THE FECAL INCONTINENCE QUALITY OF LIFE SCALE (FIQL): IMPROVING OUTCOMES MEASUREMENT

by

Alexander Peterson

B.A., Reed College, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Population and Public Health)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2017

Abstract

The Fecal Incontinence Quality of Life scale (FIQL) is a patient reported outcome measure (PROM) that is used to measure the effect that fecal incontinence has on quality of life, and has previously demonstrated high reliability and validity. It measures four domains of quality of life: lifestyle, coping/behavior, depression/self-perception, and embarrassment. Despite its wide use, previous studies have not applied rigorous modern methods to evaluate the FIQL's psychometric properties at the item and test level.

This thesis used a cohort of prospectively recruited patients from an elective surgical registry and applied methods from classical test theory (CTT), exploratory factor analysis (EFA), item response theory (IRT), and differential item and test functioning (DIF) to identify strengths and weaknesses in the FIQL. Specifically, this thesis aims to 1) confirm the reliability of the instrument, 2) describe the domains of quality of life measured by the instrument, 3) identify high and low quality items, and 4) determine whether one's score on the FIQL is influenced by gender or surgical procedure.

Out of 317 completed questionnaires from 880 total eligible patients, 236 were included for analysis. Reliability for all four domains was high as measured by Cronbach's α. Exploratory factor analysis failed to identify the four domains the FIQL claims to measure. Individual items demonstrated high discrimination but most had low difficulty. Items 2c, 2l, 3a, and 3h failed to demonstrate good separation between response categories. Five item pairs demonstrated local item dependence, most from question 3. Only item 2g demonstrated differential item functioning, based on gender. Differential test functioning was minimal.

The FIQL demonstrated a high degree of reliability, and the lifestyle domain can be used as is or with minor improvements. The FIQL can be improved by making response options

consistent, distributing items from different domains evenly throughout the instrument, adding items with higher difficulty and better response separation, and removing items 2c, 2l, 3a, and 3h. Further research is needed before the FIQL can be used confidently as a stand-alone measure of fecal incontinence-related quality of life.

Preface

This thesis is based on analysis of patient-reported outcomes data collected by a grant funded by the Canadian Institutes for Health Research (CIHR) whose principal investigator was Dr. Jason M. Sutherland, Centre for Health Service and Policy Research, School of Population and Public Health, University of British Columbia. I was solely responsible for data preparation, statistical analysis, interpretation, and writing the manuscript. The research was approved by the University of British Columbia's Behavioural Research Ethics Board (BREB; H12-02062).

Table of Contents

List of Tables

List of Figures

List of Equations

List of Abbreviations

CFI:	Comparative fit index
CTT:	Classical test theory
DIF:	Differential item functioning
DTF:	Differential test functioning
EFA:	Exploratory factor analysis
FI:	Fecal incontinence
FIQL:	Fecal Incontinence Quality of Life Scale
GRM:	Graded response model
ICC:	Item characteristic curve
IRT:	Item response theory
MPT:	Mean P-value threshold (anchor selection method)
PCA:	Principal components analysis
PROM:	Patient reported outcome measure
RMSEA:	Root mean square error of approximation
sDTF:	Signed differential test functioning statistic
SRMR:	Standardized root mean square residual
TLI:	Tucker Lewis index
uDTF:	Unsigned differential test functioning statistic
VCH:	Vancouver Coastal Health

Acknowledgements

I would first like to thank my adviser, Dr. Jason M. Sutherland, whose support and guidance made this thesis possible.

I would also like to express my gratitude to my thesis committee, Drs. Charlyn Black, Ahmer Karimuddin, and Mieke Koehoorn, for their valuable feedback and insight.

Lastly, I would like to thank my partner, Evin, my parents, and my family and friends who have provided encouragement and unconditional support throughout the process of preparing for and writing this thesis. This accomplishment would not have been possible without them.

Dedication

To Evin

Chapter 1: Introduction

Fecal incontinence (FI), or the accidental leakage of gas or stool, is a burdensome condition that significantly impacts quality of life. Patients have reported that FI affects psychosocial functioning, work and travel, and patients may face social stigma, embarrassment and depression (1–4). It is also a relatively common condition, whose prevalence in the United States may range as high as 7% to 15% (5). In Canada, studies have found prevalence for FI of 3.7% in a community-based sample (6), 3% among secondary school teenage girls (7), 4% among older adults (8), and 7.7% among post-partum women (9).

Since FI is prevalent and has a large impact on quality of life, it is important to develop instruments that can accurately quantify this impact and lead to better treatments and interventions for FI.

This thesis evaluates an existing instrument, known as the Fecal Incontinence Quality of Life scale (FIQL) (2), that was developed to measure FI-related quality of life. To do so, a cohort of patients with FI was created from a sample of prospectively recruited patients waiting for elective surgical treatment. The cohort is composed of patients waiting for surgery for conditions associated with fecal incontinence, and who have completed the FIQL.

In addition to classical test theory (CTT), modern validation methodologies such as item-response theory (IRT) and exploratory factor analysis (EFA) will be used to identify the strengths and weaknesses in the instrument.

The findings from this thesis allow researchers to use the FIQL more appropriately. The strengths of the FIQL identified from this thesis should generate more confidence among physicians and researchers to assess the impact of FI and treatment for FI on patient quality of

life. Limitations in in reliability, validity, and other measurement properties provide a path forward for improving the FIQL.

1.1 Patient Reported Outcome Measures (PROMs)

1.1.1 Generic Instruments

Many instruments have been developed to measure health-related quality of life, called patient-reported outcome measures (PROMs). These PROMs generally fall into one of three categories: generic instruments, specialized instruments, and condition-specific instruments (10). Generic instruments are available, such as the Medical Outcomes Study Short Form General Health Survey (SF-36) (11), Nottingham Health Profile (12), or EuroQol-5D (13). These scales are useful for comparisons across diseases and populations, but may not be sufficiently responsive to detect small changes over time for individuals with a specific disease or condition. For example, one study of the SF-36 found that the instrument performed sub-optimally as a repeated measure on the same individuals, and is much better suited to comparisons between individuals at a single point in time (14).

1.1.2 Specialized Instruments

Specialized instruments may be used to measure single aspects of quality of life, such as depression, anxiety or coping. Examples include the Beck Depression Inventory (15), Hospital Anxiety & Depression Scale (HADS) (16), or the COPE scale for measuring psychological coping (17). While these measures are well-validated and easy to interpret, they may only measure one or two aspects of the impact of health status on quality of life.

1.1.3 Condition-Specific Instruments

Finally, there are condition-specific instruments for measuring quality of life, such as the Gastrointestinal Quality of Life Inventory (GIQLI) (18), Inflammatory Bowel Disease Questionnaire (IBDQ) (19), or the FIQL (2). These instruments provide a multidimensional assessment of quality of life, and are designed to be sensitive to changes in condition-specific symptoms and treatment.

1.1.4 Evaluating PROMs

Deshpande et al. have identified a number of desirable properties of PROMs (20). These include specificity to the construct being measured, an evidence-based conceptual framework, and an optimal number of items. PROMs should also be easy to understand, confidential, and reproducible.

In addition, PROMs should be based on a thorough theoretical and empirical understanding of the construct of interest. Messick has provided a detailed framework for evaluating instruments not just on the basis of statistical characteristics, but with a focus towards assessing the "adequacy and appropriateness of interpretations and actions based on test scores" (21). By validating instruments like PROMs, we can learn about the theoretical constructs that the PROM intends to measure, the characteristics of the instrument itself, and the ramifications that the use and interpretation of a given PROM has for patients, physicians, and researchers.

If the construct of interest is composed of several domains, these domains should be identified through review of research and input from patients and physicians. Items measuring each domain should be included in the PROM, and responses to these items should span all levels of the domain. Instrument validation can then be used to provide empirical evidence

regarding the meaning and adequacy of the identified domains. The relationships between domains, outcomes, and other phenomena can be examined to ensure that the PROM is functioning as intended, and new findings or inadequacies can be addressed in new or modified versions of the PROM (22).

Certain statistical properties can be used to assess these characteristics of a PROM. To connect statistical methods to the larger question of validity, it is instructive to consider two broad statistical properties of instruments: reliability and validity.

Reliability is a measure of the precision of an instrument – an instrument with high reliability will produce similar results for the same individual under similar conditions.

Instruments demonstrate high validity if they measure what they claim to measure. Validity may be demonstrated through statistically significant correlations with similar measures and expected outcomes, and lack of correlation with unrelated measures and outcomes. Validity can also be demonstrated through the logical grouping of item responses in factor analysis. PROMs should be interpretable, with meaningful minimum or maximum thresholds identified, and a well-established value for minimally important change (i.e. the smallest change in instrument score that a patient would consider significant). Finally, if any of these properties vary across subgroups it should be well documented and investigated (22).

Modern views of instrument validation disagree with this strict dichotomy between reliability and validity, instead arguing that validation studies should be focused on integrating all available information into assessing the "meaningfulness, appropriateness, and usefulness of the specific inferences made from test scores" (23). Therefore, this thesis is intended as a starting point for continued research into the development, use, and interpretation of the FIQL.

1.2 Background

1.2.1 Fecal Incontinence Quality of Life Scale

For measuring the effects of fecal incontinence on quality of life, Rockwood et al. proposed the Fecal Incontinence Quality of Life Scale in 2000 (2). The scale contains 4 questions and 29 items. Questions 1 and 4 are each a single item. Question 2 contains 13 items (a – m) and question 3 contains 14 items (a – n). The instrument measures four domains: lifestyle (10 items), coping/behavior (9 items, subsequently referred to as "coping"), depression/self-perception (7 items, subsequently referred to as "depression"), and embarrassment (3 items). Items for questions 2 and 3 have four response categories, ranging from "Most of the time" to "None of the time," and "Strongly agree" to "Strongly disagree," respectively. These responses are graded from 1 to 4. Item 1 is answered on a scale from 1 to 5 ("Excellent" to "Poor"), and item 4 from 1 to 6 ("Extremely so – to the point I have just about given up" to "Not at all").

Responses are then averaged for each domain. Item 1 is reverse coded before averaging. Lower scores indicate lower quality of life. There is no established method for calculating an overall score.

1.2.2 Properties of the FIQL

The FIQL was developed to measure the impact of FI on quality of life, separate from symptom severity (2). To develop the instrument, a panel of experts was convened that identified six domains of quality of life that they suspected would be impacted by FI. Forty-one items were developed to measure these domains and pilot tested on 50 patients. After pilot testing, two domains were eliminated and the number of items was reduced to 29. See Appendix A for details on scoring.

The FIQL demonstrated adequate psychometric properties, including test-retest reliability as demonstrated by non-significant *t*-tests, internal reliability (Cronbach's alpha = 0.8 to 0.96), statistically significant correlations with 6 selected subscales of the SF-36, and the ability to discriminate between patients with FI and healthy controls.

The FIQL has been translated into French (24), Portuguese (25), Italian (26), Spanish (27), Turkish (28), Japanese (29–31), Norwegian (32), German (33), Dutch (34,35), and Chinese (36). These translations have demonstrated generally good internal and test-retest reliability, responsiveness, and convergent and discriminant validity. Furthermore, the FIQL is relatively short, with many patients completing the survey in about 10 to 12 minutes (24,30).

However, this thesis appears to be the first study to evaluate the FIQL in English in its original format (with the single alteration that the "N/A" response option was removed, see Section 3.1.1: Missing Data in the FIQL for details).

1.2.3 Limitations of the FIQL

While the original development and subsequent translation studies have shown generally good characteristics, these studies also revealed several conceptual, psychometric, linguistic, and cultural shortcomings of the instrument.

The FIQL may have conceptual limitations due to its method of development. Patients were not included in identifying the domains impacted by fecal incontinence, nor in the creation of items to measure these domains, and were only involved in pilot testing and validation (2). This raises questions about the relevance and interpretation of this instrument from patients' perspectives. Furthermore, the authors did not identify which samples were used for confirmatory factor analysis to eliminate items and identify domains, and what the demographics

of those samples were, making it unclear for whom this instrument may be best suited, and whether its psychometric properties should be consistent across samples with different demographics.

There are statistical limitations to the original development as well. Only 47 individuals were included for the test-retest validation and no power analysis was conducted. In fact, despite statistical non-significance, the lifestyle domain declined from 3.28 to 2.23 (range 1 to 4) in the two testing windows. And while the FIQL domain scores were almost always statistically significantly correlated with subscales of the SF-36, this only suggests that the measures were not uncorrelated, but does not test whether the correlation was within an expected range.

This was true for translation studies as well, with wide ranging Pearson correlation coefficients that were sometimes quite low. For example, embarrassment with social functioning ($r = 0.12 – 0.48$), lifestyle with social functioning ($r = 0.24 – 0.65$), and coping with mental health ($r = 0.29 – 0.54$) (2,27,30,33,36). One study found correlations no higher than 0.3 for embarrassment with all subscales of the SF-36, and no higher than 0.4 for coping with all subscales of the SF-36 (37). These findings indicate that either the FIQL embarrassment domain does a poor job of measuring similar constructs to subscales of the SF-36, or that it measures constructs that are distinct from SF-36 subscales.

Translation studies have revealed other limitations as well. Bols et al. studied the responsiveness and interpretability of the FIQL after translating it into Dutch (35). While they concluded that the FIQL was the best available tool for measuring quality of life for patients with FI, they also found that the depression domain had inadequate internal responsiveness and longitudinal construct validity, and raised concerns about the instrument's complicated

answering and scoring scheme. Simpler versions of the FIQL have been proposed, although are not in wide use (38,39).

