rasch Score and its Change from Baseline

Anchored		No Change		Minimal Improvement	
Questions	Visit	Mean (SD)	n	Mean (SD)	n
15-Item Question	Week 0 (BL)	0.16 (1.526)	9	0.44 (1.829)	19
	Week 6	-0.02 (1.231)	9	0.81 (1.542)	17
	CFB	-0.17 (1.462)	9	0.92 (1.262)	16
	Difference (CI)	1.10 (-0.05, 2.25)			
	P Value	0.0606			
four-Item Question	Week 0 (BL)	0.25 (1.588)	20	-0.11 (1.522)	46
	Week 6	-0.29 (1.517)	21	1.13 (1.309)	37
	CFB	0.03 (1.401)	18	1.36 (1.218)	37
	Difference (CI)	1.33 (0.59, 2.06)			
	P Value	0.0007			

BL= Baseline; CFB= Change from Baseline; CI=95% Confidence Interval

For purposes of comparison (with results in Table 2.3), the results on SST summed score from a previous study results [1] are shown in Table 2.4.

From the results shown in Table 2.3 and 2.4, the difference (95% CI) of the CFB between two groups is 1.97 (-0.17, 4.10) for the rescaled SST Rasch score vs. 1.95 (0.06, 3.85) for the SST summed score, using the 15-item anchored question. Based on the four-item anchored question, the difference (95% CI) of the CFB between the two groups is 2.38 (1.03, 3.74) in the rescaled Rasch SST score vs. 2.33 (0.99, 3.66) in the SST summed score. Note in the results, the CIs in Rasch scores were wider than CIs in summed scores because Rasch scores are nonestimable for the summed scores of zero (the answers are all NO) and twelve (the answers are all YES). For this reason, five patients were lost using 15-item anchored question and 15 patients using four-item anchored question.

Table 2.3: Rescaled SST Rasch Score and Its Change from Baseline

Anchored		No Change		Minimal Improvement	
Questions	Visit	Mean (SD)	n	Mean (SD)	n
15-Item Question	Week 0 (BL)	0.30 (2.856)	9	0.83 (3.422)	19
	Week 6	-0.03 (2.194)	9	1.45 (2.750)	17
	CFB	-0.32 (2.698)	9	1.64 (2.348)	16
	Difference (CI)	1.97 (-0.17, 4.10)			
	P Value	0.0693			
four-Item Question	Week 0 (BL)	0.46 (2.972)	20	-0.21 (2.848)	46
	Week 6	-0.53 (2.704)	21	2.02 (2.333)	37
	CFB	0.06 (2.568)	18	2.44 (2.247)	37
	Difference (CI)	2.38 (1.03, 3.74)			
	P Value	0.0009			

BL= Baseline; CFB= Change from Baseline; CI=95% Confidence Interval

Table 2.4: SST Summed Score and Its Change from Baseline

Anchored		No Change		Minimal Improvement	
Questions	Visit	Mean (SD)	n	Mean (SD)	n
15-Item Question	Week 0 (BL)	6.33 (3.000)	9	6.71 (3.690)	21
	Week 6	6.00 (2.500)	9	8.33 (3.055)	21
	CFB	-0.33 (2.958)	9	1.62 (2.012)	21
	Difference (CI)	1.95 (0.06, 3.85)			
	P Value	0.0439			
four-Item Question	Week 0 (BL)	5.83 (3.667)	24	5.76 (2.877)	46
	Week 6	6.33 (3.447)	24	8.59 (2.833)	46
	CFB	0.50 (2.396)	24	2.83 (2.783)	46
	Difference (CI)	2.33 (0.99, 3.66)			
	P Value	0.0009			

BL= Baseline; CFB= Change from Baseline; CI=95% Confidence Interval

Discussion

In comparing the results of these four analyses, the point estimates are of similar magnitude (1.97 of rescaled Rasch score vs. 1.95 of the SST summed score using the 15-item anchored question; 2.38 of rescaled Rasch score vs. 2.33 of SST summed score using the four-item anchored question). All four analyses reached statistical significance or borderline significance at α=0.05. The two sets of results, one obtained from the analysis of rescaled Rasch scores and the other obtained from the analysis of summed scores, show agreement with each other both clinically (in terms of magnitude) and statistically. Above all, it appears that the results from the Rasch model supports the conclusion in the original study, which determined that a 2 point change in the SST summed score was the MCID.[1]

Because the Rasch model can be viewed as a transformation from SST summed scores into linear space, provided that the fit to the model and several assumptions[10, 15] about the model are appropriate, Figure 2.1 shows the relationship between Rasch SST scores (i.e., person ability) and SST summed scores in our population. Although scores from the two methods are expected to be highly related, they do not form a linear relationship due in part to the logarithmic nature of the Rasch model. A plot of the two scores is expected to be represented by an ogive (sigmoidal-shaped curve) in which the curve rises gradually, has a steep central slope, and then gradually flattens out [16, 17]. The shape of the curve indicates that the Rasch scores at the upper and lower end of the range are more spread out than the corresponding the SST summed scores. In other words, the graph demonstrates that the Rasch model describes more meaningful changes for those patients at the upper and lower end of the range, which would not be indicated when using SST summed scores (i.e., equal spaces between any two scores).

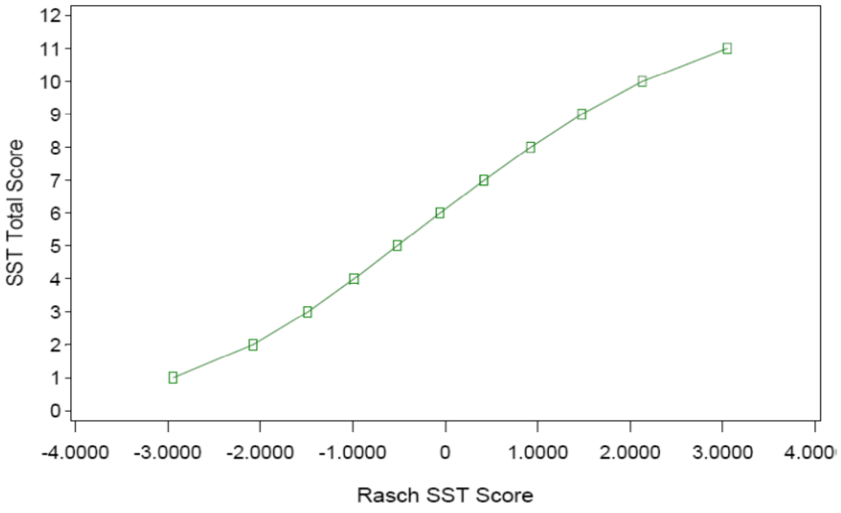

Figure 2.1: Relationship between SST Total Score and Rasch SST Score

After transformations (our advice would be to let the statistician build the Rasch model for each unique patient dataset), clinicians may conveniently use Rasch SST scores to draw conclusions. The change scores can be identified for a patient, and, in turn, they can be compared with MCID determined in terms of Rasch scores. The interpretation in Rasch scores across patients is much simpler than SST summed scores because the change scores of the same magnitude have the same meaning across all the baselines. That is, a change of two in Rasch scores for one patient with a baseline score of minus one has the same meaning as a change of two in Rasch scores for another patient with a baseline score of three because the Rasch score is the part of linear model in logit form. In contrast, the use of summed scores provides no such guarantee. In other words, Rasch scores provide the basis for clinicians to compare results across patients in order to evaluate if an intervention is effective.

In explaining to patients how they are doing after an intervention, however, clinicians may still use SST summed scores as a reference point. In that case, a clinician may say to a patient: “When you came in initially, you only indicated you could perform 6 activities. Now, after the intervention, you can perform 8 activities”

A limitation of the study is its relatively small sample size. Even though a total of 81 subjects were enrolled in the study, the sample size for the “No Change” group is 9 while the sample size for the “Minimally Improvement” Group is 21 according to 15-item question. The sample size is 24 for the “No Change” group and 46 for the “Minimally Improved” group according to four-item question. From the above two analyses, the MCID for SST summed score was determined to be a change of 2 points. In the analysis with the four-item question, the power to detect a difference of 2 points between two groups is about 80% at the current sample sizes (24 for No Change vs. 46 for Minimal Improvement).

To conclude, clinicians and outcomes researchers have relied upon the MCID, and this has profound implications for patient care. I raise the question whether it should be trusted. For this reason, I viewed the MCID from another statistical perspective. Despite small sample sizes, the consistency in results is notable: each result ascertained through Rasch transformation supported the other, and both of them were consistent with MCID’s ascertained by the SST summed score [1]. These results may be reassuring to physicians and researchers who employ the MCID of the SST because different methods of determining it (summed scores and Rasch scores) are mutually supportive.

References

1. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010; **92**: 296-303.

2. Copay AG, Subach BR, Glassman SD, Polly Jr DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal* 2007; **7**: 541-546.

3. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989; **10**: 407-415.

4. Deyo RA. Clinical research methods in low back pain. *Physical Medicine and Rehabilitation: State of the Art Reviews* 1991; **5**: 209-222.

5. Deyo RA. The quality of life, research, and care. *Ann Intern Med* 1991; **114**: 695-697.

6. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473-483.

7. Radloff L. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977; **46**: 385-401.

8. University of Washington. Simple Shoulder Test, Available at http://www.orthop.washington.edu/uw/simpleshoulder/tabID__3376/ItemID__186/PageID__356/qview__true/Articles/Default.aspx, Accessed Jan 7, 2012.

9. Sheng X, Carrière KC. An improved CML estimation procedure for the Rasch model with item response data. *Statistics in Medicine* 2002; **21**: 407-416.

10. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research* 2007; **57**: 1358-1362.

11. Rasch G. Chapter VII Notions Implied in the Structural Model for Items. *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press: Chicago, 1980.

12. de Ayala RJ. Chapter 2 The One-Parameter Model. *The Theory and Practice of Item Response Theory.* The Guildford Press: New York ondon, 2009.

13. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009; **18**: 927-932.

14. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, Haythornthwaite JA, Jensen MP, Kerns RD, Ader DN, Brandenburg N, Burke LB, Cella D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kehlet H, Kramer LD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rauschkolb C, Revicki DA, Rothman M, Schmader KE, Stacey BR, Stauffer JW, von Stein T, White RE, Witter J, Zavisic S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008; **9**: 105-121.

15. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health* 2004; **7**: S22-S26.

16. Baker FB. Chapter 4 The Test Characteristic Curve. *The Basics of Item Response Theory.* Heinemann: Portsmouth, New Hampshire, 1985.

17. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-40): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997; **50**: 451-461.

CHAPTER **3**

BAYESIAN APPROACH FOR RASCH MODEL ACCOUNTING FOR MEASUREMENT ERRORS

Abstract

Measurement errors exist whenever there is an attempt to measure. They may distort an analysis in many ways, which in turn may result in attenuation to the null, loss of power, and bias of estimates. I used a Bayesian approach through Markov Chain Monte Carlo (MCMC) methods in order to account for measurement errors in the Rasch model. In doing so, I explored a spectrum of factors that may affect the MCMC implementation. This model I used constitutes our attempt to address more general issues related to measurement errors. Importantly, its use led to the exploration of how to obtain the estimates of β*s* corresponding to extreme summed scores with reasonable variances, which remain inestimable in the classical Rasch model. Rasch model accounting for measurement errors enables us to prevent missing data in the conversion process between the summed scores and linear Rasch scores.

Introduction

Measurement errors exist whenever there is an attempt to measure.[1] Most often they consist of measurement errors in questionnaires, which may be defined as the discrepancy between respondents' true attributes and the data obtained in the questionnaire about their attributes. A subtype of measurement errors is response error, which has many sources, including the mode of interview, wording of questions, interviewer behavior, sensitivity of information requested, respondents' recall, and coding errors.[2]

Additionally, when subjects are asked to complete questionnaires, we necessarily narrow the choices with which they may respond. For example, in binary response questions, there are only two categories for them to choose (e.g., 1=yes or 2=no), which is exemplified in the simple shoulder test (SST). In other words, questionnaires, particularly those consisting of binary response categories, force subjects to make a choice among predefined response categories, even though their true choice may fall in-between these categories.

This suggests that measurement error is much more than simple recording or instrument error. It encompasses many different sources of variability.[3] Measurement errors may distort an analysis in the following ways:[4, 5]

- attenuation to null
- loss of power
- bias of estimates (i.e., real effects are hidden, observed data exhibit relationship that are not present in the error-free data, and the signs of estimated coefficients are reversed relative to the case with no measurement error)

There is substantial evidence from the literature to support the use of the Rasch model to compare outcomes measured by questionnaires, either among patients, within patients (change scores), or between treatments. According to the following studies, in relation to summed scores,

- the Rasch model has produced more accurate estimates of change in terms of estimated trait level.[6]
- Rasch scores also achieved greater relative precision when compared with conventional summed scores in many articles.[7-10]

Nevertheless, preexisting studies on Rasch Models have not dealt with virtually ubiquitous measurement errors. This chapter extends fundamental considerations of measurement errors in Rasch modeling and addresses the following issues in dichotomous response data in questionnaires: nonlinear summed scores and measurement errors. To do so, I implemented the Markov Chain Monte Carlo (MCMC) method, specifically via the Metropolis-Hastings algorithm.[11][12]

Background

Simple Shoulder Test

The SST is a self-reported questionnaire, designed to assess patient reported shoulder function, consists of 12 yes/no questions.[13] For each question in the SST, answer “Yes” is coded as one (1) and answer “No” is coded as zero (0). The sum of answers (1 vs. 0) over all SST questions for each subject is used to characterize the patient’s shoulder function, denoted as the SST summed score. The higher the SST summed score, the better the patient-reported shoulder function is. The range of SST summed score is from 0 to 12.[13]

From the SST, the summed score is used as a measurement of disease severity. Among the problems that arise from this feature are the following:

- There are no natural numerical upper or lower limits to health status. Thus, a zero value does not have an inherent meaning.
- The SST summed score is not necessarily an interval measure (i.e., linear measure). The spaces between 0, 1… and 12 may not be equal, which variously impacts change scores depending on the initial scores, i.e., 12 – 10 ≠ 3 – 1.

Rasch Model Accounting for Measurement Errors

To deal with the nonlinear property of summed scores for the SST, the Rasch model converts ordinal measures into linear measures. Considering the data structure of two category items through the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ is given by:

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)} \qquad (3.1)$$

where

X_{ni} is the response (1=Yes, 0= No) for Person n to Item i;

β_n is the ability of Person n (*n = 1, …, N),* which hereafter will be referred as the "Rasch score";

δ_i is the difficulty of Item i (*i = 1, …, I);*

P(·) is the probability that Person n has a true or observed response to Item i.

Through the logistic regression model, the dichotomous Rasch model has the following linear form:

$$logit[P(X_{ni} = 1)] = \beta_n - \delta_i \qquad (3.2)$$

The person ability β_n $(n = 1, \ldots, N)$ is to measure how able n^{th}person can perform items in SST; the item difficulty δ_i $(i = 1, \ldots, I)$ is to measure how difficult of i^{th}item.

In order to account for the effects of measurement errors, I proposed to incorporate them into the Rasch model under a preestablished framework of measurement errors.[3] Recall that the SST consists of 12 self-reported yes/no questions. Typically, the measurement errors in self-reported questionnaires may be modeled as classical measurement error. In the classical measurement error model for an item response data,[14] I introduce the following terms: an unobserved true response to Item i by Person n (X_{ni}) is measured by some individual-specific random error. The observed response W_{ni} may be different from X_{ni} because it is mainly caused by inaccurate information obtained in the self-report questionnaires.

Because the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{W}$ is difficult to model directly, I modeled two latent variables $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ instead. The $\boldsymbol{L_W}$ was defined as $logit\left(P(\boldsymbol{W})\right)$ and the $\boldsymbol{L_X}$was defined as $logit\left(P(\boldsymbol{X})\right)$**.** The Rasch model incorporating the classical measurement error structure may be modeled in the following way: the logit transformed latent variables $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ linked the true response $\boldsymbol{X}$ and the observed response $\boldsymbol{W}$**.**

$$\boldsymbol{L_W} = \boldsymbol{L_X} + \boldsymbol{U} \qquad (3.3)$$

$$L_{X_{ni}} = \beta_n - \delta_i \qquad (3.4)$$

where each component U_{ni} of $\mathbf{U}$ are i.i.d. random variables of measurement error

$\boldsymbol{U} \sim N(0\mathbf{1}, \sigma_u^2\boldsymbol{I})$, where σ_u^2 is the variance of the classical measurement error.

