EVALUATION OF THE INTERNATIONAL CONSULTATION ON INCONTINENCE QUESTIONNAIRE SHORT-FORM: EVIDENCE FROM A SURGICAL POPULATION

by

Zuzanna Alicja Kurzawa

B.A., Queen's University, 2014

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)

November 2017

Abstract

The International Consultation on Incontinence Questionnaire Short-Form (ICIQ-UI-SF) is a four-item patient-reported outcome (PRO) measure. Its intended use is screening for incontinence, assessing impact of incontinence on quality of life, and facilitating patient-clinician discussions. Evaluations of this instrument to date have relied on a simple set of analytical tools—limiting user's confidence of the instrument's validity and reliability. The purpose of this thesis was to conduct a comprehensive evaluation of the ICIQ-UI-SF.

The analyses were conducted on 177 completed ICIQ-UI-SF instruments by men with chronic urinary incontinence waitlisted for urological surgery for treatment of their condition. This comprehensive evaluation included application of the following methods: confirmatory factor analysis (CFA), principal component analysis, measures of reliability (classical test theory (CTT) and McDonald's coefficient), item response theory (IRT), and differential item functioning (DIF). A supplemental investigation examined previously constructed ICIQ-UI-SF severity categories. Specific goals included assessing: instrument characteristics (dimensionality, ceiling effects), reliability, performance of individual items, whether socioeconomic status influences patients' ICIQ-UI-SF scores, and concordance with other commonly collected PROs (EQ-5D-3L, Visual Analogue Scale).

Responses to all items were left skewed and ceiling effects were identified. Model fit could not be assessed through the CFA, however the factor loadings of items one and two differed significantly ($p < 0.0002$) from item three indicating possible multidimensionality. The PCA contrastingly provided some, albeit limited evidence that the ICIQ-UI-SF is unidimensional. Reliability was low/moderate as measured by Cronbach's alpha (0.63) and McDonald's coefficient (0.65). The IRT revealed the third item's reliability may be improved by

collapsing response levels. The instrument does not discriminate between individuals with high incontinence burden. There was no DIF by socioeconomic status. Supplemental investigations demonstrated the ICIQ-UI-SF discriminates between surgical patients with mild/moderate incontinence versus severe/very severe—but not beyond these dichotomies.

Directly comparing ICIQ-UI-SF scores of urological surgery patients with high incontinence burden is not recommended. If unchanged, the ICIQ-UI-SF can be used as a complement to other data, such as reporting aggregated surgical outcomes, or as a starting point for patient-clinician discussions when applied to a surgical population. For application to surgical triage, this analysis recommends amendments to the ICIQ-UI-SF.

Lay Summary

The International Consultation on Incontinence Questionnaire Short-Form (ICIQ-UI-SF) is a questionnaire used to measure the symptoms, health status, treatment outcomes, and quality of life associated with urinary incontinence. Responses to the ICIQ-UI-SF are summed to produce a total score; a higher score implies higher incontinence burden. Despite its widespread use, evaluations of this questionnaire are limited. In particular, it is unclear whether the ICIQ-UI-SF is appropriate to administer to urological surgery patients.

This thesis conducts a comprehensive evaluation of the ICIQ-UI-SF on a sample of urological surgery patients in Canada. It was found that if unchanged, the ICIQ-UI-SF can be used as a complement to other data or as a starting point for patient-clinician discussions. Directly comparing ICIQ-UI-SF scores of urological surgery patients with high incontinence burden is however not recommended. For application to surgical triage, this study recommends changes to the ICIQ-UI-SF.

Preface

This thesis is based on analysis of patient-reported outcomes data from the Value and Limitations in Hospital Utilization and Expenditures (VALHUE) project, which is a partnership between Vancouver Coastal Health (VCH), Providence Health Care (PHC), and the University of British Columbia (UBC). The data collection was supported through a grant funded by the Canadian Institute for Health Research (CIHR), whose principal investigator was Dr. Jason Sutherland, Centre for Health Services 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 of this manuscript. The research was approved by the University of British Columbia's Behavioural Research Ethics Board (BREB approval number: H12-02062).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CFA:	Confirmatory Factor Analysis
CRF:	Category Response Function
CTT:	Classical Test Theory
DIF:	Differential Item Functioning
EQ-5D-3L:	EuroQol Five-dimensions – Three-Level
GRM:	Graded Response Model
ICIQ-UI-SF:	International Consultation on Incontinence Questionnaire Short-Form
IIF:	Item Information Function
IRT:	Item Response Theory
LI:	Local Independence
MCID:	Minimally Clinically Important Difference
ORMIS:	Operating Room Management Information Systems
PHC:	Providence Health Care
PROs:	Patient-Reported Outcomes
SES:	Socioeconomic Status
TIF:	Test Information Function
UBC:	University of British Columbia
UI:	Urinary Incontinence
VALHUE:	Value and Limitations in Hospital Utilization and Expenditures
VAS:	Visual Analog Scale
VCH:	Vancouver Coastal Health

Acknowledgements

I would like to thank and acknowledge Dr. Jason Sutherland, whose unwavering support, academic insights, and countless revisions made this work possible. I am grateful for the way you invest in your students, and provide us with opportunities to engage in all aspects of the research process.

I would also like to thank and acknowledge Dr. Chris Richardson and Dr. Christopher McLeod, whose feedback and expertise guided this research to be methodologically sound and meaningful in the context of population health.