Rullier et al. found that items in the embarrassment domain showed higher correlation with depression and coping than the total embarrassment score (24). All the translated versions of the FIQL demonstrated lower reliability for the embarrassment domain compared to the other domains (Cronbach alpha range = 0.51 – 0.85, ICC range = 0.59 – 0.89) (24,27,28,30,31,33,34,36). This could be due to cultural differences in concepts of embarrassment or shame (29), the small number of questions in that domain, or because item 2l does not appear to be a direct measure of embarrassment, although the literature is unclear if removing the question would have increased or decreased reliability (30,31).

Cultural considerations around religion and sexuality resulted in changes to many of the translated versions. One Japanese version removed items that ask about sex (items 3h and 3k), and many of the versions that retained these items reported high rates of missing responses (24,33,36). Item 2d, which references going to church, was altered or removed in the Turkish and Japanese versions (28–31). Three translations removed the "N/A" response option in order to avoid confusion (24,30,33). Several versions also replaced agree/disagree questions with frequency, so that all items had the same response options (27,29,31).

In addition to cultural considerations, there were concerns about the complexity of scoring and interpretation of the FIQL. The instrument does not provide an overall score, making use and interpretation more difficult. In addition, it is unclear how strongly each domain affects overall quality of life, since "there is no *a priori* reason to believe that each scale [domain] is of equal importance" (2). Hashimoto et al. used estimated item thresholds to produce an alternative

scoring algorithm for a reduced version of the FIQL, but found it performed very closely to a simple sum score (29). More research in this area is needed.

In addition, and importantly, men have been underrepresented in almost all translation and validation studies of the FIQL, including the original development of the instrument (2,24,27–32,34–36,38). Many clinic-based studies have reported high proportions of women with fecal incontinence, attributing this difference to obstetric causes. However, this sex difference is not confirmed in population studies, and may be due to different health-seeking behaviors and lack of consistent screening (5,6,40). Since the FIQL was largely developed and validated with majority-female studies, and since fecal incontinence may have different effects on quality of life for men compared to women (40,41), care should be taken to ensure the FIQL performs similarly for men and women. It is also possible that differences in patient physiology, diagnosis, or surgical procedure could impact how one responds to the FIQL, and should be investigated.

Finally, previous studies of the FIQL have relied exclusively on CTT methods for their analyses. These analyses have only examined broad characteristics of the instrument, and have not provided item-level information on which aspects of the FIQL may be lacking or how to improve them. The one exception is a Japanese translation by Hashimoto et al. that used an IRT model (29). The results of this study cannot be readily applied to the English version of the FIQL, because it was a Japanese translation with significant alterations to the instrument, including reducing the number of items to 14 and changing response options.

Despite the limitations identified in the translation studies and minimal research on the English version, the FIQL is already in wide use to evaluate treatments for fecal incontinence. It has been used to examine the impact of treatments such as biofeedback (42,43), bulking

injections (44,45), sacral nerve stimulation (46,47), transanal minimally invasive surgery (48), and anal anastomosis (49), among others.

This study uses both CTT and modern IRT methods to thoroughly investigate the strengths and limitations of the English version of the FIQL for patients with FI awaiting elective colorectal surgery, and to provide detailed information on the overall and item-level characteristics of the instrument and recommendations for improvement. This thesis aims to: 1) confirm previous research on the reliability of the instrument using Cronbach's α, 2) describe the domains of quality of life that the instrument measures with EFA, 3) identify high and low quality items with IRT, 4) and determine whether one's score on the FIQL is influenced by gender or by expected surgical procedure using DIF analysis.

The strengths of the instrument identified in this study can provide clinicians and researchers more confidence in the appropriate use and interpretation of the FIQL, and limitations provide a map for improving the FIQL and measurement of FI-related quality of life.

Chapter 2: Data

2.1 Participant Recruitment, Inclusion and Exclusion Criteria

Patients were prospectively recruited from a convenience sample of 14 general and colorectal surgeons practicing in Vancouver Coastal Health (VCH) authority hospitals as part of a larger study of surgical waitlists (50). Patients were identified from VCH's electronic waitlist registry. Once eligible patients were enrolled in the elective surgical waitlist, they were contacted by phone by a trained VCH surveyor to explain the study. Patients were eligible for participation if the surgical procedure for which they were waiting was elective, they were over the age of 18, could speak and read English, and had a mailing address. Patients were not recruited if they could not be reached via the contact information provided through VCH, were not English-speaking, or were scheduled for surgery in less than two weeks (emergent cases). Patients were contacted within two weeks of being assigned to VCH's registry (50).

2.2 Survey Administration

Among patients agreeing to participate, a survey package was mailed. The package included the FIQL, along with the Fecal Incontinence Severity Index (FISI), a commonly used measure of the type and frequency of incontinent episodes (51). Patients' age, gender, and self-reported chronic health conditions were collected at the same time. Participants were provided the option of completing the surveys online, and surveys received by mail were manually entered into a database. Patient privacy was ensured through a Privacy Impact Assessment performed by the VCH Legal and Privacy Office, and the dataset was de-identified prior to analysis.

Participants with a score of 0 on the FISI, indicating no incontinence, in addition to participants who did not complete any of the FISI, were excluded from analysis. Patients were included regardless of suspected or confirmed cancer diagnosis.

Chapter 3: Methods

Four statistical methodologies were used to evaluate the FIQL. Reliability was assessed with Cronbach's α statistic, and the dimensionality and validity of the instrument investigated with EFA. IRT was employed to assess item- and test-level characteristics. Finally, both differential item functioning and differential test functioning for gender and expected surgical procedure were evaluated using IRT. In each of these methodologies, a strategy for handling missing response data was identified and applied.

3.1 Missing Data

Missing data is a common problem in health research focused on patient reported outcome measures (52). There are three kinds of missing data that have different implications for statistical inference. Data may be missing completely at random (MCAR), which means that the probability of a variable being missing is not dependent on any other variable in the dataset or the value of the variable itself. Excluding data that is MCAR will not bias results, but will decrease power and precision. Data may also be missing at random (MAR), meaning that its probability of missing is dependent on observed variables. If one conditions on observed variables, the probability of the variable being missing becomes MCAR. That is, other variables in the dataset can account for the missingness, and can be used to adjust for loss of power or bias. Finally, data may be missing not at random (MNAR). In this situation, the probability of data being missing is dependent on the unobserved variables' values. This is the most serious kind of missing data, since one cannot easily adjust for its effect on statistical inference, which may produce biased results (52,53).

The most common modern method for handling missing data is multiple imputation. Multiple imputation involves using an algorithm, often Markov Chain Monte Carlo or iterative chained equations, to impute a set of plausible values for the missing data based on observed variables in the dataset. This process allows one to compensate for lack of power if data is MCAR, and remove potential bias if data is MAR (54). Even in cases where data is MNAR and as much as 30% of data are missing, multiple imputation can minimize bias to less than 5% (53).

For this reason, multiple imputation will be employed where feasible to ensure estimates are precise and unbiased. An important consideration in multiple imputation is the number of imputations needed to obtain estimates with acceptable levels of precision and bias. While previous recommendations based on relative efficiency have suggested as few as 3 to 5 imputations may be adequate, simulations have shown that a larger number of imputations is necessary to ensure sufficient power, with 100 imputations likely more than adequate for the purpose of this thesis (55). All IRT and DIF analyses were performed on 100 imputed datasets, and the results of all imputations were plotted in each figure to evaluate the sensitivity of the findings to the number of imputations.

3.1.1 Missing Data in the FIQL

The original FIQL scale included "N/A" as a response category for questions 2 and 3, with the instructions "If it is a concern for you for reasons other than accidental bowel leakage then check the box under Not Apply (N/A)." The scoring guidelines then state "Not Apply is coded as a missing value in the analysis for all questions."

This response category was not included in the FIQL administered to patients in this study, and was the only change made to the instrument prior to administration. There are three

reasons for this alteration: this response option was only included for questions 2 and 3 even though no distinction is made between these and the other questions when scoring, it conflates a response of "N/A" with item non-response, and it has the potential to confuse participants. By omitting this category, all missing data can be treated consistently through imputation, since missing data only occurs when a participant fails to answer an item.

There is also ambiguity in calculating and interpreting the depression domain. Since questions 1 and 4 are answered on different scales (1 to 5, and 1 to 6, respectively) from questions 2 and 3 (1 to 4), if missing values are excluded from the mean calculation, there are two maximum scores for this domain: 4.43 if answering all questions (average of highest response to each item), or 6 if answering only question 4 and omitting the rest. Adding to the confusion, the original explanation of the scoring algorithm states that domains "range from 1 to 5" (2).

When interpreting the demographics table, the maximum observed score was 5.5. In subsequent analyses this issue can be safely ignored, since item response theory and differential item and test functioning rely on sum (rather than mean) scores, and missing values were imputed prior to analysis.

3.2 Classical Test Theory and Exploratory Factor Analysis

3.2.1 Classical Test Theory: Reliability & Cronbach's α

Classical test theory is used to evaluate the reliability and validity of instruments. It is based upon the theoretical mathematical relationship

$$X = T + E \tag{3.1}$$

where X is the observed test score, which is composed of the sum of T, the true value of the construct of interest, and E, random error.

Reliability (ρ) is a measure of the accuracy of an instrument, measured as the ratio of true score variance between parallel tests (σ_T^2) and total score variance (σ_X^2), written

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}. \tag{3.2}$$

Since the total score variance is composed of true score variance and error variance (σ_E^2), as error decreases, reliability approaches 1. Instruments that accurately and consistently measure the construct of interest (have low error variance) will demonstrate high reliability.

In practice, true score variance is not observed. One statistic used to estimate reliability is Cronbach's α (see equation 3.3). It can be used to measure the internal consistency of a given instrument, and is considered a lower-bound estimate of reliability. The estimated, standardized coefficient α can be calculated as

$$\alpha = \frac{p}{p-1} \times \frac{\sum_{i \neq j} \widehat{Cov}(X_i, X_j)}{\widehat{Var}(X)} \tag{3.3}$$

where p is the number of test items, X_i is the ith item on a test, and $X = \sum_{i=1}^{p} X_i$. The test statistic α estimates ρ when items on an instrument measure the same underlying construct (unidimensional), in the same scale on the true score (tau-equivalent), and error scores are uncorrelated (22). For this reason, α was calculated for each of the FIQL domains. Values of α between 0.7 – 0.8 are generally deemed acceptable, 0.8 – 0.9 good, and greater than 0.9 as excellent when measuring internal consistency, although a high value for α can be due to reasons other than internal consistency (56). Failure to achieve a high value for α may indicate a need to add additional test items, or remove items that are poor measures of the underlying construct. All

available pairwise correlations were used to calculate Cronbach's α, to minimize the loss of information from missing data.

3.2.2 Exploratory Factor Analysis

EFA is a tool used to understand the relationships between observed variables and the latent constructs the items purport to measure, called factors. The goal of EFA in the context of instrument validation is to determine the number of factors measured by an instrument, their meaning, and the relationships between factors and individual items. Mathematically, factor analysis is written as solutions to the system of equations

$$\underline{X} = \boldsymbol{A}\underline{f} + \underline{u} \tag{3.4}$$

where $\underline{X}$ is a vector of observed item responses, $\boldsymbol{A}$ is a matrix of coefficients (factor loadings), $\underline{f}$ is a vector of factors, and $\underline{u}$ is a vector of item-specific errors. The factor loadings are typically standardized to range between -1 and 1. This model represents each observed item score as a linear combination of underlying factors and a unique error term. The model assumes that errors are uncorrelated with the factors, and among themselves.

For example, an individual's observed response on the 29 items on the FIQL can be represented as the application of factor loadings ($\boldsymbol{A}$) to the vector of factors ($\underline{f}$). If the number of factors is equal to the number of items, then each factor simply represents a single item, and the error term goes to zero. However, PROMs generally have multiple items measuring a single underlying construct, with a single factor representing this construct. In the FIQL, four factors (one for each domain) are expected to explain a majority of the variance in the data. In this case, $\underline{X}$ would be a 29 x 1 vector of item responses for a single individual, $\boldsymbol{A}$ would be a 29 x 4 matrix

of factor loadings, $\underline{f}$ would be a 4 x 1 vector of factors, and $\underline{u}$ would be a 29 x 1 vector of item-specific error terms.

Factor loadings represent how strongly correlated an item is with a factor. Items with higher factor loadings (in absolute value) have a stronger influence on the factor's value, and are said to "load" on that factor. While there is no strict cut-off, items with a factor loading greater (in absolute value) than 0.3 are often considered to meaningfully measure a factor (57).

Once identified, factors can be used to describe the structure of the observed data. Well-performing instruments should demonstrate a structure consistent with their stated goals. In other words, items should have high loadings on the factors associated with the domain they claim to measure (e.g. an item asking about behavior should load highly on the factor that represents behavior). Furthermore, items should not carry high loadings on unrelated factors (e.g. an item asking about behavior should not load highly on the factor that represents embarrassment). When an item has high loadings on multiple factors, this is termed cross-loading.

There are infinitely many solutions to the system of equations described in equation 3.4 because factors are unobserved. EFA proceeds by using the information from an initial solution to determine the appropriate number of factors, then rotating those factors according to optimization criteria provided by the researcher to produce a single, most-interpretable solution, and estimating standardized factor loadings and correlations.