$\boldsymbol{L_X} \sim N(\boldsymbol{\mu_{L_X}}, \sigma_{L_X}^2\boldsymbol{I})$, where $\boldsymbol{\mu_{L_X}}$ is the mean of $\boldsymbol{L_X}$, and $\sigma_{\boldsymbol{L_X}}^2$ is the variance of $\boldsymbol{L_X}$

Therefore, the joint distribution of $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ is:

$$\begin{pmatrix} \boldsymbol{L_X} \\ \boldsymbol{L_W} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu_{L_X}} \\ \boldsymbol{\mu_{L_X}} \end{pmatrix}, \begin{bmatrix} \sigma_{L_X}^2\boldsymbol{I} & \sigma_{L_X}^2\boldsymbol{I} \\ \sigma_{L_X}^2\boldsymbol{I} & \left(\sigma_{L_X}^2+\sigma_u^2\right)\boldsymbol{I} \end{bmatrix}\right) \tag{3.5}$$

and the conditional distribution of $\boldsymbol{L_X}$ given $\boldsymbol{L_W}$ is:

$$\boldsymbol{L_X}|\boldsymbol{L_W} \sim \boldsymbol{N}\ \left(\boldsymbol{\mu_{L_X}} + \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2+\sigma_u^2}\boldsymbol{I}\left(\boldsymbol{L_W} - \boldsymbol{\mu_{L_X}}\right),\ \ \sigma_{L_X}^2\left(1 - \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2+\sigma_u^2}\right)\boldsymbol{I}\right) \tag{3.6}$$

In order for the implementation of the above formula, I further assume that $\boldsymbol{\mu_{L_X}}$is equal to $\boldsymbol{L_W}$, and $\sigma_{\boldsymbol{L_x}}^2$ is equal to $\sigma_{\boldsymbol{L_W}}^2$ estimated as $var\ (\boldsymbol{L_W})$, and $\sigma_{\boldsymbol{U}}^2$ is equal to the proportion (denoted as a%) of $\sigma_{\boldsymbol{L_X}}^2$. Therefore, Formula (3.6) became

$$\boldsymbol{L_X}|\boldsymbol{L_w} \sim N\left(\boldsymbol{L_W}\mathbf{1}, \sigma_{\boldsymbol{L_W}}^2\left(\frac{0.a}{1.a}\right)\boldsymbol{I}\right) \tag{3.7}$$

Description of Source Data and Simulated Data

The SST data were obtained from a larger dataset collected for a study whose Principal Investigator was Robert Z. Tashjian, MD, Department of Orthopedics, University of Utah, School of Medicine, Salt Lake City, UT.[15, 16] Prior to data collection, Dr. Tashjian's study received approval from the University of Utah Institutional Review Board.

Eighty-one patients with rotator cuff tendonitis or tearing were enrolled and treated with nonoperative modalities. All of them provided SST data both at Screening and Week 6 follow-up. Permission to use these data was obtained from Dr. Tashjian (i.e., PI of the study for which data were collected).

Due to the difficulty in modeling the likelihood of the Rasch model along with measurement errors directly, I used simulations to demonstrate the utilization of our methods. I first estimated βs and δs from classical Rasch model based on observed SST data at week 0 without accounting for measurement errors. I then used these estimates as true parameters. Starting with true βs and $\delta s,$ I introduced a proportion of variance of L_X as the measurement errors into Rasch model and then generated simulated datasets. One hundred datasets were generated with each measurement error scenario: 10%, 50%, and 100% of the variance of $\boldsymbol{L_X}$. I will refer to three scenarios of measurement errors as, respectively, 10% level, 50% level, and 100% level of measurement errors.

For each above simulated dataset with known measurement errors, I used a Bayesian approach through Markov Chain Monte Carlo (MCMC) method[11, 17, 18] to estimate the parameters βs and δs from the Rasch model incorporating measurement errors. In doing so, I adopted Gibbs sampler in conjunction with a Metropolis-Hasting steps to evaluate the effect of measurement errors in Rasch model. After the MCMC simulations, I summarized posterior distribution of Rasch model accounting for measurement errors and compared the results with true parameters.

Methods

Overview of Definition and Properties of Markov Chains

Consider a distribution π from which a sample must be drawn via Markov chains. $P(y|x)$ is transitional probability (or conditional transitional density) and

$$P(y|x) = Pr(\theta(n+1) = y\ |\theta(n) = x) \qquad \textit{(3.8)}$$

The conditional transitional density

$$p(y|x) = \frac{\partial P(y|x)}{\partial y}, for\ x, y\ \in S \qquad \textit{(3.9)}$$

where S is the state space.

According to Section 6.2 of Gamerman,[19] a transition kernel $p(y|x)$ must be constructed in a way such that π is the equilibrium distribution of the chain. A simple way to do this is to consider reversible chains where the kernel p satisfies[19]

$$\pi(x)p(x|y) = \ \pi(y)p(y|x), for\ all\ x\ and\ y \qquad \textit{(3.10)}$$

As seen in Section 4.6 in the text on Markov chain Monte Carlo,[20] Equation (3.10) is the reversibility condition of the chain, which is also referred to as the detailed balance equation. Even though this is not a necessary condition for convergence, it is a sufficient condition in order that π be the equilibrium distribution of the chain.

The kernel $p(y|x)$ consists of 2 elements:

- an arbitrary transition kernel $q(y|x)$
- a probability of $\alpha(x, y)$

$$p(y|x) = q(y|x)\alpha(x, y), if\ x \neq y \tag{3.11}$$

So the transitional kernel defines a density $p(\cdot\ |x)$ for every possible value of the parameter different from x. Therefore, there is a positive probability left for the chain to remain at θ: $p(x|x) = 1 - \int q(y|x)\ \alpha(x, y)dy$

The general form for transitional kernel is:

$$p(A|x) = \int_A q(y|x)\alpha(x, y)\ dy + I(x \in A)\left[1 - \int_A q(y|x)\alpha(x, y)\ dy\right] \tag{3.12}$$

Hastings [21] proposed to define the acceptance probability in such a way that when combined with arbitrary transition kernel, it defines a reversible chain as follows:

$$\alpha = min\left\{1, \frac{\pi(y)q(y|x)}{\pi(x)q(x|y)}\right\} \tag{3.13}$$

Those algorithms, based on chains with transitional kernel (3.12) and acceptance probability (3.13), are Metropolis-Hastings algorithms. The transitional kernel q defines only a possible move that can be confirmed according to the value of acceptance probability α. Thus, q is generally referred to as the proposal kernel or proposal (conditional) density.

It is crucial that the proposal kernels are easy to draw from, because the Metropolis-Hastings method replaces the difficult generation of π by many generations proposed from q. Another equally important requirement to be met by q is the correct tuning of the moves it proposes to ensure that moves cover the parameter space and may be accepted in the computing power. Except for a few technical restrictions in the previous paragraph, there is a total flexibility for the choice of the proposal transitional kernel q.

Bayesian Methods through MCMC Process

As pointed out above, it is not practical to model the joint distribution of true and observed response (i.e., **X** and **W**) directly. For this reason, I employed Logit transformation to convert **X** and **W** into $\boldsymbol{L_X}\ and\ \boldsymbol{L_w}$. This allows us to assume that measurement errors are normally distributed in the Logit dimension. The Bayesian approach is a very efficient and effective way to obtain the model parameter estimates. According to the Bayesian method, the complete likelihood function needs to be constructed for the Rasch model accounting for measurement errors. The complete likelihood function for the model (see Appendix A for details) is

$$LH(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{L_X}) = \{f(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\delta})\} \times \{f(\boldsymbol{\beta}) * f(\boldsymbol{\delta})\} \times \{f(\boldsymbol{X}|\boldsymbol{L_X})\} \times \{f(\boldsymbol{L_X}|\boldsymbol{L_W})\} \qquad (3.14)$$

I generated a Gibbs Sampler for each parameter according to the following outlined steps. After a Gibbs sampler is generated at each step as a candidate θ, Metropolis-Hastings Algorithm to accept this candidate θ from a proposed distribution q(.) is based on the following probability:

$$\alpha = min\left\{1, \frac{LH\ (\theta_{candidate})q(\theta_{current}|\theta_{candidate})}{LH\ (\theta_{currrent})q(\theta_{candidate}|\theta_{currrent})}\right\} \qquad (3.15)$$

Then compare α with probability generated from uniform distribution. If α is greater than the probability from uniform distribution, then update $\theta_{current}$ with $\theta_{candidate}$.

Otherwise, keep $\theta_{currrent}$ as it is. From $LH(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{L_X})$, the steps to generate Gibbs samplers are:

1. Obtain initial starting values of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, e.g. estimates the two sets of parameters from the simulated dataset using Rasch model ignoring measurement error. Set these values as the current values $\boldsymbol{\beta}_{current}$ and $\boldsymbol{\delta}_{current}$.
2. Generate a candidate of the true Logit $\boldsymbol{L_X}$ according to the distribution of $\boldsymbol{L_X}|\boldsymbol{L_W}$, given the observed $\boldsymbol{L_W}$ equal to $\boldsymbol{\beta}_{current}$ minus $\boldsymbol{\delta}_{current}$.
3. Generate a candidate of true response $\boldsymbol{X}_{candidate}$ based on updated $\boldsymbol{L_X}_{current}$. $\boldsymbol{X}_{candidate}$ follows logistic distribution.
4. Generate a candidate of $\boldsymbol{\beta}$ based on current $\boldsymbol{\beta}_{\text{current}}$ using the normal distribution $N(\boldsymbol{\beta}_{curr}, \sigma^2_{\boldsymbol{\beta}\,proposed})$.
5. Generate a candidate of $\boldsymbol{\delta}$ based on current $\boldsymbol{\delta}_{current}$ using the normal distribution $N(\boldsymbol{\delta}_{curr}, \sigma^2_{\boldsymbol{\delta}\,proposed})$.

Repeat Step 2 to 5 for a large number of times, or until convergence. For each generated Gibbs sampler, the associated acceptance ratio can be found in Appendix B.

Markov Chain Monte Carlo Implementation

As we know, many factors will affect the implementation of a MCMC algorithm, such as convergence and acceptance rate. The focus of this chapter is to obtain the true posterior distribution of $\boldsymbol{\beta}$. Other model parameters $\boldsymbol{L_X}, \boldsymbol{X}, and\ \boldsymbol{\delta}$ are considered as the nuisance parameters.

General Exploration of Factors Affecting MCMC implementation

In our case, the following factors may affect the implementation of MCMC algorithm: the variance of measurement errors ($\sigma_{\boldsymbol{u}}^2$) equal to the proportion of $\sigma_{L_w}^2$ (a%) and prior variance ($\sigma_{\boldsymbol{\beta}}^2$ and $\sigma_{\boldsymbol{\delta}}^2$) and proposal variances for $\boldsymbol{\beta}$ *and* $\boldsymbol{\delta}$.
After carefully examining the chains of $\mathbf{L_X}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$, I found that the variance of measurement errors, expressed as a% × $\sigma_{\mathbf{L_w}}^2$, only dictated the generation of $\mathbf{L_X}$ chains but not the acceptance ratio of $\mathbf{L_X}$ because the acceptance ratio for $\mathbf{L_X}$ did not involve any $\mathbf{L_X}$ portion; the chains of $\mathbf{X}$ were derived from $\mathbf{L_X}$ and the acceptance ratio for $\mathbf{X}$ depends on current $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. For both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, prior variance and proposal variance are deciding factors for a candidate to be accepted. Therefore, I focused on the behaviors of chains for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Several combinations of the factors ($\sigma_{\boldsymbol{\beta}}^2, \sigma_{\boldsymbol{\beta}\,\text{proposed}}^2, \sigma_{\boldsymbol{\delta}}^2, \sigma_{\boldsymbol{\delta}\,\text{proposed}}^2$) were used to check the acceptance rate and convergence from trace plots of $\boldsymbol{\beta}$ *and* $\boldsymbol{\delta}$. The explored ranges for these parameters were $\sigma_{\boldsymbol{\beta}}^2$: 1 to 4; $\sigma_{\boldsymbol{\beta}\,\text{proposed}}^2$: 0.01 to 0.5; $\sigma_{\boldsymbol{\delta}}^2$:1 to 4; $\sigma_{\boldsymbol{\delta}\,\text{proposed}}^2$: 0.01 to 0.5. Generally speaking, no matter what these four factors are, they did not affect the average of acceptance rates for $\mathbf{L_X}$ (98%) and $\mathbf{X}$ (20%) and the pattern of trace plots of these two sets of model parameters. However, when prior variances for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ increased from 1 to 4, the average acceptance rates were slightly rising to about 95-96% and 89-91% respectively. For $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, if I increased the proposed variances from 0.01 to 0.5, the average acceptance rates decreased to about 78% for $\boldsymbol{\beta}$ and around 60% for $\boldsymbol{\delta}$. In short, since the prior variances for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ did not substantially affect acceptance rates, I chose sample variances of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ as the prior variances for further explorations. Because the sample variance for $\boldsymbol{\delta}$ is relatively small, so I chose the proposed $\sigma_{\boldsymbol{\delta},\text{proposed}}^2$ equal to 0.1. However, according to the following

exploration results in Table 3.1, I chose the proposed equal to 0.5 such that the average acceptance rate is below 80%.

Exploration of Factor $\boldsymbol{\sigma^2}$ for $\boldsymbol{\beta}$ in the Implementation of MCMC

I arbitrarily chose two simulated data sets (labeled as 12th and 73rd dataset) from each level of measurement errors (see Section: Description of Source Data and Simulated Data for generation of simulated datasets). In this exploration, I used the mean squared error (MSE) of an estimator to quantify the difference between values implied by an estimator and the true values of the quantity being estimated. MSE is defined as variance plus square of bias. I ran 100,000 MCMC iterations for each selected dataset. The first 50,000 iterations were discarded and the last 50,000 iterations were used for the calculations of MSE. From the distribution of 81 true βs (corresponding to 81 subjects) in Week 0 data, I selected five true β parameters according to quartiles, minimum, and maximum values. In Table 3.2 to Table 3.7, I present the results for the selected five true βs from Rasch model and the MCMC results. To facilitate the description, I also classified βs into two categories: extreme β vs. nonextreme β. Extreme β refers to the β that came from subject's summed score of 0 or 12 while nonextreme β refers to the β that came from subject's summed score of 1 to 11. Other terms in the tables are described as below:

- Classical Rasch Model is a Rasch model (Eq. 3.1) without accounting for measurement errors

Table 3.1: Acceptance Rate for βs on MCMC Results from 100,000 Iterations for a Simulated Dataset (12) with 10% Level of Measurement Errors

$\sigma^2_{\beta,\ proposed}$	Statistic	β
0.01	Mean	0.9957
	25th, 75th	0.9951 , 0.9957
	Median	0.9955
	Min, Max	0.9942 , 0.9987
0.05	Mean	0.9784
	25th, 75th	0.9759 , 0.9772
	Median	0.9764
	Min, Max	0.9749 , 0.9950
0.1	Mean	0.9572
	25th, 75th	0.9526, 0.9542
	Median	0.9534
	Min, Max	0.9506 , 0.9902
0.25	Mean	0.8941
	25th, 75th	0.8828, 0.8867
	Median	0.8844
	Min, Max	0.8791 , 0.9741
0.5	Mean	0.7947
	25th, 75th	0.7731 , 0.7792
	Median	0.7753
	Min, Max	0.7694 , 0.9467

σ^2_β =41.22 is the sample variance from all βs, and σ^2_β =1.89 is the sample variance of βs excluding extreme βs corresponding to the summed score of 0 and 12.

Table 3.2: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (12) with 10% Level of Measurement Errors

		Classical Rasch Model			MCMC I			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-18.84	981561.4	981562.2	-7.16	10.67	126.58	-7.85	10.98	112.57	-3.60	1.08	206.40
β_3 (25th)	-1.49	-1.94	0.14	0.34	-6.18	12.18	34.22	-1.40	1.31	1.31	-2.16	1.76	2.21
β_6 (50th)	-0.06	0.42	0.05	0.28	3.08	23.83	33.69	0.37	1.40	1.59	0.80	2.44	3.18
β_2 (75th)	1.48	0.86	0.08	0.46	5.17	20.20	33.85	0.60	1.13	1.90	1.25	2.16	2.21
β_{58} (Max)	18.06	18.85	488908.4	488909	7.45	12.12	124.62	7.32	10.87	126.27	3.52	1.13	212.52
Mean Acceptance Rate for βs in MCMC					90.23%			79.47%			80.40%		

MCMC I: σ_{β}^2 = 41.22 was used for the chains of βs.

MCMC II: σ_{β}^2 = 41.22 was used for the chains when initially estimated βs were extreme βs; σ_{β}^2 = 1.89 was used for the chains when initially estimated estimated βs were nonextreme βs.

MCMC III: σ_{β}^2 = 4, arbitrarily chosen variance, was used for the chains of all βs.

Table 3.3: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (73) with 10% Level of Measurement Errors

		Classical Rasch Model			MCMC I			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-17.75	277640.4	277640.4	-6.42	7.66	140.09	-6.34	6.63	140.74	-3.47	1.08	210.11
β_3 (25th)	-1.49	-0.43	0.03	1.14	-2.27	9.96	10.57	-0.32	1.14	2.51	-0.68	2.39	3.04
β_6 (50th)	-0.06	-0.03	0.05	0.05	0.16	12.84	12.88	0.09	1.31	1.34	0.13	2.27	2.30
β_2 (75th)	1.48	0.81	0.06	0.51	3.75	9.86	15.04	0.87	1.04	1.40	1.37	1.60	1.61
β_{58} (Max)	18.06	17.86	541141.8	541141.8	6.58	6.69	138.40	6.54	6.81	139.54	3.64	1.32	209.08
Mean Acceptance Rate for βs in MCMC					87.91%			78.14%			80.50%		

MCMC I: σ_{β}^2 = 25.18 was used for the chains of βs.

MCMC II: σ_{β}^2 = 25.18 was used for the chains when initially estimated βs were extreme βs; σ_{β}^2 = 1.66 was used for the chains when initially estimated estimated βs were nonextreme βs.

MCMC III: σ_{β}^2 = 4, arbitrarily chosen variance, was used for the chains of all βs.

Table 3.4: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (12) with 50% Level of Measurement Errors

		Classical Rasch Model			MCMC I			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-16.59	229927.7	229929.5	-4.45	3.13	184.67	-4.35	3.09	187.33	-3.16	1.24	219.23
β_3 (25th)	-1.49	-0.35	0.03	1.32	-1.44	9.97	9.97	-0.19	0.68	2.36	-1.09	3.26	3.41
β_6 (50th)	-0.06	-0.00	0.02	0.02	0.09	11.38	11.41	0.04	0.62	0.63	-0.21	4.09	4.12
β_2 (75th)	1.48	1.13	0.08	0.20	3.29	4.43	7.72	0.67	0.58	1.22	2.09	1.93	2.31
β_{58} (Max)	18.06	16.59	459039.3	459041.4	4.36	2.90	190.62	4.43	3.00	188.72	3.17	1.19	222.87
Mean Acceptance Rate for βs in MCMC					84.95%			71.21%			79.81%		

MCMC I: σ_{β}^{2} = 10.92 was used for the chains of βs.

MCMC II: σ_{β}^{2} = 10.92 was used for the chains when initially estimated βs were extreme βs; σ_{β}^{2} = 0.67 was used for the chains when initially estimated βs were nonextreme βs.

MCMC III: σ_{β}^{2} = 4, arbitrarily chosen variance, was used for the chains of all βs.

Table 3.5: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (73) with 50% Level of Measurement Errors

		Classical Rasch Model			MCMC I			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-1.63	0.61	266.03	-3.52	2.93	210.27	-1.14	0.52	282.11	-2.56	1.51	237.69
β_3 (25th)	-1.49	-0.00	0.02	2.23	-0.46	7.39	8.45	-0.04	0.73	2.84	-0.02	4.36	6.52
β_6 (50th)	-0.06	-1.12	0.05	1.16	-3.12	2.77	12.18	-0.83	0.56	1.16	-2.51	1.83	7.83
β_2 (75th)	1.48	-0.00	0.02	2.20	-0.13	7.70	10.28	-0.01	0.77	2.98	0.12	3.87	5.70
β_{58} (Max)	18.06	2.43	0.27	244.62	3.67	2.43	209.45	1.31	0.41	280.98	2.94	1.34	230.03
Mean Acceptance Rate for βs in MCMC					83.74%			71.73%			80.33%		

MCMC I: σ_{β}^{2} = 7.62 was used for the chains of βs.

MCMC II: σ_{β}^{2} = 7.62 was used for the chains when initially estimated βs were extreme βs; σ_{β}^{2} = 0.77 was used for the chains when initially estimated estimated βs were nonextreme βs.

MCMC III: σ_{β}^{2} = 4, arbitrarily chosen variance, was used for the chains of all βs.

Table 3.6: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (12) with 100% Level of Measurement Errors

		Classical Rasch Model			MCMC I from _1			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-2.42	0.60	240.98	-1.21	0.27	279.51	-1.20	0.30	280.02	-3.12	1.25	220.38
β_3 (25th)	-1.49	-1.11	0.03	0.17	-0.93	0.32	0.63	-0.94	0.30	0.60	-2.77	1.40	3.06
β_6 (50th)	-0.06	-0.34	0.02	0.10	-0.36	0.43	0.53	-0.40	0.40	0.52	-1.71	2.38	5.12
β_2 (75th)	1.48	-0.70	0.02	4.77	-0.72	0.39	5.21	-0.72	0.38	5.20	-2.46	1.77	17.25
β_{58} (Max)	18.06	1.63	0.05	270.02	1.07	0.29	289.00	1.08	0.30	288.43	2.97	1.41	228.95
Mean Acceptance Rate for βs in MCMC					68.51%			68.53%			81.01%		

MCMC I: $\sigma_{\boldsymbol{\beta}}^2$ = 0.48 was used for the chains of βs.
MCMC II: $\sigma_{\boldsymbol{\beta}}^2$ = 0.48 was used for the chains when initially estimated βs were extreme βs; σ_{β}^2 = 0.48 was used for the chains when initially estimated estimated βs were nonextreme βs.
MCMC III: σ_{β}^2 = 4, arbitrarily chosen variance, was used for the chains of all βs.

Table 3.7: Results from Classical Rasch Model and MCMC Models (from 100,000 Iterations) with Different Variances of Prior βs for a Simulated Dataset (73) with 100% Level of Measurement Errors

		Classical Rasch Model			MCMC I			MCMC II			MCMC III		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-0.71	0.03	296.37	-1.87	2.08	259.95	-0.53	0.43	303.01	-2.02	1.87	254.75
β_3 (25th)	-1.49	-0.00	0.02	2.23	0.11	2.98	5.55	0.00	0.52	2.75	0.02	3.69	5.98
β_6 (50th)	-0.06	-0.00	0.02	0.02	-0.12	3.45	3.45	0.01	0.50	0.51	0.11	3.73	3.76
β_2 (75th)	1.48	-1.12	0.07	6.83	-2.42	1.46	16.62	-0.74	0.43	5.32	-2.44	1.57	16.93
β_{58} (Max)	18.06	2.44	1.10	245.06	2.80	1.23	234.15	1.18	0.34	285.15	2.95	1.40	229.74
Mean Acceptance Rate for βs in MCMC					79.71%			69.43%			80.30%		

MCMC I: $\sigma_{\boldsymbol{\beta}}^2$ = 3.56 was used for the chains of βs.
MCMC II: $\sigma_{\boldsymbol{\beta}}^2$ = 3.56 was used for the chains when initially estimated βs were extreme βs; σ_{β}^2 = 0.54 was used for the chains when initially estimated estimated βs were nonextreme βs.
MCMC III: σ_{β}^2 = 4, arbitrarily chosen variance, was used for the chains of all βs.

- MCMC I is Rasch model accounting for measurement errors through the Bayesian method (hereafter, refers to MCMC Rasch model as a general term), in which I assumed that the variance ($\sigma_{\boldsymbol{\beta}}^2$) of prior distribution for βs is the sample variances of all initial estimates of βs from the classical Rasch model based on a simulated dataset;
- MCMC II is MCMC Rasch model, in which I assumed that the variance ($\sigma_{\boldsymbol{\beta}}^2$) of prior distribution for extreme βs is the variances of all initial estimates of βs from classical Rasch model based on a simulated dataset; the variance ($\sigma_{\boldsymbol{\beta}}^2$) of prior distribution for nonextreme βs is the variance of initial estimates of βs after excluding extreme β estimates from classical Rasch model based on a simulated dataset;
- MCMC III is MCMC Rasch model, in which, I arbitrarily assumed that the variance ($\sigma_{\boldsymbol{\beta}}^2$) of prior distribution is equal to 4.

Additionally, for MCMC I to MCMC III, other factors were as follows: $\sigma_{\boldsymbol{\beta}\,\text{proposed}}^2$=0.5, $\sigma_{\boldsymbol{\delta}}^2$= Sample variance of all initial estimates of δs from Rasch model, $\sigma_{\boldsymbol{\delta}\,\text{proposed}}^2$=0.1, and the variance of measurement errors ($\sigma_{\boldsymbol{u}}^2$) in terms of $\sigma_{\boldsymbol{LW}}^2$ (a%) assumed as what I introduced

Comparing the results among MCMC Rasch models (Table 3.2 to Table 3.7), I made the following observations:

With 10% level of measurement errors, the MSE from MCMC II are smaller for nonextreme βs than MCMC I and MCMC III for both datasets. For extreme βs, MSEs of MCMC II are similar to MCMC I and better than MCMC III.

With 50% level of measurement errors, the results were a little bit different between two datasets. For both datasets, MSE from MCMC II were smaller for nonextreme βs than MCMC I and MCMC III. However, for extreme βs in Dataset 12, MSEs of MCMC II are similar to MCMC I and better than MCMC III. For extreme βs in Dataset 73, MSEs of MCMC II were slightly worse than MCMC I and MCMC III because the introduction of 50% measurement errors changed the true extreme βs to nonextreme βs.

Similarly, with 100% level of measurement errors, MSE from MCMC II were smaller for nonextreme βs than MCMC I and MCMC III for both datasets. MSE from MCMC II were a little bigger for extreme βs than MCMC I and MCMC III because true extreme βs became nonextreme βs in the simulated datasets. In terms of the acceptance rates, MCMC II for three measurement error scenarios had slightly lower acceptance rates than both MCMC I and MCMC III, ranged from 69% to 79% with one exception. In this case (Dataset 12 with 100% level of measurement errors), the acceptance rate of MCMC II was almost exactly the same as that of MCMC I, but better than that of MCMC III.

Therefore, in terms of MCMC models, I recommended the setting in Table 3.8 for the assumptions on the factors: prior $\beta\ and\ \delta$ and proposed $\beta\ and\ \delta$ for the further exploration.

Table 3.8: Setting of prior β and δ and proposed β and δ for MCMC Implementation

Prior β[1,2]	Proposed β	Prior δ[3]	Proposed δ
N (0, σ_{β}^{2})	N (β_{curr}, 0.5)	N(0, σ_{δ}^{2})	N (δ_{curr}, 0.1)

1. For observed extreme βs in a simulated dataset, σ_{β}^{2} is the sample variance of all initially estimated βs from Rasch model

2. For nonextreme βs in a simulated datasets, σ_{β}^{2} is the sample variance of initially estimated βs from Rasch model excluding extreme βs

3. σ_{δ}^{2} is the sample variance of all initially estimated δs from Rasch model

Additionally, I compared the results from classical Rasch model with the results from MCMC Rasch models. I observed three patterns:

- In a situation where a true extreme β remained an extreme β after measurement errors were introduced, the MSE for classical Rasch model is much larger than that for MCMC Rasch models (MSEs are tens-of-thousands for classical Rasch model vs. MSEs are around 300 for MCMC Rasch models).
- In the situation where a true extreme β became nonextreme β after measurement errors were introduced in the datasets, MCMC Rasch models produced a smaller MSEs in most cases.
- In the case where a true nonextreme β remained nonextreme β after measurement errors were introduced in the datasets, MCMC Rasch models performed better for some βs while classical Rasch model performed better for other βs. However, the differences between the best MCMC Rasch model (i.e., MCMC II) and classical Rasch for nonextreme βs are small.

Generally speaking, then, MCMC Rasch models are better than classical Rasch models when measurement errors exist in the datasets.

Markov Chain Monte Carlo Simulations and Results

To recapitulate the previous sections, I have explored prior variances and proposed variances for $\boldsymbol{\beta}\ and\ \boldsymbol{\delta}$. From the practical point of view, another important factor in the MCMC Rasch model implementation is to decide how much measurement errors exist in the real observed dataset. The purpose of the following section is to examine assumptions of measurement errors in the implementation of MCMC Rasch models.

In order to do so, I examined different levels of measurement errors, which I introduced into the dataset (i.e., 10%, 50%, and 100%). Each level of measurement errors was defined as a proportion of variance of $\mathrm{L_X}$ (see Section on Description of Source Data and Simulated Data). Each level of measurement errors consisted of 100 simulated datasets. I examined three levels and, in all, the total number of simulated datasets examined was 300.

Described below was the MCMC scheme with the assumption that measurement error only exists in the dataset at 10% of variance of $\mathrm{L_X}$. Hereafter, I refer to the MCMC Rasch model based on this assumption simply as the "MCMC 10% model" (this is to be distinguished from the 10% level of measurement errors introduced in 100 simulated datasets).

For *each* simulated dataset ($D = 1, \dots, 100$) within each of three levels of measurement errors ($L = 10\%, 50\%, 100\%$).

- First, I ran 10,000 iterations for MCMC Rasch model. In the simulations, I used the above recommended setting for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in Table 3.8.

- Second, I discarded the first 9000 iterations and averaged the last 1000 iterations to estimate $\mathrm{\beta_{n,D,L}}$ $(\mathrm{n} = 1, \dots, \mathrm{N})$ for each of the subjects in that dataset.

I repeated the above two bulleted procedures for 300 simulated datasets (i.e., 100 simulated datasets within each level of measurement errors).

Finally, I averaged $\mathrm{\beta_{n,D,L}}$ over 100 datasets $\left(\mathrm{i.e.}, \frac{\sum_{\mathrm{D}=1}^{100} \mathrm{\beta_{n,D,L}}}{100}\right)$ with each of three levels of measurement errors as the estimate of $\mathrm{\beta_{n,L}}$ for L level of measurement errors. The associated variance of mean for each $\mathrm{\beta_{n,L}}$ was calculated as variance$\left(\frac{\sum_{D=1}^{100} \beta_{n,D,L}}{100}\right)$.

The above scheme, which pertains to the assumption that measurement errors exist in the dataset at 10% of variance of $\mathrm{L_X}$, were repeated for the assumption that measurement errors exist in the dataset at 50% and 100% of variance of $\mathrm{L_X}$.(referred to as, respectively, “MCMC 50% model ” and “MCMC 100% model”)

Similarly, for each dataset within each level of measurement errors, I estimated $\mathrm{\beta_{n,D,L}}$ for each of the subjects in that dataset from classical Rasch model. I averaged $\mathrm{\beta_{n,D,L}}$ across 100 datasets $\left(i.e., \frac{\sum_{D=1}^{100} \beta_{n,D,L}}{100}\right)$ to estimate of $\mathrm{\beta_{n,L}}$ for L level of measurement errors.

Table 3.9 presents the means across 100 datasets for the same five βs in the previous section as well as associated variances and MSEs with 10% level of measurement errors. Table 3.10 presents the same statistics across 100 datasets with 50% level of measurement errors. Finally, presented in Table 3.11 are those statistics across100 datasets with 100% level of measurement errors.

As seen in Table 3.9 for the 10% level of measurement errors, the MCMC 10% model performed best for βs: 25th, 50th, 75th, and maximum among all the MCMC

Table 3.9: Results from Classical Rasch Model and MCMC Models with Three Assumptions on Measurement Errors Exist in the Datasets Across 100 Datasets with 10% Level of Measurement Errors

		Classical Rasch Model			MCMC 10% Model			MCMC 50% Model			MCMC 100% Model		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-18.16	4737.65	4737.71	-6.98	0.06	119.91	-6.98	0.07	119.76	-6.62	0.07	127.81
β_3 (25th)	-1.49	-1.56	126.90	126.90	-0.97	0.01	0.28	-0.63	0.02	0.74	-0.52	0.01	0.95
β_6 (50th)	-0.06	-0.1	0.001	0.002	-0.003	0.01	0.01	0.06	0.01	0.03	0.01	0.01	0.02
β_2 (75th)	1.48	1.46	61.89	61.89	1.09	0.01	0.16	0.55	0.01	0.87	0.45	0.01	1.06
β_{58} (Max)	18.06	18.24	5062.09	5062.12	7.20	0.07	117.96	6.98	0.07	122.74	6.63	0.07	130.72

MCMC 10% were MCMC results if measurement errors in the simulated datasets were assumed 10% of the variance of L_X.

MCMC 50% were MCMC results if measurement errors in the simulated datasets were assumed 50% of the variance of L_X.

MCMC 100% were MCMC results if measurement errors in the simulated datasets were assumed 100% of the variance of L_X.

Table 3.10: Results from Classical Rasch Model and MCMC Models with Three Assumptions on Measurement Errors Exist in the Datasets Across 100 Datasets with 50% Level of Measurement Errors

		Classical Rasch Model			MCMC 10% Model			MCMC 50% Model			MCMC 100% Model		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-7.80	1011.77	1114.34	-2.54	0.01	236.78	-2.40	0.01	240.91	-2.33	0.01	243.29
β_3 (25th)	-1.49	-0.20	0.000	1.65	-0.17	0.01	1.75	-0.13	0.01	1.85	-0.1	0.01	1.94
β_6 (50th)	-0.06	0.07	0.000	0.02	0.05	0.01	0.02	0.03	0.01	0.01	0.03	0.01	0.01
β_2 (75th)	1.48	0.32	0.000	1.34	0.29	0.01	1.41	0.22	0.01	1.58	0.18	0.01	1.69
β_{58} (Max)	18.06	9.23	1743.11	1820.98	2.84	0.02	231.63	2.75	0.02	234.38	2.62	0.02	238.38

MCMC 10% were MCMC results if measurement errors in the simulated datasets were assumed 10% of the variance of L_X.

MCMC 50% were MCMC results if measurement errors in the simulated datasets were assumed 50% of the variance of L_X.

MCMC 100% were MCMC results if measurement errors in the simulated datasets were assumed 100% of the variance of L_X.