There are two commonly used methods for identifying factors: maximum likelihood and principal components analysis (PCA). Maximum likelihood allows for estimation of standard errors of factor loadings and inference on model fit, but assumes that the data follow a multivariate normal distribution. PCA requires no distributional assumptions, but does not allow

for inference based on the multivariate normal distribution (58). PCA was used for model fitting because the item responses did not meet the assumptions for ML estimation.

Once factors are identified, one must determine the number to retain before it is possible to produce estimates of factor loadings and correlations. One can make this decision *a priori* for theoretical reasons, or select the number of factors based on criteria. Since one goal of this analysis was to determine if there is evidence for four factors, the decision was not made *a priori.* Factors with an eigenvalue greater than 1 were retained, since this implies the factor accounts for more than one item's variance. The model fit was evaluated descriptively by the percent variance explained by the retained factors.

Once the number of factors have been identified they are then rotated to obtain the most interpretable fit. There are two main categories of factor rotation. Orthogonal rotation keeps factors uncorrelated. Oblique rotation allows factors to rotate freely, often resulting in a more interpretable fit to the data, but may produce factors that are correlated. In the case of the FIQL, oblique rotation based on the oblimin criteria (with tau=0) was used since there is no reason to believe that the domains measured by the FIQL (lifestyle, coping, depression, and embarrassment) should be uncorrelated. The oblimin option in SAS allows for high factor correlation, but produces better row parsimony and fewer item cross-loadings, increasing interpretability (59,60).

EFA was performed on the correlation matrix of all available pair-wise correlations to minimize the effect of missing data.

3.3 Item Response Theory

Item response theory (IRT) was developed as a method to evaluate the performance of individual items on an instrument, and consists of a variety of statistical models. These models first estimate an underlying trait θ that the item or instrument is supposed to measure, like the factors identified in factor analysis. One can then represent item response patterns (called item characteristic curves, ICC), expected test score, item and test information, and the standard error of estimation of an instrument. The appeal of IRT over CTT methods is that these values are functions of the underlying trait θ, rather than a constant averaged across a population, so they provide a more detailed analysis. This allows one to determine the range of values for θ for which each item, or the instrument, is most accurate. One can also determine how well individual items and the instrument can measure this underlying trait θ.

In the case of the FIQL, the underlying trait θ can be thought of as the true value of an individual's quality of life along one domain, and is assumed normally distributed. For example, if an IRT model determines that $\theta = -2$ for an individual on the lifestyle domain, this indicates that the individual's true lifestyle-related quality of life is two standard deviations below the average lifestyle-related quality of life in the sample. Item response theory can then be used to determine if that individual is responding to the items of the instrument in a manner consistent with that level of lifestyle-related quality of life.

3.3.1 Graded Response Model

While there are a variety of IRT models, Samejima's graded response model (GRM) is appropriate for measuring ordinal response patterns, such as those found in the FIQL. This model

uses the logistic cumulative distribution function to estimate the probability of responding at a certain level to each item, written:

$$P_{jk}^{*}(\theta) = \frac{1}{1 + \exp\left[-a_j\left(\theta - b_{jk}\right)\right]}$$

and

$$P_{jk}(\theta) = P_{jk}^{*}(\theta) - P_{j(k+1)}^{*}(\theta) \tag{3.5}$$

where θ is the continuous, normally distributed underlying trait, measured by j items, each with k levels. The parameters of interest are a_j, which is the slope of the logistic curve of the item, termed "discrimination", and b_{jk} which is the boundary between two response levels on the same item, termed "difficulty." Because one can only respond in one of k ways to a given item, the probability function $P_{jk}(\theta)$ is the probability of choosing that response level, minus the probability of choosing the next response level $(k + 1)$. The probability of responding below the first response is set to 1, and of responding above the highest level is 0 (61).

Each item has one discrimination parameter, a_j, and $k - 1$ difficulty parameters, b_{jk}. The discrimination parameter measures the slope of the logistic curve for that item. High discrimination indicates a clear division between adjacent response categories. This is desirable because it means that individuals with levels of the underlying trait that are close to a boundary will respond consistently each time. In the case of low discrimination, the choice between two adjacent options will be less consistent.

Difficulty measures the levels of the underlying trait at which one is ambivalent between choosing adjacent response categories. Since the underlying trait θ is assumed to have mean 0 and variance 1, the units can be interpreted as standard deviations. Consider a question with 4 response options and high item difficulty with thresholds at $b_1 = 0.5, b_2 = 1,$ and $b_3 = 1.5$. This

would indicate that individuals with an underlying trait value $\theta = -1$ will be most likely to respond to the item by choosing the lowest response category, since $-1 < 0.5$. If the discrimination parameter is also high, even individuals near $\theta = 0.5$ will respond consistently. For example, individuals with $\theta = 0.4$ will still have a high probability of choosing the lowest response option, while those with $\theta = 0.6$ will have a low probability of choosing the lowest response option, and a high probability of choosing the next option.

Finally, one can produce information functions for both items and the instrument. The information function is an estimate of the reliability of the item or instrument. Unlike CTT statistics, information is a function of the underlying trait, rather than a single value, making it comparable across populations with different distributions of the trait. Items with higher discrimination and minimal variance produce higher information functions. The test information function for the instrument is simply the sum of the individual item information functions. For test information functions, the standard error of estimation, SE(θ), for each level of the underlying trait can be calculated as:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.6}$$

where $I(\theta)$ is the test information function.

The GRM requires two assumptions of the data: unidimensionality (all items in the scale measure one underlying construct), and local item independence (within the scale, an examinee's response on one item does not affect their response on another item) (62).

To evaluate the FIQL per its scoring algorithm, four unidimensional graded response models were fit to the data, one for each of the FIQL's domains. Unidimensionality was assessed through exploratory factor analysis on each domain. The domain was considered unidimensional

if EFA produced 1 factor with eigenvalue greater than 1, representing a large proportion of the variance (63). Each item should also carry a high loading on this factor.

These graded response models were fit using the expectation-maximization (EM) algorithm, and fit was assessed through a generalized M_2 statistic (64), which tests the null hypothesis that the model fits exactly; root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR), which are indices of absolute fit where 0 indicates perfect fit; and the comparative fit index (CFI) and Tucker Lewis index (TLI), which measure fit relative to a null model with penalties for additional complexity. Acceptable values for these statistics are provided in Table 4.6.

Item fit was determined from the $S - X^2$ statistic, with $p > 0.05$ indicating acceptable fit (65). Items were evaluated for local item dependency (LID) using Pearson's X^2 statistic, which measures the correlation between deviations from expected and observed responses on two items (66). In line with other research, residual correlations >0.20 were considered evidence of local dependence (63,67,68).

Model and item fit statistics, and item difficulty and discrimination parameters were calculated following 100 imputations of missing data. Following model fitting, items were evaluated qualitatively based on difficulty, discrimination, and information. Each domain was also assessed for overall information.

3.4 Differential Test and Item Functioning

An important concern for evaluating PROMs is whether they perform differently for different subgroups of a population, even after controlling for underlying quality of life θ. If an instrument demonstrates this characteristic, one may reach erroneous conclusions about

differences between these subgroups, or over- or under-estimate the reliability of the instrument in populations with different relative sizes of these subgroups. These differences can occur for individual items, on the instrument as a whole, or both.

3.4.1 Differential Item Functioning

Differential item functioning (DIF) refers to the differential performance of an individual item between two or more subgroups after adjusting for ability level. In general, items that exhibit DIF should be removed or modified. However, if the combination of all items on an instrument do not produce differential test functioning (DTF, see section 3.4.2), removal or modification of these items may be unnecessary (69).

There are many methods available for detecting DIF. In general, they fall into two categories: IRT-based methods and non-IRT-based methods. Non-IRT-based methods include Mantel-Haenszel, standardization, logistic regression, and Breslow-Day procedures, among others (70). These models typically use the observed test scores to adjust for ability level.

In IRT-based methods for evaluating DIF, comparisons are adjusted for the true underlying quality of life θ, rather than observed test score. There are two main IRT-based methods for evaluating DIF in graded response models.

The first IRT-based method is a likelihood ratio test. In this method, a model with parameters constrained to be equal across groups is compared with a model in which these parameters vary. However, the likelihood ratio test may produce inflated type I errors, especially in the case of poor model fit (71,72).

The second method, Lord's Wald test (73), tests for differences in parameters directly. It is based on the test statistic Q_j for item j, defined as

$$Q_j = \left(\underline{\nu}_{jR} - \underline{\nu}_{jF}\right)' \left(\boldsymbol{\Sigma}_{jR} - \boldsymbol{\Sigma}_{jF}\right)^{-1} \left(\underline{\nu}_{jR} - \underline{\nu}_{jF}\right) \quad (3.7)$$

where $\underline{\nu}_{jR}$ is the vector of parameter estimates for item j in the reference group R, and $\boldsymbol{\Sigma}_{jR}$ is the variance-covariance matrix for item j in the reference group. This statistic follows a χ^2 distribution with degrees of freedom equal to the number of parameters estimated in the model. For most items in the FIQL, 4 parameters were estimated (one for discrimination, and 3 difficulty parameters), although questions 1 and 4 had 5 and 6 parameters, respectively. This test statistic can be used to evaluate the null hypothesis that the difference in parameter estimates between two groups is 0 (70), and performs well in the context of the graded response model (74).

However, Lord's method requires that item parameters have a common metric to differentiate DIF from distributional differences between groups. To produce this metric, certain "anchor items" must be assumed DIF-free. In addition, the presence of DIF in one item can affect the estimation of DIF in other items, leading to type I errors (75). To correct for these problems, the means and variances for the reference and focal groups should be allowed to differ, and items without DIF should be identified and used as anchors (70).

There are three main anchor strategies to determine the number of anchor items. One can select a constant number of anchor items (e.g. 4, 10%, or 20%), one can use all other items as anchors, or one can select the number of items iteratively, either forward or backward, based on some criteria. This thesis will use iterative forward selection based on the mean p-value threshold criteria (MPT) suggested in Kopf et al., since it performed better than other strategies, at least in the context of a dichotomous Rasch model (75). This strategy proceeds in two steps:

1. Calculate a critical value that can be used to select anchors:

a. For k items, perform Lord's Wald test using each of the items as a single anchor, such that every item will have $k - 1$ associated p-values
b. Calculate the mean of these p-values for each item
c. Sort these item-specific mean p-values from largest to smallest
d. Select the $(0.5 * k)^{th}$ p-value as a threshold. If there are an odd number of items, round up
e. Finally, the critical value c_j for each item is the number of p-values higher than this threshold value

2. Use iterative forward selection to select items:
    a. Items can now be sorted from largest to smallest by c_j
    b. In the iterative forward selection process, one begins by using the item with the largest c_j value as an anchor, then adding additional items in order of descending c_j. In the case of ties, an item is chosen randomly.
    c. The iterative process continues while the number of anchor items is less than the number of non-significant p-values

3.4.2 Differential Test Functioning

Once appropriate anchor items have been identified, one can estimate test score functions (T(θ)) for each subgroup. These test score functions represent the score an individual with a given quality of life level θ would be expected to receive if administered the test. Differential test functioning occurs when this test function differs between two subgroups.

Chalmers et al. (69) have proposed two statistics for measuring differential test functioning: the signed DTF (sDTF) and unsigned DTF (uDTF):

$$sDTF = \int [T(\theta, \psi_R) - T(\theta, \psi_F)] g(\theta) d\theta$$
$$uDTF = \int |T(\theta, \psi_R) - T(\theta, \psi_F)| \, g(\theta) d\theta \quad (3.8)$$

where ψ_R is the collection of parameters for the reference (or baseline) group, ψ_F is the same for the focal group (the group in which DIF is suspected), and $g(\theta)$ is a weighting function such that $\int g(\theta) \, d\theta = 1$. These statistics are calculated using the estimated item parameters $\hat{\psi}_R, \hat{\psi}_F$ and evaluated numerically using discrete quadrature nodes. These statistics measure the magnitude of the difference between the test score functions over the range of θ.

The sDTF statistic ranges from $-TS$ to TS where TS is the maximum test score (sum of all items), and is a measure of the average amount of scoring bias that exists between the two groups across the range of θ. This statistic is in the same units as the instrument (when measured as a sum score), and negative values indicate that the reference group scores lower, on average, than the focal group. For example, an sDTF score of 1 when comparing female (reference) to male (focal) participants would indicate that scores for women will be 1 unit lower, on average, than for men, after adjusting for underlying ability level.

The uDTF statistic ranges from 0 to TS, and represents the average area between the two expected test functions. It describes the absolute deviations in item properties over the test. A value of 0 indicates that the expected test functions are identical and no DTF is present. Both the uDTF and sDTF can be represented as percentages of the total score. Furthermore, because the sDTF and uDTF statistics are evaluated over a range of θ, one can investigate what levels of θ have high or low levels of DTF.

3.4.3 DIF and DTF in the FIQL

The FIQL was evaluated for DIF and DTF based on gender and surgical procedure. When testing for DIF by surgical procedure, individuals waiting to undergo prolapse repair (n = 24) and recto-vaginal fistula repair (n = 7) were excluded from the analysis, since there were too few individuals to make meaningful comparisons. Those waiting to undergo tumor excision were compared to those waiting to undergo fissurectomy or fistulectomy, two subgroups with an approximately equal number of patients.

Differential test functioning was examined through qualitative investigation of expected test function plots, and quantified with sDTF and uDTF statistics.

DIF was be determined using Lord's method. Anchors were selected using the iterative forward MPT procedure described above. Due to the complexity of this anchor selection method, it was only performed on a single imputed dataset. Once anchors were selected, results were based on analysis of 100 imputations.

3.4.4 Statistical Software

Demographics, missing data, Cronbach's α, and exploratory factor analysis were all generated using SAS software, Version 9.4 for Windows (SAS Institute Inc., Cary, NC, USA). Item response theory and differential item and test functioning were performed using R, Version 3.3.1 (76), using the mirt package (77). Figures were produced using the ggplot2 package (78).

Chapter 4: Results

4.1 Sample Demographics

A total of 880 patients were eligible for participation. Among eligible participants, the response rate was 43.4%. Of these, 4 duplicated observations were removed, resulting in 317 completed questionnaires.

Sixty-eight patients' responses were excluded for having a score of 0 on the Fecal Incontinence Severity Index (indicating no incontinence). An additional 8 were excluded for not completing any items of the FISI. Of those, 5 also did not complete any of the items of the FIQL. Four individuals had two surgeries on different dates, at least 90 days apart, and filled out the questionnaire each time, and were included twice in the dataset and treated as independent observations. While this introduces some amount of bias into the model, the number of correlated observations is very small relative to the number of patients completing the instrument. A total of 236 participants were included for analysis.

Demographic characteristics are provided in Table 4.1. Mean age was 58 years and approximately half were male. Using the patient-ranked scoring scale, the mean FISI score was 27 (range 3 – 61, higher values indicate more severe incontinence).

On the FIQL participant scores were highly left-skewed, with many participants responding in the highest categories for most questions. Based on a simple sum score across the entire instrument, 50 participants (21.2%) scored at least 115 out of 119 possible points. Using mean scores, participants scored lowest on the coping domain of the FIQL (indicating lower quality of life) with a mean of 3.02 out of 4, followed by embarrassment (mean = 3.16) and lifestyle (mean = 3.28). The mean score for the depression domain was 3.41. Most patients were either waiting to have an anal fissurectomy or fistulectomy (44%), or a transanal excision of a

rectal tumor or lesion (43%), with a smaller number planning to undergo a repair of rectal prolapse (10%) or recto-vaginal fistula (7%). In addition, most patients were not suspected of having cancer (70%), while 22% and 8% had suspected and diagnosed cancer, respectively.

Table 4.1: Sample Demographics

Variable		N	Mean (SD)	Percent	Missing
Age		236	58 (16.8)		0
FISI* Scores		217	27 (15.4)		19
FIQL** Scores	Coping	236	3.02 (0.95)		0
	Embarrassment	234	3.16 (0.91)		2
	Lifestyle	236	3.28 (0.88)		0
	Depression***	236	3.41 (0.93)		0
Gender	Male	131		54%	0
	Female	110		46%	
Procedure	Anal Fissurectomy/Fistulectomy	106		44%	0
	Transanal Excision of Rectal Tumor/Lesion	104		43%	
	Repair Rectal Prolapse	24		10%	
	Recto-Vaginal Fistula	7		3%	
Cancer	Not Suspected	167		70%	1
	Suspected	54		22%	
	Diagnosed	19		8%	

** FISI is the Fecal Incontinence Severity Index, higher values indicate more incontinence*

*** FIQL is the Fecal Incontinece Quality of Life scale, higher values indicate better quality of life*

**** Observed score range of 1 to 5.5. All other scales range 1 to 4.*

4.2 Missing Data

Table 4.2 provides information on item non-response. The two items concerned with sex (3h, "I have sex less often than I would like to," and 3k, "I am afraid to have sex") demonstrated the highest rates of non-response, at 11% and 9.3%, respectively. There were 9 individuals who omitted only those two questions. There were 8 additional participants who answered all questions except for the items in question 3, perhaps indicating an error in instrument administration, comprehension, or response fatigue. The remaining items had low rates of non-response.

Missing data did not appear to significantly impact the results of the analysis, as evidenced by the high degree of overlap across imputations (see Figures 4.1 – 4.5 and Appendix B). Only in identifying DIF for item 2g was there sufficient variability across imputations to reach a different conclusion, see Section 4.6.1.1 for details.

4.3 Cronbach's α

Cronbach's α was consistent with other validation studies of the FIQL, demonstrating excellent internal consistency for three domains, and a much lower value for the embarrassment domain, as shown in Table 4.3. Examination of the effect of individual items ("α if item deleted") reveals two items that would increase α if deleted, indicating poor correlation with their respective domains. Item 3h ("I have sex less often than I would like to") only had a 57% correlation with the coping domain, and item 2l ("I leak stool without even knowing it") was only 47.5% correlated with the embarrassment domain. Item 2l has demonstrated this property in other validation studies as well (24,31).

Table 4.2: Item Non-Response

Item	Missing	Percent Missing	Mean Response
1	4	1.7%	3.19
2a	4	1.7%	3.38
2b	4	1.7%	3.47
2c	7	3.0%	3.39
2d	5	2.1%	3.43
2e	6	2.5%	3.28
2f	6	2.5%	3.03
2g	4	1.7%	2.94
2h	7	3.0%	3.26
2i	3	1.3%	3.06
2j	5	2.1%	3.17
2k	5	2.1%	3.34
2l	4	1.7%	3.37
2m	4	1.7%	3.21
3a	10	4.2%	3.00
3b	10	4.2%	3.01
3c	11	4.7%	2.89
3d	10	4.2%	3.22
3e	10	4.2%	3.04
3f	10	4.2%	3.00
3g	12	5.1%	2.92
3h	26	11.0%	2.68
3i	10	4.2%	3.10
3j	8	3.4%	2.91
3k	22	9.3%	3.11
3l	11	4.7%	3.28
3m	9	3.8%	3.40
3n	10	4.2%	2.90
4	0	0.0%	5.06

Table 4.3: Cronbach α Estimates

Domain	Standardized α
Lifestyle	0.95
Coping	0.94
Coping without 3h	0.95
Depression	0.91
Embarrassment	0.76
Embarrassment without 2l	0.81

4.4 Exploratory Factor Analysis

Examination of eigenvalues indicated a three-factor solution, with a large drop in eigenvalue between factors 3 and 4, from 1.35 to 0.93. These three factors accounted for 71% of the variance. It is not surprising that only 3 factors were retained, since the embarrassment domain only contains 3 items, which may be insufficient to identify a common factor (58).

After rotation, only factors 1 and 2 were well correlated, with a correlation coefficient of 0.63. The other factors demonstrated less correlation at 0.33 for factors 1 and 3, and 0.27 for factors 2 and 3.

Factor loadings for each item are provided in Table 4.4, along with the domain ascribed to each item. The gray bars indicate the magnitude of the factor loadings graphically. The criteria of simple structure was not met, since many items cross-loaded on more than one factor. Factor 1 was composed primarily of items from question 3, and factor 2 with items from question 2. This could be due in part to the visual layout of the questionnaire, since the items are grouped by a response matrix. It could also be explained by the uneven distribution of domains in the two sections. Question 2 was primarily composed of items belonging to lifestyle and coping, and focuses on behaviors and actions. Question 3, however, contains all the items related to depression, and 2 of the three items for embarrassment. It is possible factor 1 picks up on the emotional impacts of fecal incontinence, while factor 2 measures behavioral impacts.

Factor 3 is more difficult to explain, since it contains generally low factor loadings, as well as items with negative loadings. Items from the beginning of the FIQL appeared to load more highly. Most likely this factor is not inherently meaningful, but accounted for variability in the item responses not well explained by the other two factors.

Table 4.4: Exploratory Factor Analysis

Item	Text	Domain	Factor1	Factor2	Factor3
1	In general, would you say your health is	Depression	44	-5	31
2a	I am afraid to go out	Lifestyle	28	53	30
2b	I avoid visiting friends	Lifestyle	32	40	37
2c	I avoid staying overnight away from home	Lifestyle	23	33	54
2d	It is difficulty for me to get out and do things like going to a movie or to church	Lifestyle	15	49	52
2e	I cut down on how much I eat before I go out	Lifestyle	5	53	49
2f	Whenever I am away from home, I try to stay near a restroom as much as possible	Coping	14	64	28
2g	It is important to plan my schedule around my bowel pattern	Lifestyle	8	64	29
2h	I avoid traveling	Lifestyle	17	45	52
2i	I worry about not being able to get to the toilet in time	Coping	0	91	0
2j	I feel I have no control over my bowels	Coping	-4	91	3
2k	I can't hold my bowel movement long enough to get to the bathroom	Coping	-8	94	-6
2l	I leak stool without even knowing it	Embarrassment	4	77	-18
2m	I try to prevent bowel accidents by staying very near a bathroom	Coping	16	69	21
3a	I feel ashamed	Embarrassment	71	18	-17
3b	I can not do many of the things I want to do	Lifestyle	64	28	9
3c	I worry about bowel accidents	Coping	53	51	-21
3d	I feel depressed	Depression	87	-8	8
3e	I worry about others smelling stool on me	Embarrassment	80	15	-34
3f	I feel like I am not a healthy person	Depression	87	-6	1
3g	I enjoy life less	Depression	85	2	3
3h	I have sex less often than I would like to	Coping	83	-5	1
3i	I feel different from other people	Depression	86	4	-4
3j	The possibility of bowel accidents is always on my mind	Coping	60	41	-19
3k	I am afraid to have sex	Depression	85	-6	-1
3l	I avoid traveling by plane or train	Lifestyle	72	6	19
3m	I avoid going out to eat	Lifestyle	61	18	18
3n	Whenever I got someplace new, I specifically locate where the bathrooms are	Coping	42	52	-4
4	During the past month, have you… wondered if anything was worthwhile	Depression	61	-8	29

** Factor loadings are rounded and multiplied by 100*

4.5 Item Response Theory

4.5.1 Assessing assumptions, model and item fit

Within each domain, exploratory factor analysis produced results consistent with the assumption of unidimensionality. In all cases, only one eigenvalue was greater than 1, and the first factor explained 70% of the variance on the lifestyle and coping domains, 64% of the depression domain, and 67% of the embarrassment domain. No item factor loading was less than 0.6 in its respective domain.

Table 4.5 provides information on local item dependence. Only one item in the lifestyle domain, and no items in the depression or embarrassment domains met the threshold for evidence of local dependence. The coping domain contained four item-pairs that met this threshold, indicating caution should be used in interpreting the results for this domain.

Results of the graded response model fit for each of the domains is provided in Table 4.6. Only the SRMSR indicated good model fit for the lifestyle and coping domains. Model fit statistics for the embarrassment and depression domains cannot be calculated since they have too few items relative to the number of response categories producing negative degrees of freedom for these statistics (64,79).

Parameter estimates and the $S - X^2$ item fit statistic P-value are provided in Table 4.7. Only two items indicated lack of adequate fit to the observed data: items 2f ("Whenever I am away from home, I try to stay near a restroom as much as possible") and 2i ("I worry about not being able to get to the toilet in time") from the coping domain.

Table 4.5: Local Item Dependence

Item Pair	X^2 Correlation	P-value
Lifestyle Domain		
3l, 3m	0.256	<0.001
Coping Domain		
2j, 3c	-0.224	<0.001
2m, 3c	-0.24	<0.001
3c, 3h	-0.227	<0.001
3c, 3j	0.206	0.001
Depression Domain		
(none)		
Embarrassment Domain		
(none)		

Table 4.6: Graded Response Model Fit Statistics

Domain	M^2 P-value	RMSEA	SRMR	TLI	CFI
(Ideal)	>0.05	<0.08	<0.08	>0.95	>0.95
Lifestyle	<0.001	0.12	0.06	0.83	0.9
Coping	<0.001	0.1	0.06	0.85	0.93
Depression	*	*	*	*	*
Embarrassment	*	*	*	*	*

** Not calculated, since degrees of freedom were too low*

4.5.2 Descriptive Item Results

As shown in Table 4.7, all items in all four domains had high discrimination parameters (a). All values for a were significantly different from 0 (P-values not provided here), except for item 3e ("I worry about others smelling stool on me"). Those with the lowest discrimination include question 1 ("In general, would you say your health is:"), 2l ("I leak stool without even knowing it"), and 3h ("I have sex less often than I would like to"), all with $a < 2$. However, discrimination values above 1.35 are still generally considered high discrimination (80).

All items except question 1 demonstrated low difficulty, with difficulty parameters (b_i) ranging from -2.68 at the lowest for question 4, to 0.25 at the highest for question 3h. This indicates that items perform best among individuals who have low levels of the underlying trait (low quality of life), but will be less accurate in distinguishing between individuals with higher quality of life.

Item characteristic curves (ICC) generally fell into one of three categories. These included ICCs showing good separation between response categories, separation only between three response categories, and limited separation between categories. Examples of each are provided in Figure 4.1. In this and all subsequent figures, all 100 imputations were plotted. Item 2i demonstrated excellent separation between categories, 3c only distinguished between 3 categories ("strongly agree," "agree", and "disagree"), and 2c showed poor separation. See Appendix B for ICCs for all items.