Table3.11: Results from Classical Rasch Model and MCMC Models with Three Assumptions on Measurement Errors Exist in the Datasets Across 100 Datasets with 100% Level of Measurement Errors

		Classical Rasch Model			MCMC 10% Model			MCMC 50% Model			MCMC 100% Model		
Selected β	True β	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$	$\hat{\beta}$	$\widehat{var}(\hat{\beta})$	$\widehat{MSE}(\hat{\beta})$
β_{40} (Min)	-17.92	-1.87	105.28	363.12	-1.03	0.004	285.33	-1.01	0.004	286.13	-1.01	0.004	286.24
β_3 (25th)	-1.49	-0.13	0.000	1.85	-0.10	0.004	1.92	-0.10	0.004	1.94	-0.09	0.004	1.96
β_6 (50th)	-0.06	-0.03	0.001	0.001	-0.03	0.004	0.01	-0.04	0.004	0.005	-0.03	0.004	0.01
β_2 (75th)	1.48	0.15	0.000	1.75	0.12	0.004	1.84	0.13	0.004	1.81	0.11	0.005	1.87
β_{58} (Max)	18.06	2.06	122.66	378.46	1.09	0.004	287.92	1.10	0.004	287.67	1.07	0.004	288.70

MCMC 10% were MCMC results if measurement errors in the simulated datasets were assumed 10% of the variance of L_X.

MCMC 50% were MCMC results if measurement errors in the simulated datasets were assumed 50% of the variance of L_X.

MCMC 100% were MCMC results if measurement errors in the simulated datasets were assumed 100% of the variance of L_X.

models. Regarding the minimum β, the MCMC 50% model provided a smaller MSE, although the difference between this model and other MCMC models was not substantial. In general, differences in MSE results obtained from MCMC models were small and were within random sampling variability. In comparing the results from classical Rasch model with the results from MCMC models, MCMC models provided substantially smaller MSEs for the βs with exception of 50th β. Notably, this was especially the case for extreme βs. For example, the MSE for minimum β was 4737.71 for classical Rasch model vs.119.76 to 127.81 for MCMC models. The MSE for maximum β was 5062.12 for classical Rasch model vs. 117.96 to 130.72 for MCMC models. Although classical Rasch model provided the better estimate for the 50th β, the difference between the two are minimal in terms of MSEs (0.002 for classical Rasch model vs. 0.01 to 0.03 for MCMC models).

In addition, for 25th and 50th βs, the MSEs from classical Rasch model were also distinctly bigger than those from MCMC Rasch models. The reason for this was that after 10% level of measurement error was introduced, nonextreme 25th true β was changed to extreme β in three of 100 simulated datasets and nonextreme 75th true β changed to extreme β in two of 100 simulated datasets. The Monte Carlo estimates of the two βs from Classical Rasch model were associated with much bigger variances while the Monte Carlo estimates of these βs from MCMC Rasch models were associated with much smaller variances.

In the case of 50% level of measurement errors (Table 3.10), the MSEs from MCMC 10% Model provided the best estimates for minimum, 25th, 75th, and maximum βs among all three MCMC models. The MSE for 50th β was smallest from MCMC 50%

or 100% models. However, the difference in MSE for 50th β is very small across three MCMC scenarios. Again, the MSEs were similar in all MCMC models. Compared with the results from classical Rasch model, the MSEs from MCMC models were very similar for nonextreme βs. And, the differences were miniscule. Once more, for the extreme βs, the MSEs were much smaller for the results from MCMC models (in the hundreds) in relation to classical Rasch model (thousands).

In the case of 100% level of measurement errors (Table 3.11), the MSEs from MCMC 10% model provided the best estimates for minimum and 25th βs among all three MCMC models; the MSEs for 50th, 75th, and maximum β were smallest from MCMC 50% model. Regardless of the different MSEs observed in three MCMC models, the differences were not significant. Compared with classical Rasch model, the results from MCMC models were also better estimates for the extreme βs. For nonextreme βs, classical Rasch model provided a little bit better estimates.

Discussion

Because the extreme βs are nonestimable using the classical Rasch model, I found that the variances of the estimates for extreme βs are tremendously large (its magnitude is in the tens-of-thousands, which triggers the algorithm to stop under the convergence criteria). For this reason, I used Bayesian methods to incorporate measurement errors into the Rasch model (resulting in what I refer to as “MCMC Rasch model”). Most importantly, I obtained the estimates of extreme βs at a better range than classical Rasch model, and these estimates had reasonable variances (in the hundreds). This enables us to make formal statistical inferences. The now estimable extreme βs prevent the missing scores in Rasch field so that the statistical power is retained, which allows statistical analysis based on the β scores. In other words, this prevents the missing data in the conversion between the summed scores and linear Rasch scores. However for nonextreme βs, MCMC Rasch models sometimes performed a little worse than classical Rasch model. But the differences are miniscule and well within the randomly sampling variation.

I compared MCMC 10%, 50%, and 100% models and found that the results in terms of MSE were very similar. So, in order for us to account for measurement errors in a real observed dichotomous response dataset, how much measurement errors may be assumed for the MCMC appears to be inconsequential so long as we use Rasch model accounting for measurement errors. Using this procedure, we obtained reasonable estimates of βs for additional statistical analysis.

Conclusions

I explored factors that affected MCMC implementation. In doing so, I found that the following factors largely determined the behavior of chains: prior variance of β, the proposed variance for β, and the proposed variance for δ. The optimal setting of these three factors was:

- the proposed variance for β is 0.5;
- the proposed variance for δ is 0.1;
- prior variance of β is either the sample variance of all initially estimated βs from Rasch model for extreme βs or the sample variance of initially estimated βs from Rasch model excluding extreme βs for nonextreme βs.

Using the optimal setting, I further explored the measurement errors, which may be assumed as 50% of variance of observed $\mathbf{L_W}$ for the implementation of MCMC.

In the next chapter, I will use the setting of factors identified in this chapter and apply it to the SST data collected by Dr. Tashjian to evaluate the effect of the measurement errors on minimum clinically important difference.

Appendices

Appendix A: Likelihood Functions

In the Rasch model for dichotomous data, the probability of outcome $X_{ni} = 1$ is given by

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)}$$

where

X_{ni} is the response (1=Yes, 0= No) for Person n to Item i;

β_n is the ability of person n (*n = 1, …, N);*

δ_i is the difficulty of Item i (*i = 1, …, I);*

P(·) is the probability that Person n has a true or observed response to Item i.

If measurement errors are not considered in the Rasch model, the likelihood function based on the observed data is

$$f(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\delta}) = LH(\boldsymbol{\beta},\boldsymbol{\delta}) = \prod_{n,i} P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta})^{X_{ni}} \times \left(1 - P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta})\right)^{1-X_{ni}}$$

where X_{ni} is the response by n^th^ subject for i^th^ question.

If we introduced measurement errors into Rasch model as in the form of

$$\boldsymbol{L_W} = \boldsymbol{L_X} + \boldsymbol{U}$$

the complete likelihood function based on the observed data is

$$LH(\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{X},\boldsymbol{L_X}) = \{f(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\delta})\} \times \{f(\boldsymbol{\beta}) * f(\boldsymbol{\delta})\} \times \{f(\boldsymbol{X}|\boldsymbol{L_X})\} \times \{f(\boldsymbol{L_X}|\boldsymbol{L_W})\}$$

That is $\{Rasch\ Model\} \times \{Prior\} \times \{\text{Bernoulli}\} \times \{Conditional\ \}$

where

$$f(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\delta}) = LH(\boldsymbol{\beta},\boldsymbol{\delta}) = \prod_{n,i} P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta})^{X_{ni}}(1 - P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta}))^{1-X_{ni}}$$

$$f(\boldsymbol{\beta}) \propto exp\left(\boldsymbol{\beta}^T {\textstyle\sum}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}/2\right)$$

$$f(\boldsymbol{\delta}) \propto exp(\boldsymbol{\delta}^T {\textstyle\sum}_{\boldsymbol{\delta}}^{-1} \boldsymbol{\delta}/2)$$

$$f(\boldsymbol{X}|\boldsymbol{L_X}) = \prod_{ni} P\left(X_{ni}|L_{X_{ni}}\right)^{X_{ni}} \times \left(1 - P\left(X_{ni}|L_{X_{ni}}\right)\right)^{1-X_{ni}}$$

$$\boldsymbol{L_X}|\boldsymbol{L_W} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{L_X}}\mathbf{1} + \frac{\sigma_{\boldsymbol{L_X}}^2}{\sigma_{\boldsymbol{L_X}}^2 + \sigma_u^2}\boldsymbol{I}\left(\boldsymbol{L_W} - \boldsymbol{\mu}_{\boldsymbol{L_X}}\right), \sigma_{\boldsymbol{L_X}}^2\left(1 - \frac{\sigma_{\boldsymbol{L_X}}^2}{\sigma_{\boldsymbol{L_X}}^2 + \sigma_u^2}\right)\boldsymbol{I}\right)$$

Appendix B: Gibbs Samplers and Metropolis-Hastings Algorithm

The Gibbs samplers along with the Metropolis-Hastings algorithm were used to generate estimates of the parameters from the complete likelihood function. For each Gibbs sampler generated at following steps, the general form of Metropolis-Hastings Algorithm to accept a Gibbs Sampler θ from a proposed distribution q(.) is:

$$\alpha = min\left\{1, \frac{LH\ (\theta_{candidate})q(\theta_{current}|\theta_{candidate})}{LH\ (\theta_{current})q(\theta_{candidate}|\theta_{current})}\right\}$$

Then compare α with probability generated from uniform distribution. If α is greater than the probability from uniform distribution, then update $\theta_{current}$ with $\theta_{candidate}$. Otherwise, keep $\theta_{current}$ as it is.

The steps to generate Gibbs Samplers and associated Metropolis-Hastings Algorithms are:

1. Obtain initial starting values of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, e.g. estimates the two parameters from the simulated dataset using Rasch model ignoring measurement error. Set these values as the current values $\boldsymbol{\beta}_{current}$ and $\boldsymbol{\delta}_{current}$.
2. Generate a candidate of the true Logit $\boldsymbol{L_X}$ according to the distribution of $\boldsymbol{L_X}|\boldsymbol{L_W}$, given the observed $\boldsymbol{L_W}$ equal to $\boldsymbol{\beta}_{current}$ minus $\boldsymbol{\delta}_{current}$.

The acceptance ratio for each element of candidate L_X[j], where j= 1, 2, …, $N \times I$, is

$$\frac{LH\ (L_{Xcandidate}[j])q(L_{Xcurrent}[j]|L_{Xcandidate}[j])}{LH\ (L_{Xcurrent}[j])q(L_{Xcandidate}[j]|L_{Xcurrent}[j])} = \frac{f(X|L_{Xcandidate}[j])}{f(X|L_{Xcurrent}[j])}$$

$$= \frac{(P_{Xcandidate}[j])^{X_{current}} \times (1 - P_{Xcandidate}[j])^{(1-X_{current})}}{(P_{Xcurrent}[j])^{X_{current}} \times (1 - P_{Xcurrent}[j])^{(1-X_{current})}}$$

where

$$P_{Xcandidate}[j] = P(X_{candidate}[j] = 1) = \frac{exp(L_{Xcandidate}\,[j])}{1 + exp(L_{Xcandidate}\,[j])}$$

$$P_{Xcurrent}[j] = P(X_{current}[j] = 1) = \frac{exp(L_{Xcurrent}\,[j])}{1 + exp(L_{Xcurrent}\,[j])}$$

3. Generate a candidate of true response $\boldsymbol{X}_{candidate}$ based on updated $\boldsymbol{L}_{\boldsymbol{X}current}$. $\boldsymbol{X}_{candidate}$ follows logistic distribution. $\boldsymbol{X}_{candidate}$ follows Logistic distribution.

The acceptance ratio for each element of $X_{candidate}[j]$, where j= 1, 2, …, $N \times I$, is

$$\frac{LH\ (X_{candidate}[j])q(X_{current}[j]|X_{candidate}[j])}{LH\ (X_{current}[j])q(X_{candidate}[j]|X_{current}[j])}$$

$$= \left(exp(\beta_{current}[n] - \delta_{current}[i])\right)^{X_{candidate}[j] - X_{current}[j]}$$

where $j = (n - 1) \times 12 + i$

$$for\ n^{th} subject\ and\ i^{th} item, denoted\ as\ "ni"\ sometimes$$

4. Generate candidate of $\boldsymbol{\beta}$ based on current $\boldsymbol{\beta}_{\text{current}}$ using the normal distribution $N\left(\boldsymbol{\beta}_{current}, \sigma_{\boldsymbol{\beta}\ proposed}^2\right)$. The $\sigma_{\boldsymbol{\beta}\ proposed}^2$ should be small. The prior distribution of $\boldsymbol{\beta}$ is $N\left(0\mathbf{1}, \sigma_{\boldsymbol{\beta}}^2\boldsymbol{I}\right)$.

The acceptance ratio for each element of $\beta_{candidate}[n]$, where n= 1, 2, …, $Nth\ subject$, is

$$\frac{LH\ (\beta_{candidate}[n])q(\beta_{current}[n]|\beta_{candidate}[n])}{LH\ (\beta_{current}[n])q(\beta_{candidate}[n]|\beta_{current}[n])} =$$

$$\prod_{i=1}^{I=12}\left(\frac{exp\{\beta_{candidate}[n]-\delta_{current}[i]\}}{exp\{\beta_{current}[n]-\delta_{current}[i]\}}\right)^{X_{current}[ni]}$$

$$\times\prod_{i=1}^{I=12}\left(\frac{1+exp\{\beta_{current}[n]-\delta_{current}[i]\}}{1+exp\{\beta_{candidate}[n]-\delta_{current}[i]\}}\right)$$

$$\times\ exp\left\{-\frac{(\beta_{candidate}[n])^2-(\beta_{current}[n])^2}{2\times\sigma_{\boldsymbol{\beta}}^2}\right\}$$

5. Generate candidate of $\boldsymbol{\delta}$ based on current $\boldsymbol{\delta}_{current}$ using the normal distribution $N(\boldsymbol{\delta}_{current}, \sigma_{\boldsymbol{\delta}\,proposed}^2)$. $\sigma_{\boldsymbol{\delta}\,proposed}^2$ should be small. The prior distribution of $\boldsymbol{\delta}$ is $N(0\mathbf{1}, \sigma_{\boldsymbol{\delta}}^2\boldsymbol{I})$.

The acceptance ratio for each element of $\delta_{candidate}[i]$, where i $= 1, 2, \ldots, Ith\ item$ is:

$$\frac{LH\ (\delta_{candidate}[i])q(\delta_{current}[i]|\delta_{candidate}[i])}{LH\ (\delta_{current}[i])q(\delta_{candidate}[i]|\delta_{current}[i])} =$$

$$\prod_{n=1}^{N}\left(\frac{exp\{\beta_{current}[n]-\delta_{candidate}[i]\}}{exp\{\beta_{current}[n]-\delta_{current}[i]\}}\right)^{X_{current}[ni]}$$

$$\times\prod_{n=1}^{N}\left(\frac{1+exp\{\beta_{current}[n]-\delta_{current}[i]\}}{1+exp\{\beta_{current}[n]-\delta_{candidate}[i]\}}\right)$$

$$\times\ exp\left\{-\frac{(\delta_{candidate}[i])^2-(\delta_{current}[i])^2}{2\times\sigma_{\boldsymbol{\delta}}^2}\right\}$$

Repeat Step 2 to 5 for a large number of times, or until convergence.

From the generation of Gibbs samplers, the chain for Logit $\mathbf{L_X}$ is an independent chain. For an independent chain where proposal transitional kernel q(y|x=q(y)it may seem that the independence from the previous state disagrees with Markovian property of the chain. Actually, q is just a proposal that is combined with an acceptance probability of α to give the transitional p of the algorithm. This transition depends on the previous state, thus, preserves the Markovian properties.[19] The chains for β and δ are symmetric chains, in this case when $q(y|x) = q(|y-x|)$, e.g., the normal distribution with mean x, then we have $q(y|x) = q(x|y)$.

References

1. Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Seymour S. *Measurement Errors in Surveys.* John Wiley & Sons, 2004.

2. Joseph Gfroerer JL, and Teresa Parsley. Studies of Nonresponse and Measurement Error in the National Household Survey on Drug Abuse. *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. US National Institute on Drug Abuse, Division of Epidemiology and Prevention Research: Rockville, MD, 1997.

3. Carroll RJ, Ruppert D, Stefanski LA. Chapter 2 Important Concept. *Measurement Error in Nonlinear Models A Modern Perspective* (Second edn). Chapman & Hall/CRC, 1995.

4. Carroll RJ, Ruppert D, Stefanski LA. Chapter 3 Linear Regression and Attenuation. *Measurement Error in Nonlinear Models A Modern Perspective* (Second edn). Chapman & Hall/CRC, 1995.

5. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology.* Chapman & Hall/CRC, 2004.

6. May K, Nicewander WA. Measuring change conventionally and adaptively. *Educational and Psychological Measurement* 1998; **58**: 882-897.

7. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-40): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997; **50**: 451-461.

8. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998; **51**: 1203-1214.

9. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs. raw scores in measuring change in health. *Medical are* 2004; **42**: I25-I36.

10. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, Gregg PJ. A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Quality of Life Research* 2004; **13**: 331-338.

11. Geyer CJ. Practical Markov Chain Monte Carlo. *Statistical Science* 1992; **7**: 473-483.

12. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC, 1996.

13. University of Washington. Simple Shoulder Test, Available at http://www.orthop.washington.edu/uw/simpleshoulder/tabID__3376/ItemID__186/PageID__356/qview__true/Articles/Default.aspx, Accessed Jan 7, 2012.

14. Sheng X. A Bayesian model for measurement errors in diagnosis of rheumatoid arthritis. *Communications in Statistics - Theory and Methods* 2009; **38**: 3419 - 3431.

15. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010; **92**: 296-303.

16. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009; **18**: 927-932.

17. Gamerman D, Lopes HF. *Markov Chain Monte Carlo.* (2nd edn). Chapman &Hall/CRC, 2006.

18. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* (2nd edn). Chapman & Hall/CRC, 2004.

19. Gamerman D, Lopes HF. *Chapter 6 Metropolis-Hastings algorithm, Markov Chain Monte Carlo.* (2nd edn). Chapman &Hall/CRC, 2006.

20. Gamerman D, Lopes HF. *Chapter 4 Markov Chains, Markov Chain Monte Carlo.* (2nd edn). Chapman &Hall/CRC, 2006.

21. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**: 97-109.

CHAPTER 4

MINIMUM CLINICALLY IMPORTANT DIFFERENCE FOR SIMPLE SHOULDER TEST IN RASCH SCORE ACCOUNTING FOR MEASUREMENT ERRORS

Abstract

Measurement errors result from classical sources such as uncertainty of subjects' self-reports on questionnaires and restricted response categories offered to them. In this chapter, I took a Bayesian approach using the Markov Chain Monte Carlo (MCMC) method to account for measurement errors in the Rasch model. Our data came from patients treated conservatively for rotator cuff tendonitis or tearing. I applied Rasch model accounting for measurement errors to the Simple Shoulder Test (SST), and evaluated the effect of measurement errors in the determination of minimum clinically important difference (MCID). Patient groups were defined by anchored questions, and the MCID was defined as the statistically significant difference of mean change from baseline between patient groups. According to the 15-item anchored question, the difference (95% CI; p value) of the mean change from baseline (CFB) between two groups was 1.97 (-0.17, 4.10; 0.0693) for rescaled Rasch score and 8.54 (1.78, 15.30; 0.0156) for rescaled MCMC Rasch SST score. According to the four-item anchored question, the difference (95% CI; p value) of the mean CFB between two groups 2.38 (1.03, 3.74; 0.0009) for rescaled Rasch score and 1.20 (-7.40, 9.81; 0.7810) for rescaled

MCMC Rasch score. The inconsistencies in MCIDs between MCMC Rasch scores and classical Rasch scores may be due to the bias of estimates of Rasch scores when measurement errors are left unconsidered. Additionally, I found that the implementation of the Rasch model accounting for measurement errors model is feasible.

Introduction

Measurement errors exist whenever there is an attempt to measure. Potentially they may undermine analysis and lead to inaccurate or incorrect conclusions. Questionnaires may introduce measurement errors because typically they offer only restricted response categories. For example, binary response questions such as those that comprise the Simple Shoulder Test (SST), force subjects to make a choice among predefined response categories, even though their true choice may fall in-between categories. This suggests that measurement errors are much more than simple recording or instrument error. They encompass many different sources of variability.[1]

As is the convention for many questionnaires, the sum of answers over all SST questions is used to characterize each subject's shoulder function, denoted as SST summed score. The higher the SST summed score, the better the shoulder function is. Nevertheless, there are problems inherent in the summed score, including the following: (1) there are no natural numerical upper or lower limits to health status, and for this reason, a zero value does not have an inherent meaning; and (2) psychometric instruments (i.e., questionnaires) do not necessarily have interval characteristics (i.e., scaled in a linear fashion) . In other words, one cannot assume that the difference between scores of 10 and 12 is equal to the difference between scores of 1 and 3. This leads to difficulty in interpreting changes in the score at follow-up visits. Fortunately, to deal with the nonlinear property of summed scores for SST, the Rasch model converts ordinal measures into linear measures.[2] Generally speaking, the Rasch model is a transformation from SST summed scores into linear space, provided that the fit to the model and several assumptions about the model are appropriate.[3] Through the Rasch model for dichotomous data, the probability of the outcome $X_{ni} = 1$ is given by:

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)} \qquad \textit{(4.1)}$$

where

X_{ni} is the response (1=Yes, 0= No) for Person n to Item i;

β_n is the ability of Person n (*n = 1, …, N);* which hereafter will be referred as the "Rasch score";

δ_i is the difficulty of Item i (*i = 1, …, I);*

P(•) is the probability that Person n has a true or observed response to Item i.

In SST, the person ability β_n $(n = 1, \dots, N)$ indicates how able n^{th}person can perform items*;* the item difficulty δ_i $(i = 1, \dots, I)$ indicates the difficulty of i^{th}item.

In this chapter, I applied Rasch model accounting for measurement errors[4] to evaluate the effect of measurement errors on the determination of minimal clinically important difference (MCID) via Bayesian approach.

In order to facilitate the assessment of effectiveness of an intervention, the determination of MCID plays an important role in bringing statistical significance and clinical significance together. The MCID represents "the smallest difference in score in the domain which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management.”[5]

Background

Measurement Errors in Questionnaires

Inevitably, as with any questionnaire, measurement errors arise with patients’ self reports on the SST, including the following:[1, 6]

- the mode of interview;
- wording of questions;
- interviewer behavior;
- sensitivity of information requested;
- respondents' recall;
- coding errors;
- choice from a limited set of answers, i.e., in the case of SST questionnaire, no option between YES and No as a choice for subjects.

Measurement errors may distort an analysis due to attenuation to null and loss of power. Without accounting for measurement errors, the analyses may result in bias of estimates (i.e., real effects are hidden, observed data exhibit relationship that are not present in the error-free data, and the signs of estimated coefficients are reversed relative to the case with no measurement error). [6, 7]

As mentioned in the Introduction, the summed score of questionnaire typically is used in analyses. However, the summed score of the SST may lack linear measurement properties, and for this reason the change in scores may not reflect what we intend to measure. This is a well-known problem, and the literature provides substantial evidence to support the use of the Rasch model to compare outcomes measured by questionnaires, either among patients, within patients (change scores), or between treatments. [8-12] Nevertheless, preexisting studies on Rasch Models in the literature have not dealt with virtually ubiquitous measurement errors. This chapter extends fundamental considerations of measurement errors in Rasch modeling and addresses the following issues in dichotomous response data in SST questionnaire: nonlinear summed scores and measurement errors.

Rasch Model Accounting for Measurement Errors

In order to account for the effects of measurement errors in questionnaires, I proposed to incorporate measurement errors into the Rasch model under a preestablished framework of measurement errors.[1] Typically, measurement errors in self-reported questionnaires may be modeled as classical measurement error. In the classical measurement error model for item response data,[4] I introduce the following terms: an unobserved true response to Item i by Person n ($\mathrm{X_{ni}}$), which is measured by some individual-specific random error (X in matrix notation). The observed response $\mathrm{W_{ni}}$ (W in matrix notation), which may be different from $\mathrm{X_{ni}}$ because it is mainly caused by inaccurate information obtained in the self-report questionnaires.

Through the logistic regression model, the dichotomous Rasch model[13] has the following linear form:

$$logit[P(X_{ni} = 1)] = \beta_n - \delta_i \qquad (4.2)$$

The $\boldsymbol{L_W}$ was defined as $logit\big(P(\boldsymbol{W})\big)$ and the $\boldsymbol{L_X}$was defined as $logit\big(P(\boldsymbol{X})\big)$, where P(·) is the probability that Person n has a true or observed response equal to 1to Item i. The Rasch model accounting for the classical measurement error structure may be modeled in the following way: the logit transformed latent variables $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ linked the true response $\boldsymbol{X}$ and the observed response $\boldsymbol{W}.$

$$\boldsymbol{L_W} = \boldsymbol{L_X} + \boldsymbol{U} \tag{4.3}$$

$$L_{X_{ni}} = \beta_n - \delta_i \tag{4.4}$$

where each component U_{ni} of $\mathbf{U}$ are i.i.d. random variables of measurement error: $\boldsymbol{U} \sim N(0\mathbf{1}, \sigma_u^2\boldsymbol{I})$, where σ_u^2 is the variance of the classical measurement error; and $\boldsymbol{L_X} \sim N(\boldsymbol{\mu}_{\boldsymbol{L_X}}, \sigma_{\boldsymbol{L_X}}^2\boldsymbol{I})$, where $\boldsymbol{\mu}_{\boldsymbol{L_X}}$ is the mean of $\boldsymbol{L_X}$, and $\sigma_{\boldsymbol{L_X}}^2$ is the variance of $\boldsymbol{L_X}.$

Description of Source Data

The SST questionnaire is a well-established patient-reported measure used to evaluate shoulder function.[14] It consists of 12 yes/no questions, with the answer “yes” coded as one (1) and answer “no” coded as zero (0). The sum of answers over all SST questions is used to characterize each subject’s shoulder function, denoted as SST summed score. The higher the SST summed score, the better is the shoulder function. The range of SST summed score is from 0 to 12.

The SST data were obtained from a larger dataset collected for a study whose Principal Investigator was Robert Z. Tashjian, MD, Department of Orthopedics, University of Utah, School of Medicine, Salt Lake City, UT.[15, 16] Prior to data collection, Dr. Tashjian's study received approval from the University of Utah Institutional Review Board. Again, permission to use these data was obtained from Dr. Tashjian (i.e., PI of the study for which data were collected) through Dr. Christy Porucznik (my dissertation chair).

The dataset was based on 81 patients with rotator cuff tendonitis or tearing who were enrolled in the study and treated with nonoperative modalities. All of them completed SST questionnaires both at Screening and Week 6 follow-up and provided the response to two anchored questions at Week 6 Follow-up. The MCID for the SST summed score was determined in a study for patients with rotator cuff disease.[15] In addition, the MCID for the Rasch SST score was determined in Chapter 2.

I treated this dataset as observed response dataset and assumed that measurement errors existed in the dataset. I followed the proposed Rasch model accounting for measurement errors to obtain the person ability scores.

Methods

Bayesian Methods through Markov Chain Monte Carlo Process

Equation (4.3) allows us to assume that measurement errors are normally distributed in the Logit dimension. Because the direct model of distribution for measurement errors is not available, I adopted the Markov Chain Monte Carlo (MCMC) approach to find the estimates of Rasch model parameters. In order to follow the framework of Bayesian approach, the complete likelihood function (see Appendix A for details) in this case is

$$LH(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{L_X}) = \{f(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\delta})\} \times \{f(\boldsymbol{\beta}) * f(\boldsymbol{\delta})\} \times \{f(\boldsymbol{X}|\boldsymbol{L_X})\} \times \{f(\boldsymbol{L_X}|\boldsymbol{L_W})\}$$

To obtain the model parameter $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ estimates through the complete likelihood function, the Gibbs Sampler and Metropolis-Hastings Algorithm were used (see Appendix B)[17].

In Chapter 3, I identified the factors that affect the implementation of MCMC simulation in Rasch model accounting for measurement errors: the variance of measurement errors ($\sigma_{\boldsymbol{u}}^2$), prior variance ($\sigma_{\boldsymbol{\beta}}^2$ and $\sigma_{\boldsymbol{\delta}}^2$), and proposal variances for $\boldsymbol{\beta}\ and\ \boldsymbol{\delta}$. I applied the optimal setting for these factors to the observed SST data as specified in Table 4.1 to generate the MCMC chains for SST data at Week 0 and Week 6 separately. In doing so, I obtained the posterior distribution of $\boldsymbol{\beta}$ for both Week 0 and Week 6.

Evaluation of Convergence for MCMC Chain

After obtaining the posterior distribution of $\boldsymbol{\beta}$, I evaluated the convergence of chain for each element β_n of $\boldsymbol{\beta}$ by using the potential scale reduction, of which the idea is to test whether dispersion within chains is larger than dispersion between chains. The potential scale reduction was estimated by $\hat{R} = \sqrt{\frac{\widehat{var}^{+}(\beta_n)}{W}}$.[18, 19]

Table 4.1: Optimal Setting for the Factors that Affect the MCMC Implementation

Prior $\beta^{1,2}$	Proposed β	Prior δ^3	Proposed δ	Measurement Error U[4]
N (0, σ_β^2)	N (β_{curr}, 0.5)	N(0, σ_δ^2)	N (δ_{curr}, 0.1)	$\mathrm{N}(0\mathbf{1}, \sigma_{\mathrm{u}}^2\boldsymbol{I})$,

1. For βs corresponding to extreme summed scores (0,12), σ_β^2 is the sample variance of all initially estimated βs from Rasch model
2. For βs corresponding to the summed scores 1 to 11, σ_β^2 is the sample variance of initially estimated βs from Rasch model excluding βs corresponding to extremely summed scores.
3. σ_δ^2 is the sample variance of all initially estimated δs from Rasch model
4. σ_u^2 is assumed as 50% of the variance of observed $\boldsymbol{L_W}$ from Rasch model

In order to calculate it, I need to generate M parallel chains for each β_n with each of length L (after discarding the first half of the simulations). I labeled simulation draws as $\beta_{n(lm)}$(l=1,…,L, m=1,…,M). The between-chain variance (B) and within-chain variance (W) were calculated as

$$B = \frac{L}{M-1}\sum_{m=1}^{M}\left(\bar{\beta}_{n(.m)} - \bar{\beta}_{n(..)}\right)^2$$

$$W = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{L-1}\sum_{l=1}^{L}\left(\beta_{n(lm)} - \bar{\beta}_{n(.m)}\right)^2\right)$$

The marginal posterior variance of β_n is $\widehat{var}^{+}(\beta_n) = \frac{L-1}{L}W + \frac{1}{L}B$.

MCID in Rasch Scores Accounting for Measurement Errors

Through the Rasch model accounting for measurement errors, I obtained the estimates of Rasch scores after accounting for measurement errors for each subject, for both Week 0 and Week 6. Because I obtained these estimates via MCMC method, I refer to these as MCMC Rasch Scores from now on. Then, the minimum clinically important difference (MCID)[5, 20] was defined as the difference of the change from baseline in MCMC Rasch scores between two patient groups (i.e., No Change vs. Minimal Improvement) if the difference showed statistical significance at 0.05 level. The two patient groups were determined by 15-item and four-item anchored questions (see Appendix C for the details of the two anchored questions), which were answered by patients who completed SST questionnaires at both Week 0 and Week 6.

Results

Convergence of MCMC Chain and Acceptance Rates of β

I generated M=5 parallel chains with length L =10000 (after deleting the first half of iterations) to calculate the potential reduction scale. In order to present the potential reduction scales, I selected five β parameters according to quartiles, minimum, and maximum values from the distribution of 81 initially estimated βs from Rasch model (corresponding to 81 subjects) using Week 0 data. According to Gelman (2004), the value for $\widehat{R}$ below 1.1 is acceptable for the claim of convergence for chains).[19]

From Table 4.2, I found that the potential reduction scales were all below 1.1 for both Weeks 0 and 6 data starting from 10,000 iterations. I treated the chain as having converged at 10,000 iterations for both weeks.

The median of acceptance rates for all βs is 79% for Week 0 model and 78% for Week 6 model. According to Roberts, et al., the "optimal efficiency" is achieved at an acceptance rate of 0.234.[21] According to Gamerman [22], multiple sources indicate to the direction of acceptance rates between 20% to 50%. However, Geyer and Thompson [23]warned that attempting to reduce the acceptance rate below 70% would keep the

Table 4.2: The Potential Reduction Scales from the Five Parallel Chains

Number of Iterations	β_{40} (Min)	β_3 (25th)	β_6 (50th)	β_2 (75th)	β_{58} (Max)
Week 0					
100	11.880	1.609	1.770	2.351	11.804
500	2.346	1.187	1.325	1.079	2.878
5000	1.046	1.044	1.026	1.035	1.021
10000	1.017	1.002	1.006	1.012	1.009
20000	1.007	1.004	1.005	1.002	1.009
Week 6					
100	1.360	2.402	2.032	1.610	20.405
500	1.708	1.272	1.183	1.225	4.173
5000	1.007	1.057	1.028	1.201	1.132
10000	1.007	1.083	1.039	1.034	1.017
20000	1.015	1.002	1.042	1.044	1.038

sampler from ever visiting part of the state space. The acceptance rates for all βs in the study are therefore considered acceptable.

Rasch Scores Accounting for Measurement Errors

From one of five chains with 20,000 iterations, I dropped the first 10,000 iterations.[24] I used the average of the second 10,000 iterations for each β as the estimate of personal score for each subject. The MCMC Rasch scores along with Rasch score was shown in Figure 4.1 for Week 0 data. Figure 4.2 showed rescaled Rasch scores as well as rescaled Rasch scores for Week 0 data.

In Figure 4.1, each box on the green line represents the box-plot for several MCMC Rasch scores that corresponded to a SST summed score. Rasch model accounting for measurement errors had a similar shape to Rasch model. Due to the measurement errors, the curve was flatter for Rasch model accounting for measurement errors than the classical Rasch model. Most importantly, under the assumptions of the proposed models,

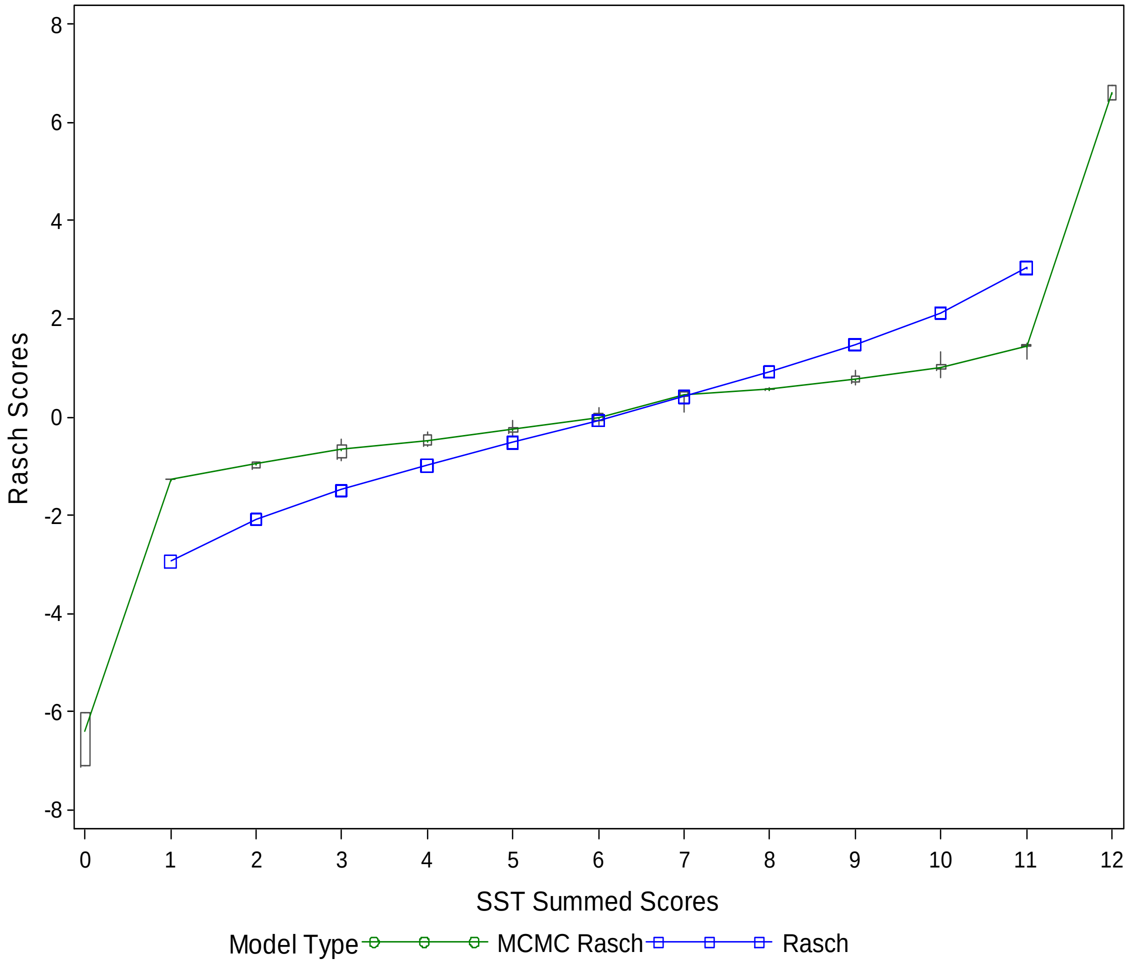

Figure 4.1: Rasch SST Scores and MCMC Rasch SST Scores

I obtained a meaningful estimate for Rasch scores for subjects whose SST summed scores were zero or twelve. For these subjects, Rasch scores were nonestimable in the classical Rasch model.

In order to obtain a fair comparison among the SST summed score, Rasch SST score, and MCMC Rasch SST score, I scaled SST Rasch scores and MCMC SST Rasch scores to match used the standard deviations of SST summed score. Figure 4.2 showed the rescaled Rasch SST scores and rescaled MCMC Rasch SST scores.

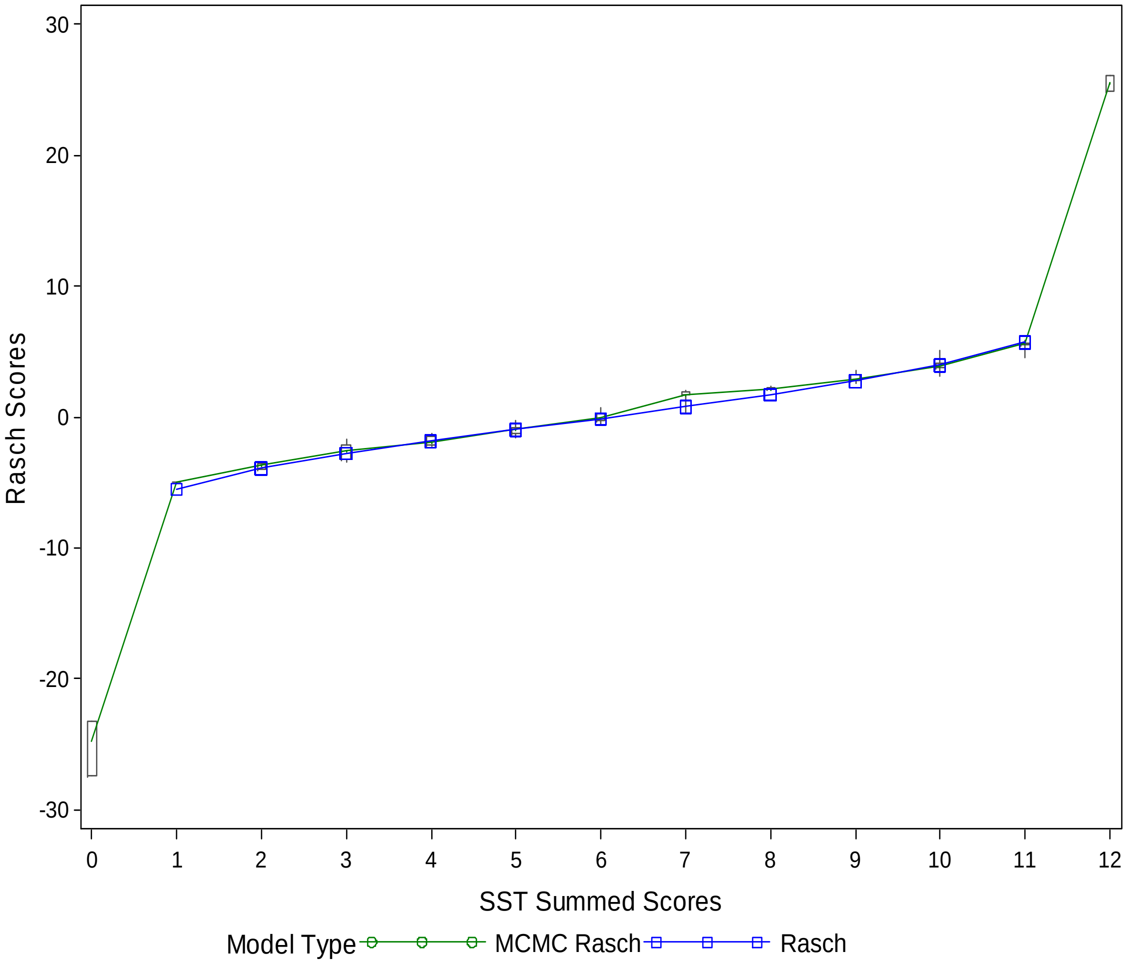

Figure 4.2: Rescaled Rasch SST Scores and Rescaled MCMC Rasch SST Scores

MCID in Rasch Scores Accounting for Measurement Errors Using the 15-item Anchored Question

Figure 4.3 plotted the change in two Rasch SST scores vs. SST summed score by patient group defined by 15-item anchored question. Several changes in the Minimal Improvement group for MCMC Rasch scores were much larger than the corresponding changes in Rasch scores. Table 4.3 presents the summary of SST summed score, Rescaled Rasch SST scores, and rescaled MCMC Rasch score and change from baseline (CFB) by anchored 15-item question. The estimated difference (95% CI) of the CFB between two groups is 1.95 (0.06, 3.85) for SST summed score; 1.97 (-0.17, 4.10) for

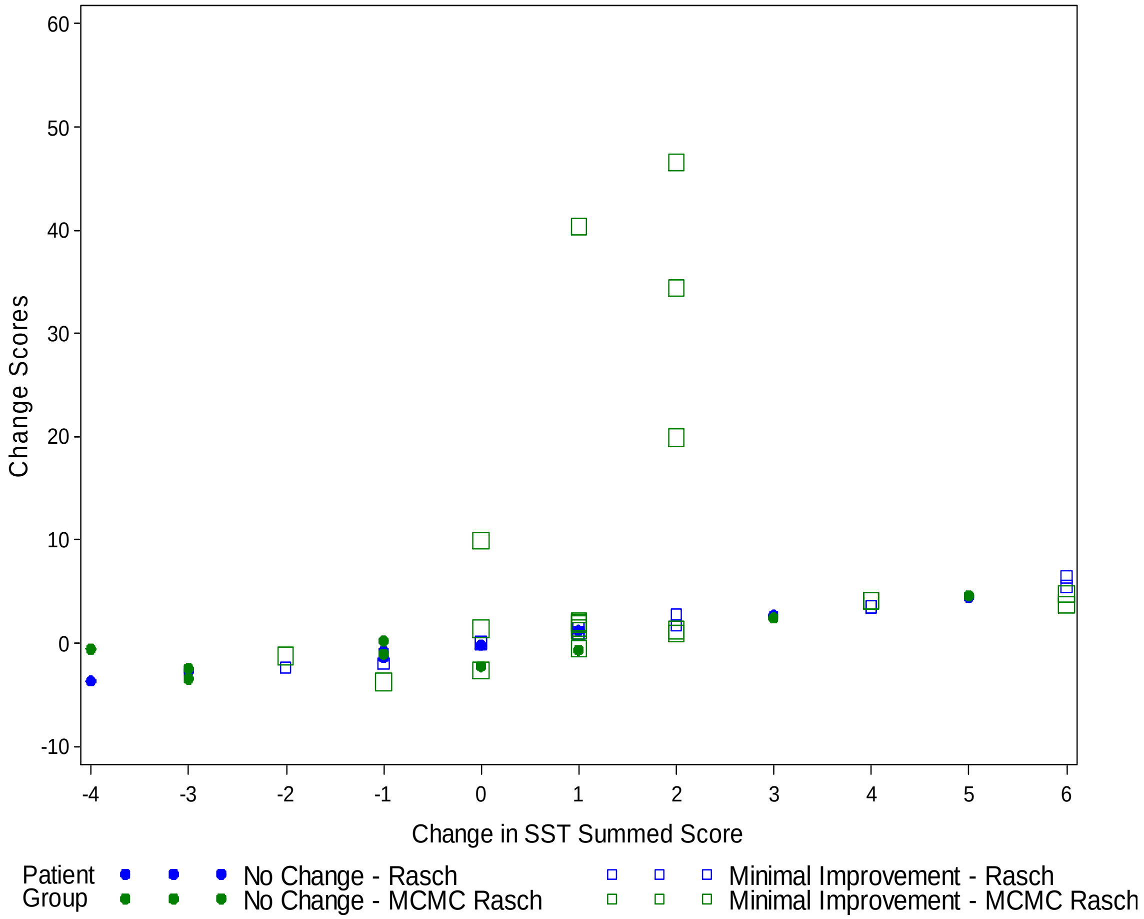

Figure 4.3: Change in Two Rescaled Rasch SST Scores vs. Change in STT Summed Score by Patient Groups Using 15-item Anchored Question

rescaled Rasch score; 8.54 (1.78, 15.30) for rescaled MCMC Rasch SST score. The p value is equal to 0.0156 for the difference of CFB in rescaled MCMC Rasch SST score between two groups, which is highly statistically significant. The Rasch model accounting for measurement errors seemed to distinguish the difference between two patient groups comparing with Rasch model and SST summed scores.

Table 4.3: Summary of SST summed score, Rescaled Rasch SST Score, and Rescaled MCMC SST Rasch Score by 15-Item Anchored Question

		No Change		Minimal Improvement	
Score Type	Visit	Mean (SD)	n	Mean (SD)	n
SST Summed Score	Week 0 (BL)	6.33 (3.000)	9	6.71 (3.690)	21
	Week 6	6.00 (2.500)	9	8.33 (3.055)	21
	CFB	-0.33 (2.958)	9	1.62 (2.012)	21
	Difference (CI)	1.95 (0.06, 3.85)			
	P Value	0.0439			
Rescaled Rasch SST Score	Week 0 (BL)	0.30 (2.856)	9	0.83 (3.422)	19
	Week 6	-0.03 (2.194)	9	1.45 (2.750)	17
	CFB	-0.32 (2.698)	9	1.64 (2.348)	16
	Difference (CI)	1.97 (-0.17, 4.10)			
	P Value	0.0693			
Rescaled MCMC Rasch SST Score	Week 0 (BL)	0.40 (2.888)	9	0.99 (8.441)	21
	Week 6	0.07 (2.421)	9	9.19 (17.052)	21
	CFB	-0.34 (2.521)	9	8.20 (14.464)	21
	Difference (CI)	8.54 (1.78, 15.30)			
	P Value	0.0156			

BL= Baseline; CFB= Change from Baseline; CI=95% Confidence Interval

MCID in Rasch Scores Accounting for Measurement Errors Using the Four-item Anchored Question

Similarly, Figure 4.4 plotted the change in two Rasch SST scores vs. SST summed score by patient group defined by four-item anchored question. Table 4.4 presents the scores and its change from baseline (CFB) by the anchored four-item question for SST summed score, Rescaled Rasch SST score, and rescaled MCMC Rasch score. The estimated difference (95% CI) of the CFB between two groups is 2.33 (0.99, 3.66) for SST summed score; 2.38 (1.03, 3.74) for rescaled Rasch score; and 1.20 (-7.40, 9.81) for rescaled MCMC Rasch score. The difference in CFB of Rescaled MCMC Rasch score between two groups is not statistically significant (p value=0.7810).

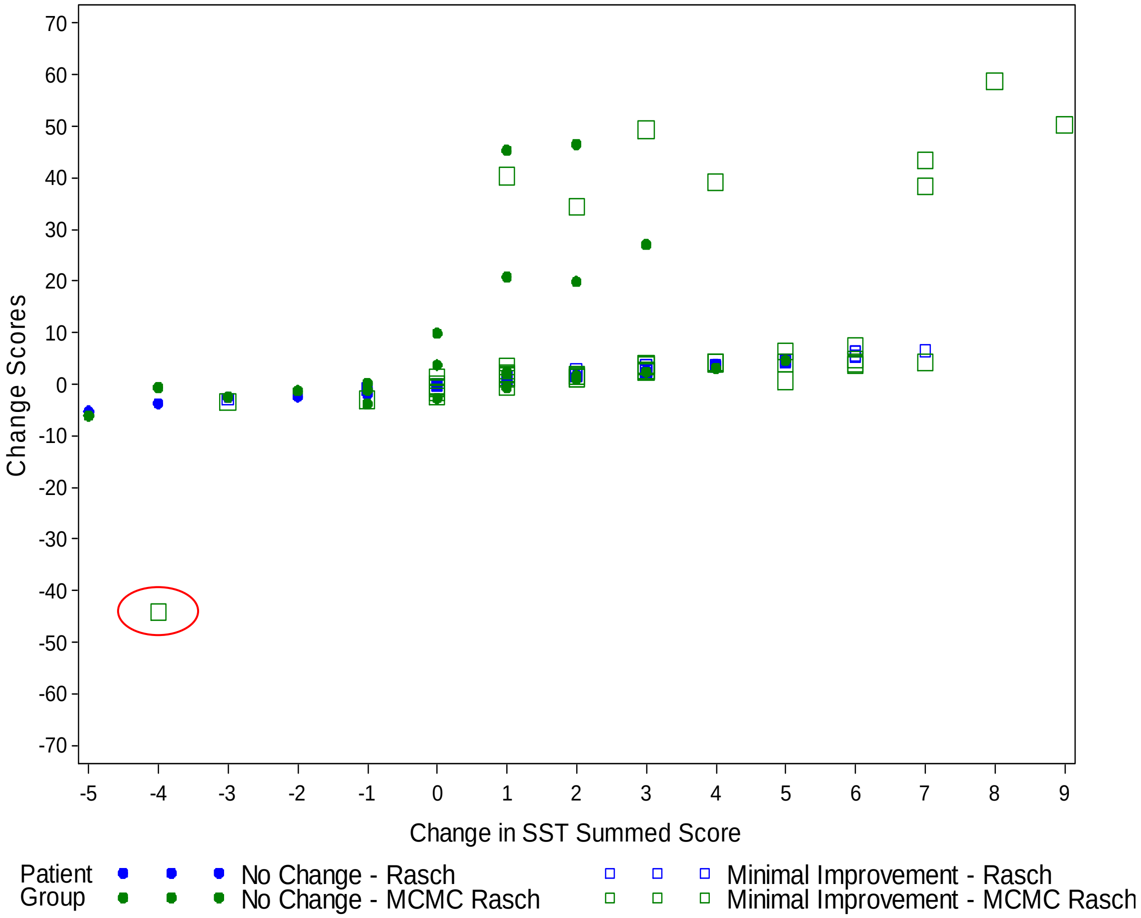

Figure 4.4: Change in Two Rescaled Rasch SST Scores vs. Change in SST Summed Score by Patient Groups Using Four-item Anchored Question