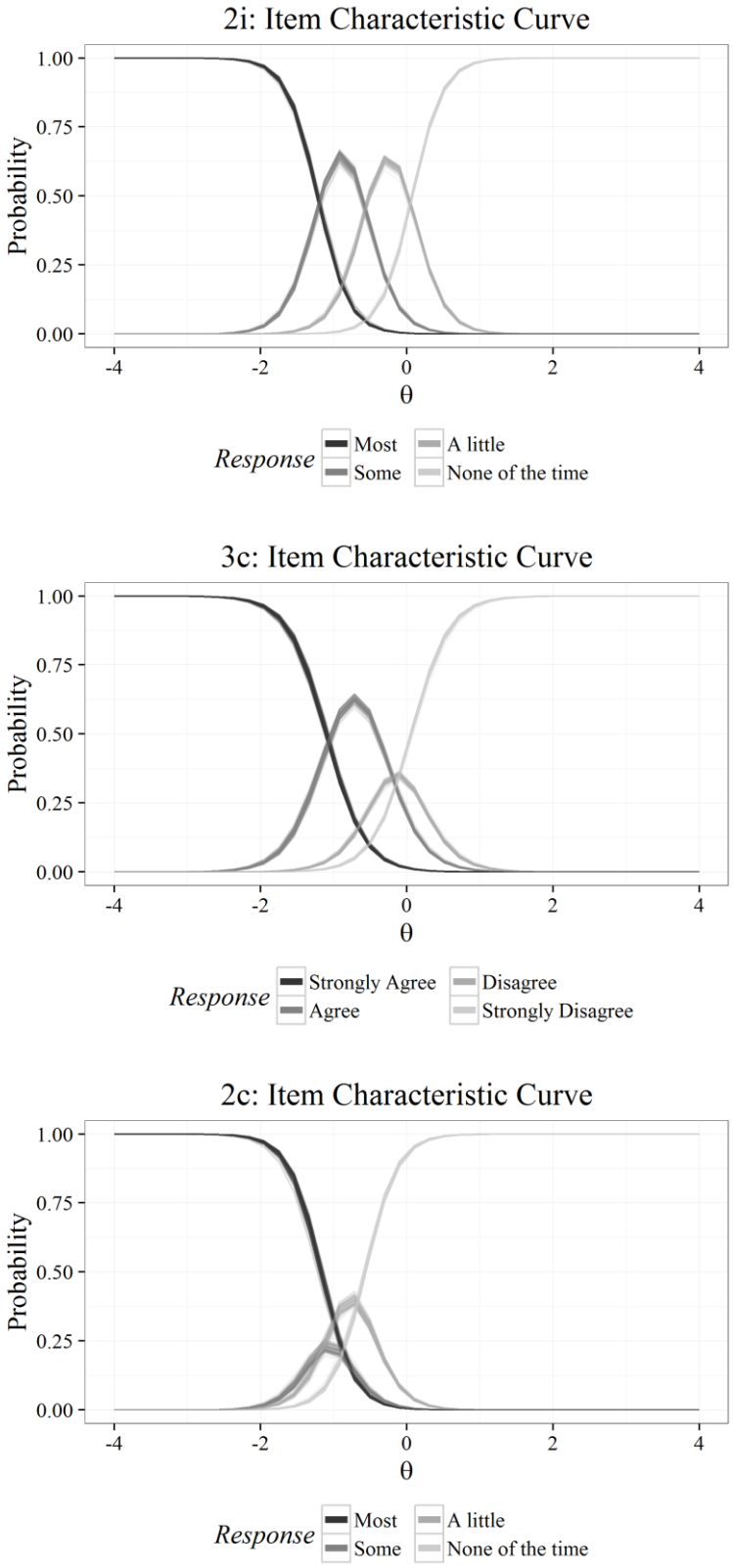
The ICCs for items 1, 2a,d,g-k,m, 3b, and 4 all demonstrated good separation. ICCs for items 2c, 2l, 3a, and 3h failed to distinguish between response categories based on underlying trait level. These items are good candidates for removal or replacement.

Table 4.7: Item Parameter Estimates and Fit Statistics

Item	a	b_1	b_2	b_3	b_4	b_5	$S-X^2$ P-value
Lifestyle Domain							
2a	5.4	-1.46	-0.91	-0.45			0.566
2b	5.52	-1.54	-0.91	-0.65			0.526
2c	4.48	-1.17	-0.96	-0.57			0.345
2d	6.23	-1.34	-1.00	-0.53			0.343
2e	3.05	-1.36	-0.80	-0.41			0.280
2g	3.03	-0.95	-0.47	0.01			0.056
2h	3.96	-1.23	-0.81	-0.35			0.263
3b	3.45	-1.19	-0.54	-0.05			0.211
3l	3.23	-1.38	-0.80	-0.45			0.253
3m	3.1	-1.68	-0.97	-0.57			0.347
Coping Domain							
2f	4	-0.92	-0.54	-0.03			0.035*
2i	4.75	-1.22	-0.57	0.07			0.025*
2j	3.76	-1.29	-0.74	-0.15			0.229
2k	3.07	-1.62	-1.00	-0.35			0.442
2m	5.58	-1.09	-0.66	-0.30			0.418
3c	3.82	-1.10	-0.33	0.06			0.090
3h	1.79	-0.82	-0.11	0.25			0.136
3j	3.11	-1.10	-0.41	0.01			0.089
3n	4.14	-1.05	-0.32	0.00			0.232
Depression Domain							
1	1.33	-2.33	-1.00	0.31	2.23		
3d	3.37	-1.63	-0.64	-0.26			0.063
3f	4.14	-1.20	-0.38	-0.06			0.123
3g	5.21	-1.08	-0.30	0.04			0.518
3i	4.33	-1.19	-0.48	-0.20			0.782
3k	2.64	-1.22	-0.61	-0.22			0.313
4	2.04	-2.68	-2.05	-1.48	-0.83	-0.06	0.345
Embarrassment Domain							
2l	1.48	-2.12	-1.26	-0.56			0.149
3a	2.53	-1.12	-0.53	-0.08			0.211
3e	6.55	-1.09	-0.43	-0.11			0.067

** Indicates significant misfit*

Figure 4.1: Item Characteristic Curves



The remaining 14 items differentiated between at least two, but not all, response categories. Eleven of these items came from question 3, and three from question 2. These questions might be improved by simplifying response categories (e.g. 3 instead of 4 response levels). Items 3c, 3h, 3j, 3l, and 3m might be considered for removal since they also demonstrated local item dependence. For example, question 3c ("I worry about bowel accidents") is very similar to, and correlated with, question 2j ("I feel I have no control over my bowels"), but did not perform as well.

4.5.3 Instrument Results

Test information functions for each of the four domains are provided in Figure 4.2. Since information is a function of the number of items in the scale, the total information provided was highest for the domains with the most items and lowest among those with fewer items. The minimum standard error of estimation for the lifestyle domain was 0.14, 0.16 for coping, 0.20 for depression, and 0.27 for embarrassment. However, due to local item dependence, the true SE(θ) for the coping domain should be interpreted with caution, as it may be larger.

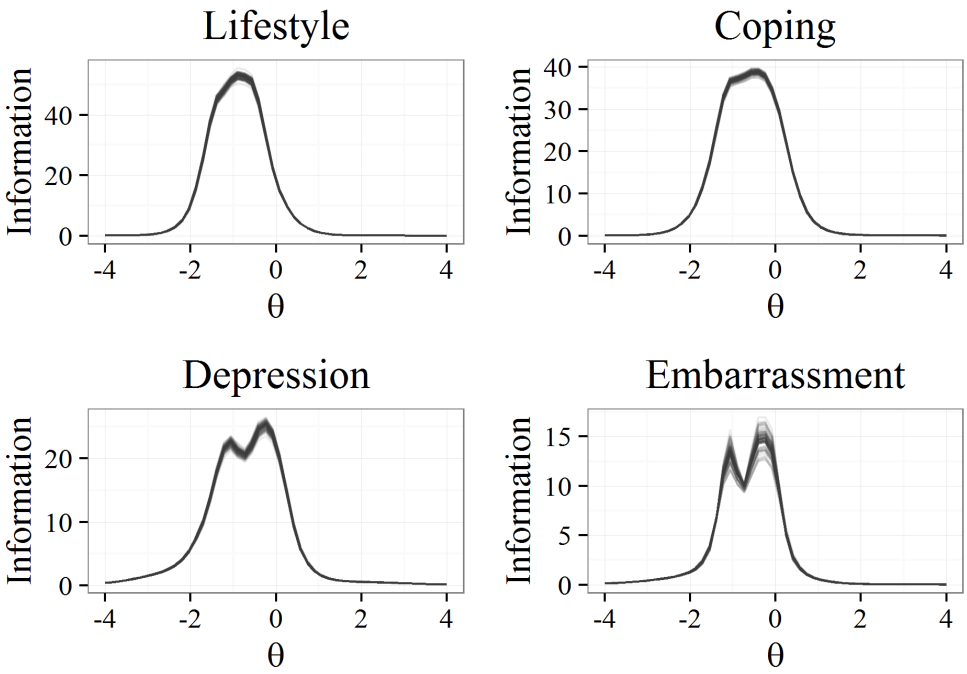
Most of the domains provide information for those with underlying quality of life θ from -3 to 1 standard deviations of the mean. The depression domain provided information for the widest range of values on the underlying trait (-4 to 1.5). This is likely because it contains questions 1 and question 4, which individually have widest range of difficulty parameters.

The dip in the information function for the depression and embarrassment domains was because several of the items in these domains did a poor job discriminating between the middle response categories. Participants with underlying quality of life near $\theta = -1$ demonstrated more

variability (and therefore lower information) in their response choices than those with lower or higher levels of quality of life, whose responses were more consistent.

These results suggest that items with higher difficulty could be added to the lifestyle, coping, and embarrassment domains to increase the range of quality of life that these domains can accurately measure. In addition, items could be removed from the lifestyle domain while still maintaining high reliability and minimizing response fatigue, and items could be added to the embarrassment and depression domains to increase reliability.

Figure 4.2: Information Functions



4.6 Differential Item and Test Functioning

4.6.1 Gender

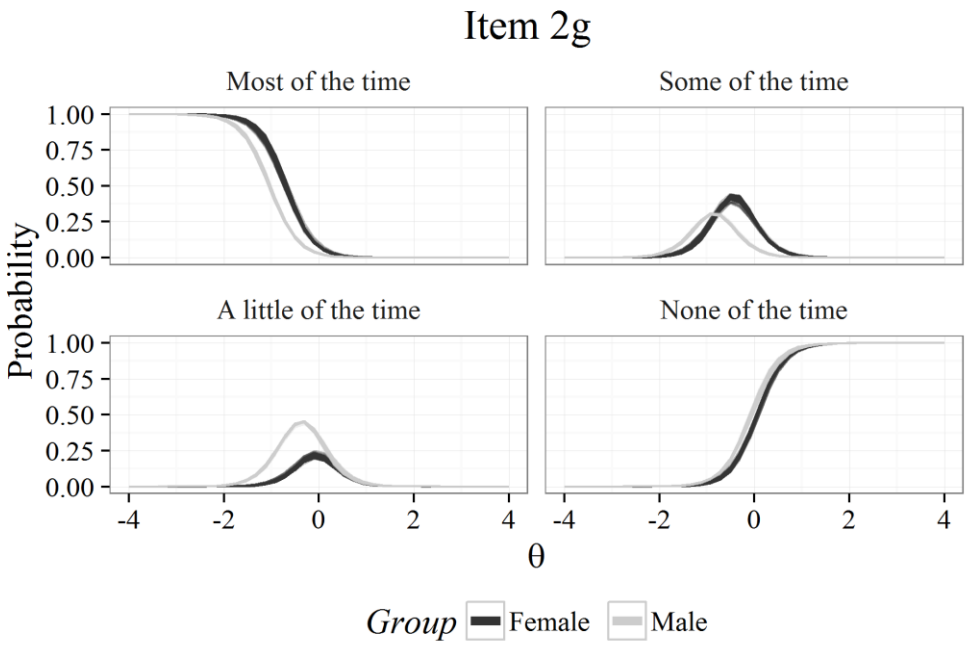
4.6.1.1 Differential Item Functioning by Gender

After selecting items 2a-d as anchors using iterative forward MPT, only question 2g ("It is important to plan my schedule (daily activities) around my bowel pattern" from the lifestyle domain demonstrated statistically significant DIF. This conclusion was dependent on imputation, with P-values ranging from 0.014 to 0.060, with the median P-value significant at 0.027.

Figure 4.3 is a plot of the estimated item characteristic curves by gender for item 2g. The plot indicates lower difficulty for female participants. That is, women were more likely to answer "most of the time" or "some of the time" compared to men with the same underlying quality of life θ, indicating that their observed scores for this item were lower than the true score would suggest.

Items 2i, 3c, 3h, and 3n were chosen as anchors for the coping domain; 3g, 3i, and 4 for the depression domain; and 3e for the embarrassment domain. Following anchor selection, no items demonstrated statistically significant DIF by gender in these domains.

Figure 4.3: Differential Item Functioning by Gender



4.6.1.2 Differential Test Functioning by Gender

Differential test functioning by gender was minimal. Visual inspection of the information functions reveals good overlap between the two groups, indicating similar accuracy for the three domains (not shown). The expected score functions for lifestyle, coping, and depression also failed to show differential test functioning (not shown). However, as indicated in Figure 4.4, there is a relatively large gender difference in test scores between men and women for the embarrassment domain for those with ability levels -6 to 0 standard deviations from the mean.

These descriptive findings were confirmed quantitatively. Scores differed between men and women, after adjusting for ability level, by less than 1 point for all domains, with women scoring slightly lower than men (indicating lower observed quality of life scores) on all domains except depression. See Table 4.8. However, for the embarrassment domain, this small difference represented a relative difference of about 1.4% in the sum score over the full range of ability levels, with the difference most pronounced for $-6 < \theta < 0$ (see Figure 4.4), which is also the range the items in the FIQL best measure. The relative difference over this range of θ was 2.7%. The larger difference in this domain is likely due to the small number of items in this domain, and may be eliminated if more items are added.