Table 4.4: Summary of SST summed score, Rescaled Rasch SST Score, and Rescaled MCMC SST Rasch Score by four-Item Anchored Question

		No Change		Minimal Improvement	
Score Type	Visit	Mean (SD)	n	Mean (SD)	N
SST Summed Score	Week 0 (BL)	5.83 (3.667)	24	5.76 (2.877)	46
	Week 6	6.33 (3.447)	24	8.59 (2.833)	46
	CFB	0.50 (2.396)	24	2.83 (2.783)	46
	Difference (CI)	2.33 (0.99, 3.66)			
	P Value	0.0009			
Rescaled Rasch SST Score	Week 0 (BL)	0.46 (2.972)	20	-0.21 (2.848)	46
	Week 6	-0.53 (2.704)	21	2.02 (2.333)	37
	CFB	0.06 (2.568)	18	2.44 (2.247)	37
	Difference (CI)	2.38 (1.03, 3.74)			
	P Value	0.0009			
Rescaled MCMC Rasch SST Score	Week 0 (BL)	-1.53 (10.845)	24	-0.14 (2.828)	46
	Week 6	5.75 (16.067)	24	8.35 (18.776)	46
	CFB	7.28 (14.380)	24	8.49 (18.368)	46
	Difference (CI)	1.20 (-7.40, 9.81)			
	P Value	0.7810			

BL=Baseline; CFB= Change from Baseline; CI= 95% Confidence Interval

According to the four-item anchored question, the Rasch model accounting for measurement errors provided different results when comparing with Rasch scores or SST summed scores. When closely inspecting the dataset, I found that one subject was able to perform four functions at Week 0, but could not perform any function at Week 6, as indicated by the point circled by red in Figure 4.4. However, this subject was classified in the Minimal Improvement group. I am wondering if the data recording was switched between Week 0 and Week 6. Nevertheless, after I excluded this subject from the analysis, I still did not obtain the statistically significant result for MCID analysis using the four-item anchored question. In other words, using MCMC Rasch SST scores, MCID cannot be determined by this four-item anchored question.

Discussion

From the preceding results, I observed the effect of measurement errors on the MCID analyses in the simple shoulder test (SST) for a population of patients with rotator cuff tendonitis or tearing. MCMC Rasch SST scores provided highly statistical significance for the difference of change from baseline (CFB) between patient groups classified by the 15-item anchored question, but the SST summed scores and Rasch score provided statistical borderline significance for the difference of CFB. Conversely, using the four-item anchored question, the difference of CFB between patient groups, was not statistically significant (p=0.718) based on MCMC Rasch SST scores, but the difference between patient groups were highly significant (both p values=0.0009) based on in SST summed scores and Rasch scores. The inconsistencies in MCIDs between MCMC Rasch scores and classical Rasch scores may be due to the bias of estimates of Rasch scores when measurement errors are left unconsidered. After accounting for measurement errors, I revealed highly significant difference of CFB in MCMC Rasch score while difference of CFB in Rasch score was almost hidden with borderline significance between patient groups by the 15-item anchored question; I could not claim significant difference of CFB in MCMC Rasch score while difference of CFB in Rasch score was determined with highly significance between patient groups by the four-item anchored question.

The first advantage of Rasch model accounting for measurement errors is that I obtained reasonable estimates of βs for those subjects with SST summed scores of 0 and 12, even though their Rasch SST scores remained infinity or nonestimable. Therefore, in determining MCID in MCMC Rasch sores, I had no missing data and the same sample size as the original analysis.[15]

Second, I suggest that the scoring system based on Rasch Model accounting for measurement errors is superior to the scoring system based on summed scores. As is discussed above, it does not assume equal increments for each additional function endorsed by a patient, and, as such, it may reflect the patient's actual experiences more accurately. Especially in two situations, the much larger magnitude of change in MCMC Rasch SST rescaled scores are observed: when responses go from no shoulder function to at least some shoulder function, and when responses go from no or some shoulder function to perfect function. For example, one subject's SST summed score is zero at Week 0 and one at Week 6; and the change score is one in terms of SST summed score, whereas the change was 20.8 in rescaled MCMC Rasch SST score. Another subject had the change from 10 to 12 in SST summed scores; but the change in Rescaled MCMC Rasch scores was 46.6. Therefore, Rasch model accounting for the measurement errors has an advantage over Rasch model in dealing with extreme SST summed scores. Moreover, this model overcomes the limitation of SST summed score restricted between 0 and 12, in which the average score was restricted between zero and one.

Still another advantage of Rasch model accounting for the measurement errors is that MCMC Rasch SST scores are now considered as "real" continuous measures. For example, if subjects endorse 7 functions on the SST, one subject's MCMC Rasch SST score was 2.5 while another subject's MCMC Rasch SST score was 1.8. This provides good justification to treat MCMC Rasch SST scores as continuous measures and to use the t-test, ANOVA, ANCOVA, regression, etc. to analyze the scores.

One limitation of this chapter is that I only assumed that measurement errors in the observed data were 50% of observed variance for $\boldsymbol{L_W}$ in this application, but I did not know with certainty if this was true for the observed data. Therefore, in the future, I could build the proportion of measurement errors as one of the parameters in the MCMC implementation to characterize it in a more accurate way. Another limitation is that findings on MCID results are based on a relatively small sample size (for No Change and Minimal Change groups, n=9 and n=21 respectively according to the 15-item anchored question; n=24 and n=46 respectively according to the four-item anchored question).

To conclude, measurement errors occur whenever there is measurement. In exploring this issue, I applied a Rasch model to account for measurement errors to data collected in a clinical setting. From this application, I observed the effects of measurement errors on the determination of MCID. Our major finding was that results obtained with the model I applied, i.e., Rasch Model accounting for measurement errors, are inconsistent with results from both the Rasch Model that does not account for measurement error and simple summed score model. This research provides a framework to explore measurement errors in real life situations. Even though the Rasch model accounting for measurement errors only pertain to questionnaires with binary response, this research may be expanded to include questionnaires with multiple response categories, which increasingly are used in clinical and health care research. Additionally, I found that the implementation of the model is feasible. The convergence of chains required 10,000 iterations, but this took only about 45 minutes of computer time on a lap top computer.

Appendices

Appendix A: Likelihood Functions

In the Rasch model[13] for dichotomous data, the probability of outcome $X_{ni} = 1$ is given by

$$P(X_{ni} = 1) = \frac{exp(\beta_n - \delta_i)}{1 + exp(\beta_n - \delta_i)}$$

where

X_{ni} is the response (1=Yes, 0= No) for Person n to Item i;

β_n is the ability of person n (*n = 1, …, N);*

δ_i is the difficulty of Item i (*i = 1, …, I);*

P(·) is the probability that Person n has a true or observed response to Item i.

If measurement errors are not considered in the Rasch model, the likelihood function based on the observed data is

$$f(\boldsymbol{X}|\boldsymbol{\beta},\boldsymbol{\delta}) = LH(\boldsymbol{\beta},\boldsymbol{\delta}) = \prod_{n,i} P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta})^{X_{ni}} \times \left(1 - P(X_{ni}|\boldsymbol{\beta},\boldsymbol{\delta})\right)^{1-X_{ni}}$$

where X_{ni} is the response by n^th^ subject for i^th^ question.

The Rasch model accounting for the classical measurement error structure may be modeled in the following way: the logit transformed latent variables $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ linked the true response $\boldsymbol{X}$ and the observed response $\boldsymbol{W}.$

$$\boldsymbol{L_W} = \boldsymbol{L_X} + \boldsymbol{U}$$

$$L_{X_{ni}} = \beta_n - \delta_i$$

where each component U_{ni} of $\mathbf{U}$ are i.i.d. random variables of measurement error

$\boldsymbol{U} \sim N(0\mathbf{1}, \sigma_u^2 \boldsymbol{I})$, where σ_u^2 is the variance of the classical measurement error.

$\boldsymbol{L_X} \sim N(\boldsymbol{\mu_{L_X}}, \sigma_{\boldsymbol{L_X}}^2 \boldsymbol{I})$, where $\boldsymbol{\mu_{L_X}}$ is the mean of $\boldsymbol{L_X}$, and $\sigma_{\boldsymbol{L_X}}^2$ is the variance of $\boldsymbol{L_X}$, the joint distribution of $\boldsymbol{L_W}$ and $\boldsymbol{L_X}$ is:

$$\begin{pmatrix} L_X \\ L_W \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{L_X} \\ \mu_{L_X} \end{pmatrix}, \begin{bmatrix} \sigma_{L_X}^2 \boldsymbol{I} & \sigma_{L_X}^2 \boldsymbol{I} \\ \sigma_{L_X}^2 \boldsymbol{I} & \left(\sigma_{L_X}^2 + \sigma_u^2\right)\boldsymbol{I} \end{bmatrix}\right)$$

Then, the conditional distribution of $\boldsymbol{L_X}$ given $\boldsymbol{L_W}$ is:

$$\boldsymbol{L_X}|\boldsymbol{L_W} \sim \boldsymbol{N}\left(\boldsymbol{\mu_{L_X}} + \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2 + \sigma_u^2}\boldsymbol{I}\left(\boldsymbol{L_W} - \boldsymbol{\mu_{L_X}}\right),\ \sigma_{L_X}^2\left(1 - \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2 + \sigma_u^2}\right)\boldsymbol{I}\right)$$

The complete likelihood function based on the observed data is

$$LH(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{L_X}) = \{f(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\delta})\} \times \{f(\boldsymbol{\beta}) * f(\boldsymbol{\delta})\} \times \{f(\boldsymbol{X}|\boldsymbol{L_X})\} \times \{f(\boldsymbol{L_X}|\boldsymbol{L_W})\}$$

That is $\{Rasch\ Model\} \times \{Prior\} \times \{\text{Bernoulli}\} \times \{Conditional\ \}$

where

$$f(\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\delta}) = LH(\boldsymbol{\beta}, \boldsymbol{\delta}) = \prod_{n,i} P(X_{ni}|\boldsymbol{\beta}, \boldsymbol{\delta})^{X_{ni}}(1 - P(X_{ni}|\boldsymbol{\beta}, \boldsymbol{\delta}))^{1-X_{ni}}$$

$$f(\boldsymbol{\beta}) \propto exp\left(\boldsymbol{\beta}^T \textstyle\sum_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta}/2\right)$$

$$f(\boldsymbol{\delta}) \propto exp(\boldsymbol{\delta}^T \textstyle\sum_{\boldsymbol{\delta}}^{-1} \boldsymbol{\delta}/2)$$

$$f(\boldsymbol{X}|\boldsymbol{L_X}) = \prod_{ni} P\left(X_{ni}|L_{X_{ni}}\right)^{X_{ni}} \times \left(1 - P\left(X_{ni}|L_{X_{ni}}\right)\right)^{1-X_{ni}}$$

$$\boldsymbol{L_X}|\boldsymbol{L_W} \sim N\left(\boldsymbol{\mu_{L_X}1} + \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2 + \sigma_u^2}\boldsymbol{I}\left(\boldsymbol{L_W} - \boldsymbol{\mu_{L_X}1}\right), \sigma_{L_X}^2\left(1 - \frac{\sigma_{L_X}^2}{\sigma_{L_X}^2 + \sigma_u^2}\right)\boldsymbol{I}\right)$$

In order for the implementation of the above formula, we can assume that $\boldsymbol{\mu_{L_X}}$ is equal to $\boldsymbol{L_W}$, and $\sigma_{L_x}^2$ is equal to $\sigma_{\boldsymbol{L_W}}^2$ estimated as $var\ (\boldsymbol{L_W})$, and $\sigma_{\boldsymbol{U}}^2$ is equal to proportion (denoted as a%) of $\sigma_{\boldsymbol{L_W}}^2$. Therefore, $\boldsymbol{L_X}|\boldsymbol{L_W} \sim N\left(\boldsymbol{L_W1}, \sigma_{\boldsymbol{L_W}}^2\left(\frac{0.a}{1.a}\right)\boldsymbol{I}\right)$

Appendix B: Gibbs Samplers and Metropolis-Hastings Algorithm

The Gibbs samplers along with the Metropolis-Hastings algorithm were used to generate estimates of the parameters from the complete likelihood function. For each Gibbs sampler[18] generated at following steps, the general form of Metropolis-Hastings Algorithm.[17, 22] to accept a Gibbs Sampler θ from a proposed distribution q(.) is:

$$\alpha = min\left\{1, \frac{LH\ (\theta_{candidate})q(\theta_{current}|\theta_{candidate})}{LH\ (\theta_{current})q(\theta_{candidate}|\theta_{current})}\right\}$$

Then compare α with probability generated from uniform distribution. If α is greater than the probability from uniform distribution, then update $\theta_{current}$ with $\theta_{candidate}$. Otherwise, keep $\theta_{current}$ as it is.

The steps to generate Gibbs Samplers and associated Metropolis-Hastings Algorithms are:

1. Obtain initial starting values of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, e.g. estimates the two parameters from the simulated dataset using Rasch model ignoring measurement error. Set these values as the current values $\boldsymbol{\beta}_{current}$ and $\boldsymbol{\delta}_{current}$.
2. Generate a candidate of the true Logit $\boldsymbol{L_X}$ according to the distribution of $\boldsymbol{L_X}|\boldsymbol{L_W}$, given the observed $\boldsymbol{L_W}$ equal to $\boldsymbol{\beta}_{current}$ minus $\boldsymbol{\delta}_{current}$.

The acceptance ratio for each element of candidate L_X[j], where j= 1, 2, …, $N \times I$, is

$$\frac{LH\ (L_{Xcandidate}[j])q(L_{Xcurrent}[j]|L_{Xcandidate}[j])}{LH\ (L_{Xcurrent}[j])q(L_{Xcandidate}[j]|L_{Xcurrent}[j])} = \frac{f(X|L_{Xcandidate}[j])}{f(X|L_{Xcurrent}[j])}$$

$$= \frac{(P_{Xcandidate}[j])^{X_{current}} \times (1 - P_{Xcandidate}[j])^{(1-X_{current})}}{(P_{Xcurrent}[j])^{X_{current}} \times (1 - P_{Xcurrent}[j])^{(1-X_{current})}}$$

where

$$P_{Xcandidate}[j] = P(X_{candidate}[j] = 1) = \frac{exp(L_{Xcandidate}\,[j])}{1 + exp(L_{Xcandidate}\,[j])}$$

$$P_{Xcurrent}[j] = P(X_{current}[j] = 1) = \frac{exp(L_{Xcurrent}\,[j])}{1 + exp(L_{Xcurrent}\,[j])}$$

3. Generate a candidate of true response $\boldsymbol{X}_{candidate}$ based on updated $\boldsymbol{L}_{\boldsymbol{X}current}$. $\boldsymbol{X}_{candidate}$ follows logistic distribution. $\boldsymbol{X}_{candidate}$ follows Logistic distribution. The acceptance ratio for each element of $X_{candidate}[j]$, where j= 1, 2, …, $N \times I$, is

$$\frac{LH\ (X_{candidate}[j])q(X_{current}[j]|X_{candidate}[j])}{LH\ (X_{current}[j])q(X_{candidate}[j]|X_{current}[j])}$$
$$= \left(exp(\beta_{current}[n] - \delta_{current}[i])\right)^{X_{candidate}[j]-X_{current}[j]}$$

where $j = (n-1) \times 12 + i,$

$$for\ n^{th} subject\ and\ i^{th} item, denoted\ as\ "ni"\ sometimes$$

4. Generate candidate of $\boldsymbol{\beta}$ based on current $\boldsymbol{\beta}_{\text{current}}$ using the normal distribution $N\left(\boldsymbol{\beta}_{current}, \sigma^2_{\boldsymbol{\beta}\,proposed}\right)$. The $\sigma^2_{\boldsymbol{\beta}\,proposed}$ should be small. The prior distribution of $\boldsymbol{\beta}$ is $N\left(0\mathbf{1}, \sigma^2_{\boldsymbol{\beta}}\boldsymbol{I}\right)$.