Figure 4.4: Differential Test Functioning for Embarrassment

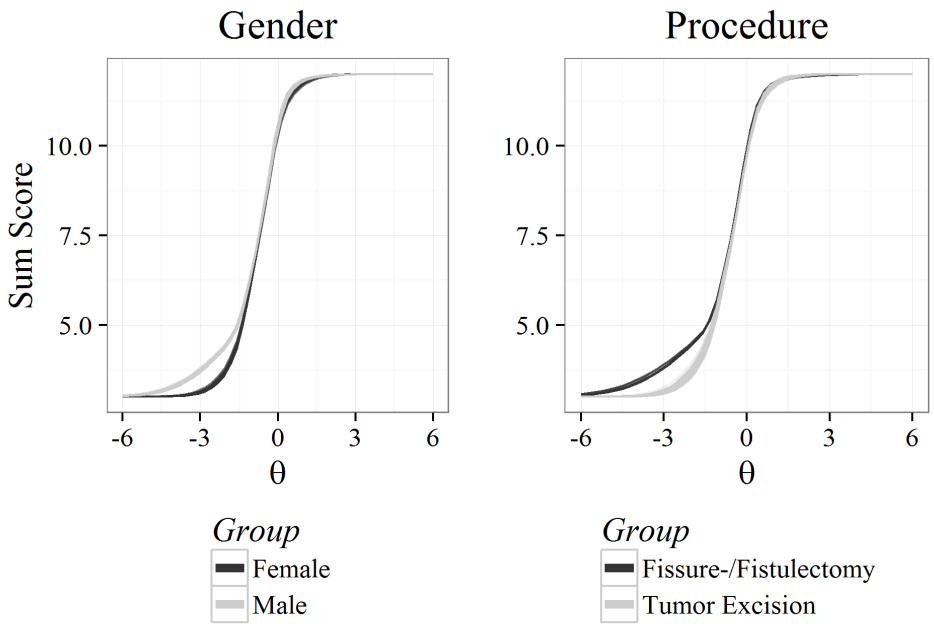


Table 4.8: DTF Statistics, Female vs. Male

	sDTF	% sDTF	uDTF	%uDTF
Lifestyle	-0.34	-0.86	0.34	0.86
Coping	-0.11	-0.30	0.11	0.30
Depression	0.03	0.11	0.21	0.67
Embarrassment	-0.17	-1.43	0.17	1.45

4.6.2 Surgical Procedure

When fitting the models as scored, the depression domain failed to produce an invertible information matrix, due to sparse response patterns to question 4. To address this model fitting problem, categories 1 and 2 ("extremely so" and "very much so") were collapsed for this analysis only. The other three domains were analyzed using all response categories.

4.6.2.1 Differential Item Functioning by Surgical Procedure

The iterative forward MPT method identified items 2a-b, 3l-m as anchors for the lifestyle domain; items 2i-j, 2m, and 3c for the coping domain; items 3d, 3f, and 3g for the depression domain; and item 3e for the embarrassment domain. After anchors were identified, no items demonstrated differential item functioning when comparing individuals waiting to undergo rectal tumor excision to individuals waiting to undergo fissurectomy or fistulectomy.

4.6.2.2 Differential Test Functioning by Surgical Procedure

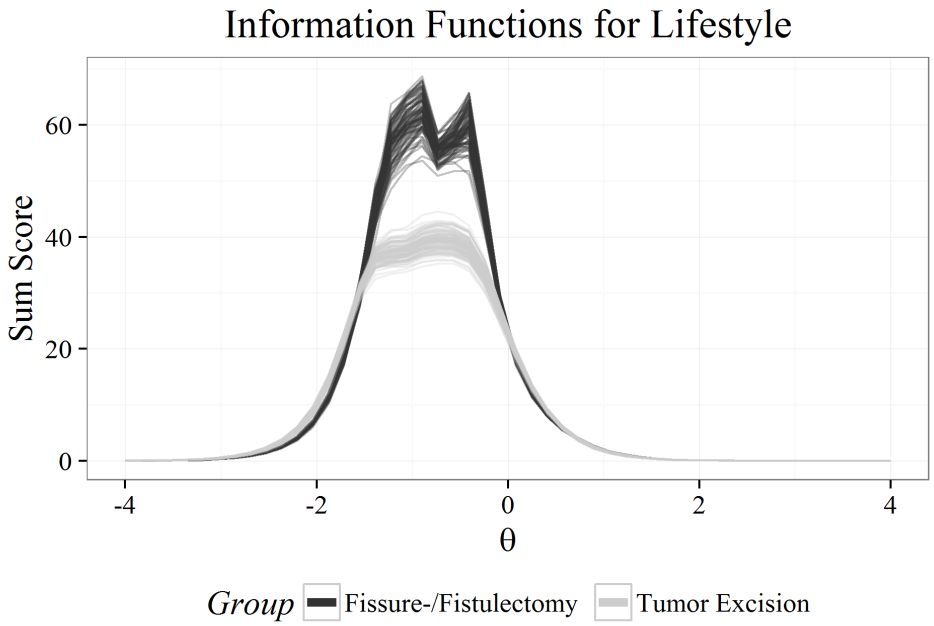
Very little DTF was detected for patients undergoing the two colorectal surgical procedures. Expected sum scores, on average, differed by less than 1 point across all levels of θ for all domains. See Table 4.9. However, since the embarrassment domain has a small range of possible scores, DTF based on surgical procedure resulted in a relative difference of 1.6% to 1.7%. In addition, this was the only domain in which the estimated DTF was positive, indicating slightly higher scores for those awaiting a fissurectomy or fistulectomy relative to tumor excision. The difference was more pronounced for lower ranges of θ, with this statistic rising to about 3.2% for $-6 < \theta < 0$. Again, this difference may disappear with the addition of items to the embarrassment domain.

Table 4.9: DTF Statistics, Fissurectomy/fistulectomy vs. Tumor Excision

	sDTF	% sDTF	uDTF	%uDTF
Lifestyle	-0.12	-0.31	0.14	0.34
Coping	-0.04	-0.13	0.06	0.17
Depression	-0.04	-0.13	0.15	0.47
Embarrassment	0.20	1.67	0.21	1.75

The information functions for the four domains did not differ meaningfully between patients prior to the two surgical procedures. All information functions between these two groups overlapped, except the lifestyle domain. The graph in Figure 4.5 indicates that lifestyle was measured with less accuracy for those undergoing tumor excision relative to fissurectomy or fistulectomy. However, this would only amount to a difference of $\frac{1}{40} - \frac{1}{65} \approx 0.01$ in the standard error of estimation at the widest gap between the functions, near $\theta = -1$.

Figure 4.5: Differential Information Functions by Procedure for Lifestyle



Chapter 5: Conclusion

The FIQL was developed to measure the impact of fecal incontinence on quality of life. It was designed as a condition-specific instrument sensitive to changes within individuals over time, and has been used as the preferred instrument for measuring these effects in a variety of studies examining treatment for FI. It has been translated into at least 10 languages. Previous studies have demonstrated the FIQL's reliability and validity using broad summary statistics such as Cronbach's α or Pearson correlation coefficients, but have not extensively evaluated its psychometric properties using item response theory or differential item and test functioning.

5.1 Study Strengths and Limitations

This thesis is the first study to examine the FIQL in English in its original format, and augments the efforts of previous studies by applying EFA, IRT, and DIF analysis in addition to CTT to produce a more comprehensive evaluation of the FIQL. Furthermore, this thesis recruited participants prospectively from a well-defined sampling frame of individuals with FI registered to an elective colorectal surgery waitlist, which limits the potential for selection bias, and enhances generalizability of the results to English-speaking, Canadian adults awaiting elective colorectal surgery. This sampling procedure also resulted in a roughly equal representation of both men and women – an important feature of this sample compared to other studies of this instrument in which men were underrepresented.

However, this study also has several limitations. Because this study was limited to patients registered to an elective surgery waitlist, the results may not be generalizable to patients undergoing medical management for FI. And while a sample size of 236 is large compared to previous studies, this is still less than simulation studies consider ideal for fitting complex IRT

models, such as the graded response model (61). It is possible a larger sample size would produce better model fit statistics, more precise parameter estimates, and be better able to identify LID and DIF. Efforts were made to minimize bias in participant recruitment and administration, however it is possible some bias remains.

5.2 Strengths of the FIQL

One of the strengths of the FIQL is its high reliability as measured by Cronbach's α, as demonstrated in this and previous studies. Furthermore, this study did not find evidence of differential test or item functioning based on gender or expected surgical procedure that would cause concern. The lone exception was item 2g ("It is important to plan my schedule (daily activities) around my bowel pattern"), which demonstrated higher difficulty for women. Since this did not affect expected test score, the item does not necessarily need to be altered or removed.

5.3 Weaknesses of the FIQL

This thesis identified several weaknesses in the FIQL and can be used to suggest improvements.

First, more research should be performed to identify what domains of quality of life are impacted by FI, along with their relative importance and inter-correlations. This research should be performed with input from patients as well as physicians and other experts. The large number of patients on the surgical waitlist who indicated nearly perfect quality of life on the FIQL suggests that there may be aspects of FI-related quality of life they wish to improve through surgery that is not being measured by the FIQL.

As these domains become better established, additional items can be developed with the help of patients. This will improve the reliability and validity of each domain and allow for easier evaluation in the future. Items that demonstrate higher difficulty thresholds should be included, to improve the instrument's accuracy at measuring individuals with above-average quality of life. This is important because it will allow for accurate measurement of improvement in quality of life following a treatment or over time. Furthermore, additional items will make it easier to identify the four domains (lifestyle, coping, depression, and embarrassment) in exploratory or confirmatory factor analysis, and improve both the construct and content validity of the instrument.

Item responses should also be made consistent (frequency rather than agreement) as they have done in several translations, to minimize the effect of the questionnaire format on response patterns. This will reduce the potential for LID, and will also make it easier to identify the four domains in factor analysis.

5.4 Evaluation of FIQL Domains

The results of this study can be used to make specific recommendations regarding the use and improvement of each of the four domains.

5.4.1 Lifestyle

The lifestyle domain performed well overall with high estimates for discrimination parameters, high reliability for individuals between -3 and 1 standard deviations of the mean, and minimal differential test functioning. This domain can be used as-is or with minimal revisions, such as removing item 2c which demonstrates limited separation between categories, and

removing item 3l (“I avoid traveling by train or plane”) since it demonstrates local item dependence with 3m (“I avoid going out to eat”), and is nearly identical to item 2h (“I avoid traveling”).

5.4.2 Depression

While the depression domain demonstrated high reliability and precision across the widest range of θ, no evidence of local item dependence, and limited differential item and test functioning, it also had the lowest total variance explained by a single factor (64%), which indicates that some of the items may be measuring something other than depression/self-perception. Furthermore, 5 of the 7 items that compose this domain were only able to distinguish between 3 response categories based on θ, which resulted in a dip in information near $\theta = -1$ for this domain. These issues could be resolved by simplifying response categories for this domain, or altering or replacing these 5 items to be better able to differentiate between all four response levels.

5.4.3 Coping

Analysis of the coping domain also revealed good overall characteristics in terms of reliability and information for θ between -3 and 1. However, several revisions should be made before this domain can be used with confidence. Item 3h (“I have sex less often than I would like to”) should be removed from this domain, since its ICC failed to distinguish between more than two response categories and is a sensitive subject with high rates of item non-response. In addition, this domain had the largest number of items demonstrating local item dependence, which should be addressed.

Yen provides a variety of reasons why local item dependence may occur on a given instrument, including external assistance in administration, speededness, fatigue, practice, item response format, passage dependence, item chaining, or content, knowledge and abilities (81). For the FIQL, it seems most likely that the LID observed in this study is due to some combination of respondent fatigue (since LID occurs most frequently later in the instrument), practice (since many items have very similar wording), and response format (items are grouped into large matrices).

To improve the coping domain, items from this domain should be more evenly distributed throughout the instrument rather than grouped near the end. In addition, item 3c ("I worry about bowel accidents") can be removed since it demonstrates LID with several other items in this domain.

5.4.4 Embarrassment

While the embarrassment domain was free from LID, it had the lowest reliability and information functions, contained no items that demonstrated good separation of response categories, and demonstrated the potential for differential test functioning at low values of θ. This domain should undergo significant revision before use.

Embarrassment should be measured by more than three items to increase precision and better define an underlying factor. Item 2l ("I leak stool without even knowing it") should be removed from this domain, since its ICC demonstrated minimal response separation, and removing this item will improve the domain's reliability as measured by Cronbach's α. The item also lacks face validity for this domain. Item 3a also did not appear to effectively measure this domain based on its ICC. Items that demonstrate good separation between response categories

and cover a wide range of difficulty thresholds should be developed and added to this domain. This domain should undergo revision before use.

These recommendations are summarized in Table 5.1.

Table 5.1 Recommended Improvements to the FIQL

Recommendations
Overall
• Use patient-centered input to develop higher difficulty items
• Use frequency rather than agreement for all items in questions 2 and 3
• Consider reducing the number of response categories
• Scoring should be based on sum, rather than mean, scores
Lifestyle
• Remove items 2c & 3l
Depression
• Consider reduced response categories for items 3d, 3f, 3g, 3i, & 3k
Coping
• Remove items 3c & 3h
• Distribute items evenly throughout instrument
Embarrassment
• Create new items for this domain
• Remove item 2l

5.5 Discussion

This study has demonstrated the importance of thorough evaluation of patient reported outcome measures. While the FIQL has some positive qualities, including excellent reliability, and a robust lifestyle domain, the instrument's overall validity, difficulty, and interpretability remain limited. The FIQL can continue to be used and interpreted with caution as a supplementary PROM. However, more development is needed before it can be used as a rigorous, comprehensive, and stand-alone measure of the impact of FI on quality of life.

More research is needed to confirm the findings of this study, and to suggest and test alterations to the instrument. In addition, additional research can address aspects of the FIQL not assessed in this study, such as optimal weighting schemes for item responses, how to produce and interpret a single summary score, or values for minimally important change.

Improvements to the FIQL will allow for valid comparisons of the impact of different treatments for fecal incontinence on patients' quality of life, as well as accurate epidemiological measurement of the burden of FI in different populations. This has wide-ranging implications for the development and evaluation of treatments, understanding the interactions between health status and psychosocial adjustment, and for healthcare priority setting and planning.

References

1. Norton NJ. The perspective of the patient. Gastroenterology. 2004;126(1):S175–9.
2. Rockwood TH, Church JM, Fleshman JW, Kane RL, Mavrantonis C, Thorson a G, et al. Fecal Incontinence Quality of Life Scale: quality of life instrument for patients with fecal incontinence. Dis Colon Rectum. 2000;43(1):9-16-17.
3. Crowell MD, Schettler VA, Lacy BE, Lunsford TN, Harris LA, DiBaise JK, et al. Impact of anal incontinence on psychosocial function and health-related quality of life. Dig Dis Sci. 2007;52(7):1627–31.
4. Bartlett L, Nowak M, Ho Y-H. Impact of fecal incontinence on quality of life. World J Gastroenterol. 2009;15(26):3276–82.
5. Bharucha AE, Dunivan G, Goode PS, Lukacz ES, Markland AD, Matthews CA, et al. Epidemiology, Pathophysiology, and Classification of Fecal Incontinence: State of the Science Summary for the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Workshop. Am J Gastroenterol. Nature Publishing Group; 2015;110(1):127–36.
6. Ilnyckyj A. Prevalence of idiopathic fecal incontinence in a community-based sample. Can J Gastroenterol. 2010;24(4):251–4.
7. Alnaif B, Drutz HP. The prevalence of urinary and fecal incontinence in Canadian secondary school teenage girls: Questionnaire study and review of the literature. Int Urogynecol J Pelvic Floor Dysfunct. 2001;12(2):134–8.
8. AlAmeel T, Andrew MK, MacKnight C. The association of fecal incontinence with institutionalization and mortality in older adults. Am J Gastroenterol. Nature Publishing Group; 2010;105(8):1830–4.

9. Tin RYT, Schulz J, Gunn B, Flood C, Rosychuk RJ. The prevalence of anal incontinence in post-partum women following obstetrical anal sphincter injury. Int Urogynecol J Pelvic Floor Dysfunct. 2010;21(8):927–32.

10. Rockwood TH. Incontinence severity and QOL scales for fecal incontinence. Gastroenterology. 2004;126(1 Suppl 1):S106–13.

11. Ware Jr. JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I . Conceptual Framework and Item Selection. Med Care. 1992;30(6):473–83.

12. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health. 1980;34:281–6.

13. The EuroQol Group. EuroQol: A new facility for the measurement of health-related quality of life. Health Policy (New York). 1990;16(3):199–208.

14. Kelly A, Rush J, Shafonsky E, Hayashi A, Votova K, Hall C, et al. Detecting short-term change and variation in health-related quality of life: within- and between-person factor structure of the SF-36 health survey. Health Qual Life Outcomes. 2015;13(199):1–12.

15. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories - IA and -II in Psychiatric Outpatients. J Pers Assess. 1996;67(3):588–97.

16. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand. 1983;67(6):361–70.

17. Carver CS, Scheier MF, Weintraub JK. Assessing coping strategies: A theoretically based approach. J Pers Soc Psychol. 1989;56:267–83.

18. Eypasch E, Williams JI, Wood-Dauphinee S, Ure BM, Schmulling C, Neugebauer E, et al. Gastrointestinal Quality of Life Index: development, validation and application of a new

instrument. Br J Surg. 1995;82:216–22.

19. Guyatt G, Mitchell A, Irvine EJ, Singer J, Williams N, Goodacre R, et al. A new measure of health status for clinical trials in inflammatory bowel disease. Gastroenterology. 1989;96(3):804–10.

20. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: A new era in clinical research. Perspect Clin Res. 2011;2(4):137–44.

21. Messick S. Validity of Test Interpretation and Use. ETS Res Rep Ser. 1990;(1):1487–95.

22. Raykov T, Marcoulides GA. Introduction to Psychometric Theory. Routledge; 2011. 348 p.

23. Hubley AM, Zumbo BD. A dialectic on validity: Where we have been and where we are going. J Gen Psychol. 1996;123(3):207–15.

24. Rullier E, Zerbib F, Marrel A, Amouretti M, Lehur P-A. Validation of the French version of the Fecal Quality of Life (FIQL) Scale. Gastroenterol Clin Biol. 2004;28(1):562–8.

25. Yusuf S, Jorge J, Habr-Gama A, Kiss D, Rodrigues J. Avaliação da qualidade de vida na incontinência anal: validação do questionário FIQL (Fecal Incontinence Quality of Life). Arq Gastroenterol. 2004;41(3).

26. Altomare D, Rinaldi M, Giardiello G, Donelli A, Petrolino M, Villani R, et al. Italian translation and prospective validation of fecal incontinence quality of life (FIQL) index. Chirogia Ital. 2005;57(2):153–8.

27. Minguez M, Garrigues V, Soria MJ, Andreu M, Mearin F, Clave P. Adaptation to Spanish language and validation of the fecal incontinence quality of life scale. Dis Colon Rectum. 2006;49:490–9.

28. Dedeli O, Fadiloglu C, Bor S. Validity and reliability of a Turkish version of the Fecal

Incontinence Quality of Life Scale. J wound ostomy Cont Nurs. 2009;36(5):532–8.

29. Hashimoto H, Shiokawa H, Funahashi K, Saito N, Sawada T, Shirouzu K, et al. Development and validation of a modified fecal incontinence quality of life scale for Japanese patients after intersphincteric resection for very low rectal cancer. J Gastroenterol. 2010;45:928–35.

30. Tsunoda A, Yamada K, Kano N, Takano M. Translation and validation of the Japanese version of the fecal incontinence quality of life scale. Surg Today. 2012;43:1103–8.

31. Ogata H, Mimura T, Hanazaki K. Validation study of the Japanese version of the Faecal Incontinence Quality of Life Scale. Colorectal Dis. 2012;14(2):194–9.

32. Dehli T, Martinussen M, Mevik K, Stordahl A, Sahlin Y, Linsetmo R, et al. Translation and validation of the Norwegian version of the fecal incontinence quality-of-life scale. Scand J Surg. 2011;100(3):190–5.

33. Ahnis A, Holzhausen M, Rockwood TH, Rosemeier HP. FLQAI - A questionnaire on quality of life in fecal incontinence: German translation and validation of Rockwood et al's (2000) Fecal incontinence quality of life scale (FIQLS). Z Gastroenterol. 2012;50(7):661–9.

34. Hoen LA 't, Utomo E, Schouten WR, Block BFM, Korfage IJ. The Fecal Incontinence Quality of Life Scale (FIQL) and Fecal Incontinence Severity Index (FISI): Validation of the Dutch Versions. Neurourol Urodyn. 2016;1–6.

35. Bols EMJ, Hendriks HJM, Berghmans LCM, Baeten CGMI, De Bie RA. Responsiveness and interpretability of incontinence severity scores and FIQL in patients with fecal incontinence: A secondary analysis from a randomized controlled trial. Int Urogynecol J Pelvic Floor Dysfunct. 2013;24(3):469–78.

36. Mak TWC, Leung WW, Ngo DKY, Lee JFY, Hon SSF, Ng SSM. Translation and validation of the traditional Chinese version of the faecal incontinence quality of life scale. Int J Colorectal Dis. 2016;31:445–50.

37. Bordeianou L, Rockwood T, Baxter N, Lowry A, Mellgren A, Parker S. Does incontinence severity correlate with quality of life? Prospective analysis of 502 consecutive patients. Color Dis. 2008;10(3):273–9.

38. Devesa JM, Vicente R, Abraira V. Visual analogue scales for grading faecal incontinence and quality of life: Their relationship with the Jorge-Wexner score and Rockwood scale. Tech Coloproctol. 2013;17:67–71.

39. de la Portilla F, Calero-Lillo A, Jiménez-Rodríguez RM, Reyes ML, Segovia-Gonzalez M, Maestre MV, et al. Validation of a new scoring system : Rapid assessment faecal incontinence score. 2015;7(9):203–7.

40. Christoforidis D, Bordeianou L, Rockwood TH, Lowry AC, Parker S, Mellgren AF. Faecal incontinence in men. Color Dis. 2011;13(8):906–13.

41. Mundet L, Ribas Y, Arco S, Clavé P. Quality of life differences in female and male patients with fecal incontinence. J Neurogastroenterol Motil. 2016;22(1):94–101.

42. Leite FR, de Lima MJR, Lacerda-Filho A. Early Functional Results of Biofeedback and Its Impact on Quality of Life of Patients With. Arq Gastroenterol. 2013;(3):163–9.

43. Bartlett L, Sloots K, Nowak M, Ho Y-H. Biofeedback therapy for symptoms of bowel dysfunction following surgery for colorectal cancer. Tech Coloproctol. 2011;15(3):319–26.

44. Altman D, Hjern F, Zetterström J. Transanal submucosal polyacrylamide gel injection treatment of anal incontinence: a randomized controlled trial. Acta Obstet Gynecol Scand.

2016;95(5):528–33.

45. Franklin H, Barrett AC, Wolf R. Identifying factors associated with clinical success in patients treated with NASHA ® / Dx injection for fecal incontinence. Clin Exp Gastroenterol. 2016;9:41–7.

46. Patton V, Abraham E, Lubowski DZ. Sacral nerve stimulation for faecal incontinence: medium-term follow-up from a single institution. ANZ J Surg. 2016;1–5.

47. Duchalais E, Meurette G, Perrot B, Wyart V, Kubis C, Lehur P-A. Exhausted implanted pulse generator in sacral nerve stimulation for faecal incontinence: What next in daily practice for patients? Int J Colorectal Dis. 2015;31:439–44.

48. Verseveld M, Barendse RM, Gosselink MP, Verhoef C, de Graaf EJR, Doornebosch PG. Transanal minimally invasive surgery: impact on quality of life and functional outcome. Surg Endosc. 2016;30(3):1184–7.

49. Park JG, Lee MR, Lim SB, Hong CW, Yoon SN, Kang SB, et al. Colonic J-pouch anal anastomosis after ultralow anterior resection with upper sphincter excision for low-lying rectal cancer. World J Gastroenterol. 2005;11(17):2570–3.

50. Sutherland JM, Crump RT, Chan A, Liu G, Yue E, Bair M. Health of patients on the waiting list: Opportunity to improve health in Canada? Health Policy (New York). 2016;120(7):749–57.

51. Rockwood TH, Church JM, Fleshman JW, Kane RL, Mavrantonis C, Thorson a G, et al. Patient and surgeon ranking of the severity of symptoms associated with fecal incontinence: the fecal incontinence severity index. Dis Colon Rectum. 1999;42(12):1525–32.

52. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal

patient reported outcomes. Stat Methods Med Res. 2013;23(5):440–59.

53. Mustillo S, Kwon S. Auxiliary Variables in Multiple Imputation When Data Are Missing Not at Random. J Math Sociol [Internet]. 2015 Apr 3;39(2):73–91. Available from: http://www.tandfonline.com/doi/full/10.1080/0022250X.2013.877898

54. Johnson DR, Young R. Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. J Marriage Fam. 2011;73(5):926–45.

55. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8(3):206–13.

56. Morera OR, Stokes SM. Coefficient α as a measure of test score reliability: Review of 3 popular misconceptions. Am J Public Health. 2016;106(3):458–61.

57. Kline P. An easy guide to factor analysis. London: Routledge; 2002.

58. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods. 1999;4(3):272–99.

59. Schmitt TA, Sass DA. Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. Educ Psychol Meas. 2011;7(1):95–113.

60. Baxter A, Arnold T. The factor procedure: Simplicity functions for rotations. SAS/STAT(R) 9.3 User's Guide. 2011.

61. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Front Psychol. 2016;7.

62. Hambleton RK, Swaminathan H. Item response theory: Principles and applications. 1985. 332 p.

63. Hospers JMB, Smits N, Smits C, Stam M, Terwee CB, Kramer SE. Reevaluation of the

Amsterdam Inventory for Auditory Disability and Handicap Using Item Response Theory. J Speech, Lang Hear Res. 2016;59:373–83.

64. Cai L, Hansen M. Limited-information goodness-of-fit testing of hierarchical item factor models. Br J Math Stat Psychol. 2013;66(2):245–76.

65. Kang T, Chen TT. Performance of the Generalized S-X^2 Item Fit Index for Polytomous IRT Models. J Educ Meas. 2008;45(4):391–406.

66. Chen W-H, Thissen D. Local dependence indexes for item pairs using item response theory. J Educ Behav Stat. 1997;22(3):265–89.

67. Yau DT, Wong MC, Lam K, McGrath C. Evaluation of psychometric properties and differential item functioning of 8-item Child Perceptions Questionnaires using item response theory. BMC Public Health. 2015;15:1–10.