The acceptance ratio for each element of $\beta_{candidate}[n]$, where n= 1, 2, …, $Nth\ subjects$, is

$$\frac{LH\ (\beta_{candidate}[n])q(\beta_{current}[n]|\beta_{candidate}[n])}{LH\ (\beta_{current}[n])q(\beta_{candidate}[n]|\beta_{current}[n])} =$$

$$\prod_{i=1}^{I=12}\left(\frac{exp\{\beta_{candidate}[n]-\delta_{current}[i]\}}{exp\{\beta_{current}[n]-\delta_{current}[i]\}}\right)^{X_{current}[ni]}$$

$$\times\prod_{i=1}^{I=12}\left(\frac{1+exp\{\beta_{current}[n]-\delta_{current}[i]\}}{1+exp\{\beta_{candidate}[n]-\delta_{current}[i]\}}\right)$$

$$\times\, exp\left\{-\frac{(\beta_{candidate}[n])^2-(\beta_{current}[n])^2}{2\times\sigma^2_{\boldsymbol{\beta}}}\right\}$$

5. Generate candidate of $\boldsymbol{\delta}$ based on current $\boldsymbol{\delta}_{current}$ using the normal distribution $N\left(\boldsymbol{\delta}_{current}, \sigma^2_{\boldsymbol{\delta}\,proposed}\right)$. $\sigma^2_{\boldsymbol{\delta}\,proposed}$ should be small. The prior distribution of $\boldsymbol{\delta}$ is $N(0\mathbf{1}, \sigma^2_{\boldsymbol{\delta}}\boldsymbol{I})$.

The acceptance ratio for each element of $\delta_{candidate}[i]$, where $\mathrm{i} = 1, 2, \ldots, Ith\ Item,$ is:

$$\frac{LH\ (\delta_{candidate}[i])q(\delta_{current}[i]|\delta_{candidate}[i])}{LH\ (\delta_{current}[i])q(\delta_{candidate}[i]|\delta_{current}[i])} =$$

$$\prod_{n=1}^{N}\left(\frac{exp\{\beta_{current}[n]-\delta_{candidate}[i]\}}{exp\{\beta_{current}[n]-\delta_{current}[i]\}}\right)^{X_{current}[ni]}$$

$$\times\prod_{n=1}^{N}\left(\frac{1+exp\{\beta_{current}[n]-\delta_{current}[i]\}}{1+exp\{\beta_{current}[n]-\delta_{candidate}[i]\}}\right)$$

$$\times\, exp\left\{-\frac{(\delta_{candidate}[i])^2-(\delta_{current}[i])^2}{2\times\sigma^2_{\boldsymbol{\delta}}}\right\}$$

Repeat Step 2 to 5 for a large number of times, or until convergence.

From the generation of Gibbs samplers, the chain for Logit $\mathbf{L_X}$ is an independent chain. For an independent chain where proposal transitional kernel q(y|x=q(y)it may seem that the independence from the previous state disagrees with Markovian property of the chain. Actually, q is just a proposal that is combined with an acceptance probability of α to give the transitional p of the algorithm. This transition depends on the previous state, thus, preserves the Markovian properties.[22] The chains for β and δ are symmetric chains, in this case when $q(y|x) = q(|y - x|)$, e.g., the normal distribution with mean x, then we have $q(y|x) = q(x|y)$.

Appendix C: Two Anchored Questions

15-Item Anchored Question

Since your last clinic visit, has there been any change in the function of your treated shoulder?

1. A very great deal worse
2. A great deal worse
3. A good deal worse
4. Moderately worse
5. Somewhat worse
6. A little worse
7. Almost the same, hardly any worse at all
8. No change
9. Almost the same, hardly any better at all
10. A little better
11. Somewhat better
12. Moderately better
13. A good deal better
14. A great deal better
15. A very great deal better

Four-Item Anchored Question

Since your last clinic visit, please rate your response to treatment.

1. None – no good at all, ineffective treatment
2. Poor – some effect but unsatisfactory
3. Good – satisfactory effect with occasional episodes of pain or stiffness
4. Excellent – ideal response, virtually pain free

According to the 15-item anchored question, no change group [15] included patients whose answers were: almost the same, hardly any worse at all; “No change”; and “Almost the same, hardly any better at all”. The minimal improvement group included patients if their answers to this question were: “A little better” and “Somewhat better”.

According to the four-item anchored question, the patients were classified as no change group [15] if their answers to this question were: “None” and “Poor”. The patients were classified as minimal improvement group [15] if their answers to this question were “Good”.

References

1. Carroll RJ, Ruppert D, Stefanski LA. Chapter 2 Important Concept. *Measurement Error in Nonlinear Models A Modern Perspective* (Second edn). Chapman & Hall/CRC, 1995.

2. Sheng X, Carrière KC. An improved CML estimation procedure for the Rasch model with item response data. *Statistics in Medicine* 2002; **21**: 407-416.

3. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research* 2007; **57**: 1358-1362.

4. Sheng X. A Bayesian model for measurement errors in diagnosis of rheumatoid arthritis. *Communications in Statistics - Theory and Methods* 2009; **38**: 3419 - 3431.

5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989; **10**: 407-415.

6. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC, 2004.

7. Carroll RJ, Ruppert D, Stefanski LA. Chapter 3 Linear Regresstion and Attenuation. *Measurement Error in Nonlinear Models A Modern Perspective* (Second edn). Chapman & Hall/CRC, 1995.

8. May K, Nicewander WA. Measuring change conventionally and adaptively. *Educational and Psychological Measurement* 1998; **58**: 882-897.

9. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-40): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997; **50**: 451-461.

10. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998; **51**: 1203-1214.

11. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Medical care* 2004; **42**: I25-I36.

12. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, Gregg PJ. A comparison of Rasch with Likert scoring to discriminate between

patients' evaluations of total hip replacement surgery. *Quality of Life Research* 2004; **13**: 331-338.

13. Rasch G. *Chapter VII Notions Implied in the Structural Model for Items, Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press: Chicago, 1980.

14. University of Washington. Simple Shoulder Test, Available at http://www.orthop.washington.edu/uw/simpleshoulder/tabID__3376/ItemID__186/PageID__356/qview__true/Articles/Default.aspx, Accessed Jan 7, 2012

15. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010; **92**: 296-303.

16. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009; **18**: 927-932.

17. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**: 97-109.

18. Gamerman D, Lopes HF. Chapter 5 Gibbs Sampling. *Markov Chain Monte Carlo.* (2nd edn). Chapman &Hall/CRC, 2006.

19. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* (2nd edn). Chapman & Hall/CRC, 2004.

20. Copay AG, Subach BR, Glassman SD, Polly Jr DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal* 2007; **7**: 541-546.

21. Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* 1997; **7**: 110-120.

22. Gamerman D, Lopes HF. Chapter 6 Metropolis-Hastings algorithm. *Markov Chain Monte Carlo.* (2nd edn). Chapman &Hall/CRC, 2006.

23. Geyer CJ, Thompson EA. Annealing Markov Chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* 1995; **90**: 909-920.

24. Gilks WR, Richardson S, Spiegelhalter DJ. Chapter 1 Introducing Markov chain Monte Carlo. *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC, 1996.

CHAPTER 5

CONCLUSION

In order to address the nonlinearity of the summed score[1-7] and account for the effects of measurement errors[8] in patients' responses to questionnaires, I proposed to incorporate measurement errors into the Rasch model under a preestablished framework of measurement errors.[9] Due to the difficulty in modeling the likelihood of the Rasch model along with measurement errors directly, I adopted the Markov Chain Monte Carlo (MCMC) approach to find the estimates of Rasch model parameters.[10-14] I referred to the Rasch model accounting for measurement errors as the MCMC Rasch model. The estimates of MCMC Rasch model parameters ($\boldsymbol{\beta}$). i.e., person abilities are called MCMC Rasch scores.

Through simulations (Chapter 3), I compared the results from classical Rasch model with the results from MCMC Rasch model with different settings that affect the MCMC implementation. I discovered the optimal setting of these factors. In addition, I observed three patterns:

- In a situation where a true extreme β remained an extreme β after measurement errors were introduced, the mean square of error (MSE) for classical Rasch model is much larger than that for MCMC Rasch models (MSEs are tens-of-thousands for classical Rasch model vs. MSEs are around 300 for MCMC Rasch models).

- In the situation where a true extreme β became nonextreme β after measurement errors were introduced in the datasets, MCMC Rasch models produced a smaller MSE.
- In the case where a true nonextreme β remained a nonextreme β after measurement errors were introduced in the datasets, MCMC Rasch models performed better for some βs while classical Rasch model performed better for other βs. However, the differences between the best MCMC Rasch model (i.e., MCMC II) and classical Rasch for nonextreme βs are small.

Generally speaking, then, MCMC Rasch models are better than classical Rasch models when measurement errors exist in datasets.

Most importantly, the MCMC Rasch model explored here provided a way to obtain the estimates corresponding to extreme summed scores with reasonable variances, which remain inestimable in the classical Rasch model. Therefore, Rasch model accounting for the measurement errors has an advantage over Rasch model in dealing with extreme SST summed scores. Furthermore, I found that results appear to be unaffected by the magnitude of measurement errors assumed in the dataset when MCMC is implemented using mean square of errors (MSE). Comparing the Rasch Model and Rasch Model accounting for measurement errors, the latter produced better estimates of Rasch model parameters (see Chapter 3).

With the application of our proposed MCMC Rasch model to a Simple Shoulder Test (SST) dataset[15] for patients with rotator cuff tendonitis or tearing, I evaluated the effect of measurement errors on the determination of the minimum clinically important difference (MCID). Again, the aim of MCID analysis is to find the statistically significant difference in change from baseline (CFB) between a No Change group and a Minimal Improvement group via two anchored questions.[16, 17] In order to facilitate the evaluation of the effect of measurement errors in this application, I also performed the MCID analyses of Rasch SST scores (Chapter 2) and MCMC Rasch SST scores in addition to the MCID determined in the SST summed score in the original study.[18]

According to the 15-item anchored question, the difference (95% CI; p value) of the CFB between two groups was 1.95 (0.06, 3.85; 0.0439) for SST summed score; 1.97 (-0.17, 4.10; 0.0693) for rescaled Rasch score; and 8.54 (1.78, 15.30; 0.0156) for rescaled MCMC Rasch SST score. According to the four-item anchored question, the difference (95% CI; p value) of the CFB between two groups was 2.33 (0.99, 3.66; 0.0009) for SST summed score; 2.38 (1.03, 3.74; 0.0009) for rescaled Rasch score; and 1.20 (-7.40, 9.81; 0.7810) for rescaled MCMC Rasch score.

As shown above, in Rasch scores, the result from MCID analysis according to the 15-item anchored question is consistent with the result according to the four-item anchored question. Furthermore, MCID as assessed through these anchors is consistent with MCID assessed by summed scores.

Nevertheless, using MCMC Rasch SST scores, the MCID cannot be ascertained by two anchored questions. In short, this finding is inconsistent with the result of MCID analysis through Rasch scores, and it may be due to the bias of estimates of Rasch scores when measurement errors are left unconsidered.

From simulations in Chapter 3, I found evidence that Rasch model accounting for measurement errors provided the estimates for both person abilities and item difficulties from questionnaires closer to the “true” model parameters in terms of MSE.

Through its application in Chapter 4, I confirmed more advantages of Rasch model accounting for measurement errors. These are as follows:

1. With reasonable estimates of βs for those subjects with SST summed scores of 0 and 12, I had no missing data and the same sample size as the original analysis in determining MCID of MCMC Rasch scores.[18]
2. I suggest that the scoring system based on Rasch Model accounting for measurement errors is superior to the scoring system based on summed scores. As is discussed above, it does not assume equal increments for each additional function endorsed by a patient, and, as such, it may reflect the patient's actual experiences more accurately. Especially in two situations, the much larger magnitude of change in MCMC Rasch SST rescaled scores are observed: when responses go from no shoulder function to at least some shoulder function, and when responses go from no or some shoulder function to perfect function. For example, one subject's SST summed score is zero at Week 0 and one at Week 6; and the change score is one in terms of SST summed score, whereas the change was 20.8 when using the rescaled MCMC Rasch SST score. Another subject had the change from 10 to 12 in SST summed scores; but the change in Rescaled MCMC Rasch scores was 46.6. Therefore, Rasch model accounting for the measurement errors has an advantage over Rasch model in dealing with extreme SST summed scores. Therefore, this model overcomes the limitation of SST summed score restricted between 0 and 12, in which the average score was restricted between zero and one.

3. MCMC Rasch SST scores may be considered as “real” continuous measures. For example, if subjects endorse 7 functions on the SST, one subject’s MCMC Rasch SST score was 2.5 while another subject’s MCMC Rasch SST score was 1.8. This provides good justification to treat MCMC Rasch SST scores as continuous measures and to use the t test, ANOVA, ANCOVA, regression, etc. to analyze the scores.

The Rasch model accounting for measurement errors as explored in this dissertation has far-reaching implications, such as:

- In health outcomes research, the estimated person abilities obtained from Rasch model accounting for measurement errors can be analyzed in the general regression model to describe the “true” relationship between the outcome measures and interventions. Evidence-based outcomes increasingly are used to formulate treatment guidelines that inform public health policy. For example, what are the outcomes of conservative treatment for rotator cuff tendonitis or tearing? Of course, the “true” relationship between an intervention and its outcomes may be useful in formulating such policy decisions.
- In clinical trials, the estimated person abilities obtained from Rasch model accounting for measurement errors may be compared between treatments to reach more accurate conclusions about the treatment effect. The question, in other words, is the extent to which patients actually benefit from an intervention. On the other hand, the model proposed here may find that an intervention may be ineffective (or worse), which would suggest that either more research on its

outcomes is necessary, or, alternatively, it should be classified as an intervention with questionable outcomes.

- The more accurate MCID may be determined using estimated person abilities from Rasch model with measurement errors. For less risky procedures and drugs that rarely result in death, the MCID may be used as a metric to establish a “volume outcome relationship”. For example, the more an institution performs a shoulder nonoperative intervention (volume), the higher may be its rate of patients that reach the MCID of the SST (outcome). If such a volume-outcome relationship is ascertainable, patients may then be channeled toward high rate, better outcome institutions. Regarding clinical trials, the MCID is required to calculate sample sizes. If a trial is overpowered (sample size too large), then expenses for it may accrue and the drug or procedure it tests may be delayed in its development (or the trial may be deemed too expensive to conduct in the first place, and the drug or procedure may remain undeveloped). On the other hand, if a trial is underpowered (sample size to small), then it may result in a false negative; in other words, the drug or procedure may be effective, but too few patients were included in the trial to ascertain statistical significance.

In addition to the above contributions to outcomes research, this dissertation investigated Rasch modeling with measurement errors in a way that would be more generally of methodological value, because the framework for incorporation of measurement errors in Rasch model may be extended to the situations of multiple response questionnaires. In questionnaires with multiple response categories (such as Beck Depression Inventory[19]), the polytomous version of the Rasch model can be used to correct the linearity problem of total scores. At the same time, we may account for measurement errors in the logit of polytomous version of the Rasch model for multiple response questionnaires. Better estimates of parameters may also be obtained through the Bayesian method for additional statistical analysis.

Finally, I found that the implementation of the model is feasible in the application of real world datasets. The convergence of chains required 10,000 iterations, but this took only about 45 minutes of computer time on a laptop computer.

References

1. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *J Clin Epidemiol* 1996; **49**: 711-717.

2. McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-40): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol* 1997; **50**: 451-461.

3. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998; **51**: 1203-1214.

4. Norquist JM, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs. raw scores in measuring change in health. *Medical Care* 2004; **42**: I25-I36.

5. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW, Gregg PJ. A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Quality of Life Research* 2004; **13**: 331-338.

6. White LJ, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. *Archives of Physical Medicine and Rehabilitation* 2002; **83**: 822-831.

7. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research* 2007; **57**: 1358-1362.

8. Carroll RJ, Ruppert D, Stefanski LA. Chapter 2 Important Concept. *Measurement Error in Nonlinear Models A Modern Perspective* (Second edn). Chapman & Hall/CRC, 1995.

9. Sheng X. A Bayesian model for measurement errors in diagnosis of rheumatoid arthritis. *Communications in Statistics - Theory and Methods* 2009; **38**: 3419 - 3431.

10. Gamerman D, Lopes HF. Chapter 4 Markov Chains. *Markov Chain Monte Carlo.* (2nd edn). Chapman & Hall/CRC, 2006.

11. Gamerman D, Lopes HF. Chapter 6 Metropolis-Hastings algorithm. *Markov Chain Monte Carlo.* (2nd edn). Chapman & Hall/CRC, 2006.

12. Gamerman D, Lopes HF. Chapter 5 Gibbs Sampling. *Markov Chain Monte Carlo.* (2nd edn). Chapman & Hall/CRC, 2006.

13. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* (2nd edn). Chapman & Hall/CRC, 2004.

14. Geyer CJ. Practical Markov Chain Monte Carlo. *Statistical Science* 1992; **7**: 473-483.

15. University of Washington. Simple Shoulder Test, Available at http://www.orthop.washington.edu/uw/simpleshoulder/tabID__3376/ItemID__186/PageID__356/qview__true/Articles/Default.aspx, Accessed Jan 7, 2012

16. Copay AG, Subach BR, Glassman SD, Polly Jr DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal* 2007; **7**: 541-546.

17. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989; **10**: 407-415.

18. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010; **92**: 296-303.

19. Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review* 1988; **8**: 77-100.