68. Mokkink LB, Knol DL, van der Linden FH, Sonder JM, D'hooghe M, Uitdehaag BMJ. The Arm Function in Multiple Sclerosis Quesionnaire (AMSQ): development and validation of a new tool using IRT methods. Disabil Rehabil. 2015;37(26):2445–51.

69. Chalmers R, Counsell A, Flora D. It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. Educ Psychol Meas. 2016;76(1):114–40.

70. Magis D, Beland S, Tuerlinkx F, De Boeck P. A general framework and an R package for the detection of dichotomous differential item functioning. Behav Res Methods. 2010;42(3):847–62.

71. Bolt D. Limited and full-information IRT estimation. In: Maydeu-Olivares A, McArdie J, editors. Contemporary Psychometrics. Hillsdale, NJ: Erlbaum; 2005. p. 2771.

72. Woods C. Empirical selection of anchors for tests of differential item functioning. Appl

Psychol Meas. 2009;33(1):42–57.

73. Lord F. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum; 1980. 271 p.

74. Cohen A, Kim S, Baker F. Detection of differential item functioning in the graded response model. Appl Psychol Meas. 1993;17(4):335–50.

75. Kopf J, Zeileis A, Strobl C. Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. Educ Psychol Meas. 2015;75(1):22–56.

76. R Core Team. R: A language environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org/

77. Chalmers RP. mirt: A multidimensional item response theory package for the R environment. J Stat Softw. 2012;48(6):1–29.

78. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2009.

79. Maydeu-Olivares A, Cai L, Hernández A. Comparing the Fit of Item Response Theory and Factor Analysis Models. Struct Equ Model A Multidiscip J. 2011;18(3):333–56.

80. Baker FB. The Basics of Item Response Theory. Second. Boston C, Rudner L, editors. ERIC Clearinghouse on Assessment and Evaluation; 2001. 172 p.

81. Yen WM. Scaling performance assessments: Strategies for managing local item dependence. J Educ Meas. 1993;30(3):187–213.

Appendices

Appendix A Scoring Algorithm for the FIQL

The score for each scale is the mean response to all items assigned to that scale. The original FIQL questionnaire had the option of responding "NA" to the items in questions 2 and 3, which was removed in this analysis. See section 3.1.1 for a detailed explanation. Items are assigned to the following domains:

Domain 1: Lifestyle: 2a, 2b, 2c, 2d, 2e, 2g, 2h, 3b, 3l, 3m

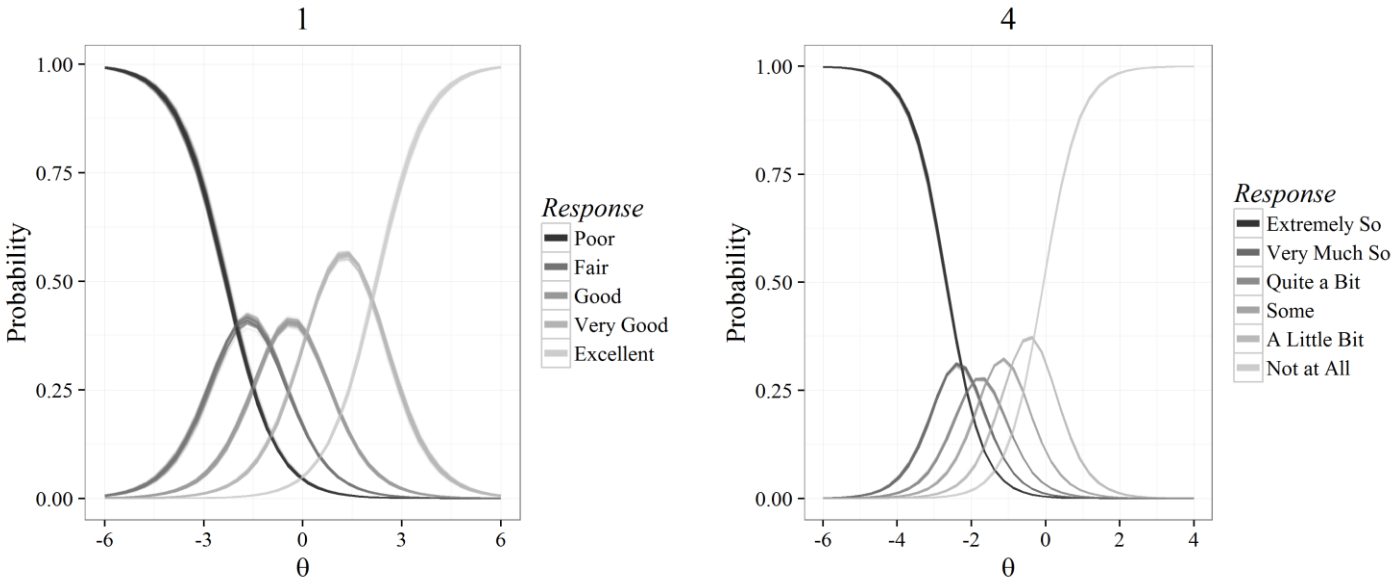
Domain 2: Coping/Behavior: 2f, 2i, 2j, 2k, 2m, 3c*, 3h, 3j, 3n

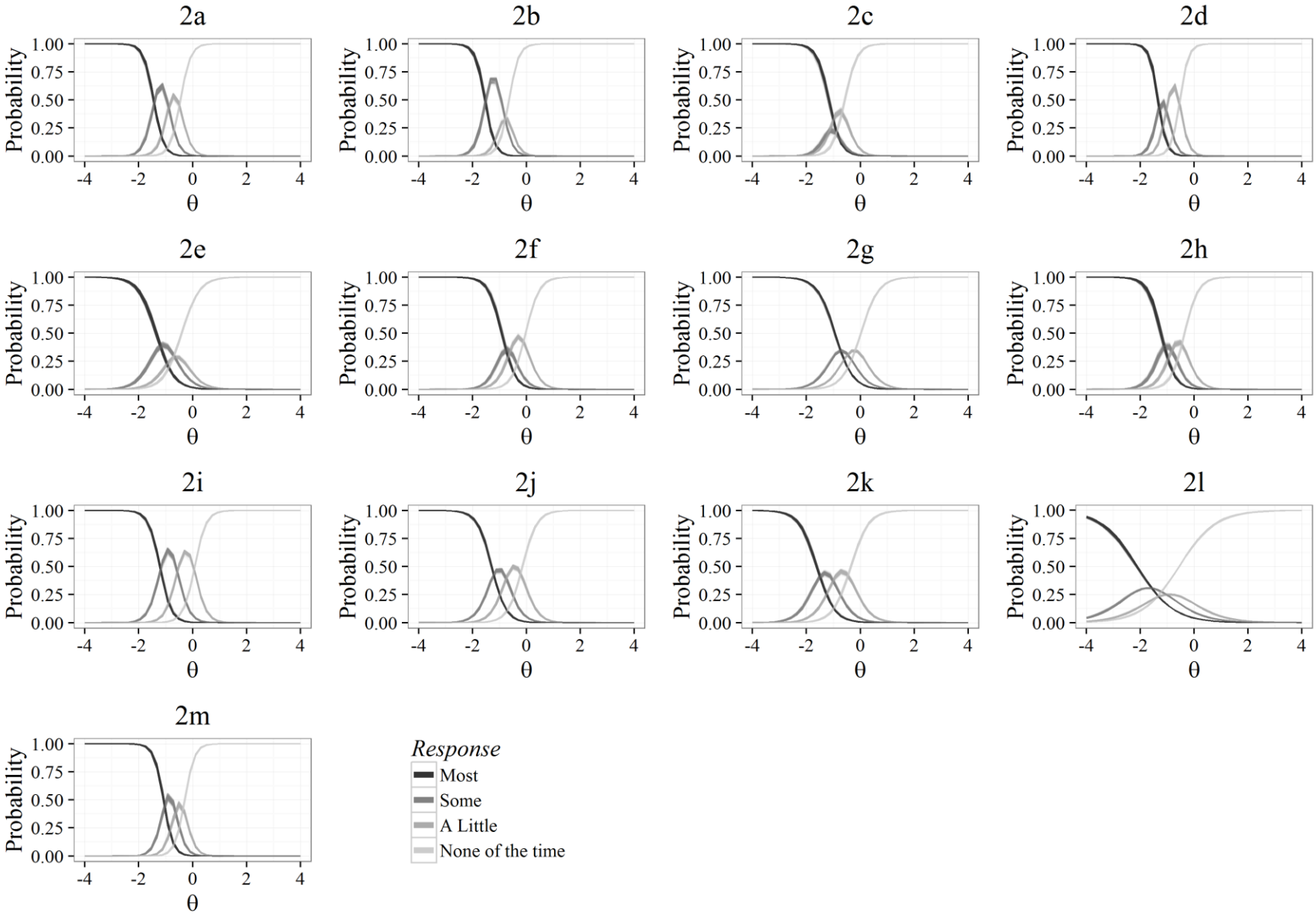
Domain 3: Depression/Self Perception: 1, 3d, 3f, 3g, 3i, 3k, 4 (note: 1 is reversed before analysis)

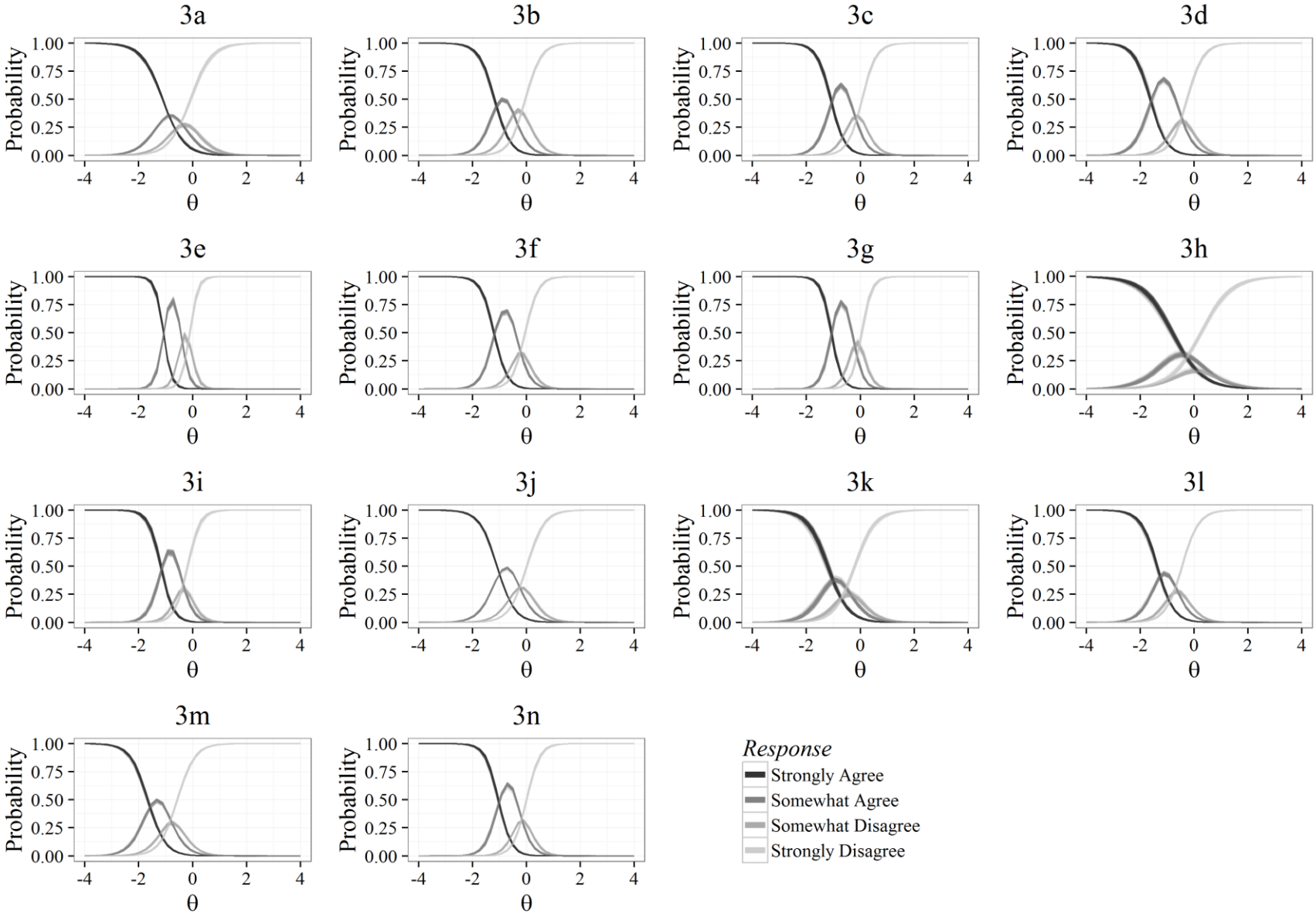
Domain 4: Embarrassment: 2l, 3a, 3e

*The original scoring algorithm contained an error in which 3d was counted on both domain 2 and domain 3. It has been changed to 3c on domain 2.

Appendix B Item Characteristic Curves




2a
2b
2c
2d
2e
2f
2g
2h
2i
2j
2k
2l
2m
Probability
1.00
0.75
0.50
0.25
0.00
-4
-2
0
2
4
θ
Response
Most
Some
A Little
None of the time


3a
3b
3c
3d
3e
3f
3g
3h
3i
3j
3k
3l
3m
3n
Probability
1.00
0.75
0.50
0.25
0.00
-4
-2
0
2
4
θ
Response
Strongly Agree
Somewhat Agree
Somewhat Disagree
Strongly Disagree