To my colleagues and friends—Ernest Lai, Kate Redfern, Alex Peterson, and Guiping Liu, thank you for your efforts with the VALHUE project, and for walking with me as I completed my thesis. I would also like to thank Andrée Chartrand for fostering my enthusiasm for research and being an endless source of encouragement.

I would like to thank the faculty, staff, and students at SPPH for equipping me with the tools needed to excel throughout my degree, and providing a warm space for exchanging ideas.

I would like to thank Jeff for supporting, teaching, and challenging me, and most importantly, ensuring I was healthy and happy throughout this whole process.

Lastly, thank you to my family who supported me throughout, and Amelia for your help editing.

Dedication

I would like to dedicate this thesis to my parents. Thank you for always building me up.

Chapter 1: Introduction

Urinary incontinence (UI) is defined as the involuntary or abnormal leakage of urine. It is estimated that three million Canadians experience UI (1). However, discrepancies regarding the prevalence of UI exist due to variations in survey methodology used (2) across studies, individuals not seeking care, and underreporting from perceived social stigma. While prevalence is higher in women aged 80 or younger, men and women are equally affected after age 80 (3). Moreover, out-of-pocket expenses associated with the care and management of UI may pose a financial burden on many, particularly seniors on low fixed incomes (4). On average, seniors living at home are estimated to spend between $1,400 and $2,100 on incontinence supplies including: adult diapers, catheter supplies, homecare services, and treatments not covered by provincial health insurance programs (4).

While not life-threatening per se, both acute and chronic forms of incontinence have significant implications on the quality of life of those affected. Patient-reported outcomes (PROs) such as the International Consultation on Incontinence Questionnaire Short-Form (ICIQ-UI-SF) are one method of measuring the burden associated with incontinence. PROs are standardized instruments (questionnaires) completed by patients that measure symptoms, function, health status, and/or quality of life (5). Although PROs are not routinely collected in Canada, many academic leaders (6), as well as the Canadian Institute for Health Information (CIHI) (7) have a vision for their routine collection, noting the value of PROs in health system assessment, quality improvement, clinical practice (screening, diagnosis, monitoring), and facilitating patient/clinician decision making (8).

The quality of existing PROs however, is variable. Some PROs are developed and evaluated using standard test development approaches, such as classical test theory/item response

theory. Others undergo limited evaluation, or there is little documentation to ascertain whether any evaluation was conducted (9,10). This variability in quality can foster skepticism among clinicians and academics regarding their validity and reliability.

The purpose of this thesis is to conduct a rigorous evaluation of the ICIQ-UI-SF instrument to provide recommendations on use of this instrument in a population of patients with chronic urinary incontinence having surgery for their condition, its limitations, and areas for future inquiry.

Following a review of the literature, data sources, and descriptive statistics, the thesis will comprise a series of analyses, each which investigates various psychometric properties of the ICIQ-UI-SF. These chapters are: confirmatory factor analysis (with principal component analysis), measures of reliability (classical test theory and McDonald's coefficient), item response theory, and a supplemental investigation. By developing recommendations regarding the use of the ICIQ-UI-SF, this research will help improve outcomes measurement in the urological surgery patient population with chronic urinary incontinence, and assist clinicians in their selection of PROs by setting a standard for evaluation.

Chapter 2: Literature Review

2.1 Urinary incontinence: risk factors, clinical assessment, and treatment

Urinary incontinence (UI) can be classified as transient or chronic (11). Common causes of transient incontinence include: delirium, infection, atrophic vaginitis, psychological disorders, pharmaceuticals, excess urine output, restricted mobility, and stool impaction (12). Once the underlying causes are addressed, the incontinence may be reversible (13). Chronic forms of incontinence do not generally resolve without intervention. The three most common types of chronic incontinence are stress, urge, and mixed. Stress incontinence is involuntary leakage due to exertion such as sneezing and coughing, often attributed to urethral hypermobility or sphincter weakness (14). Fifty percent of UI patients in Canada present with stress incontinence (4). Urge incontinence is involuntary urine leakage, paired or preceded by a sudden need to urinate that cannot be deferred. It represents 14% of UI patients (4), and is often caused by detrusor over-activity—involuntary contractions during the filling phase resulting in incomplete bladder emptying (15). On average, stress incontinence is higher in women, urge incontinence is higher in men (16). Mixed incontinence includes aspects of both urge and stress incontinence, and represents 32% of UI patients in Canada (4).

Other less common forms of incontinence include overflow, functional, and true. Overflow incontinence is typically a result of chronic bladder outflow obstruction, and occurs in men with prostatic diseases. Functional incontinence is where individuals are unable to reach the bathroom due to poor mobility or unfamiliar surroundings; it is not related to urinary system dysfunction. True incontinence is where urine leaks continuously (17).

The initial evaluation of a patient presenting with UI includes: medical history, physical examinations, and treatment expectations. Regarding patients' history, factors potentially

affecting UI include: review of storage/voiding, type and severity of incontinence, and degree to which it interferes with daily life. Presence of pain, hematuria, recurrent urinary tract infections, pelvic prolapse in women, previous pelvic radiation therapy, and suspected fistula are generally indications of other complications, and often result in consults with specialists (12). The physical examination includes: abdominal examination, cough stress test, pelvic examination in women and digital rectal examination in men. For both men and women, lifestyle factors associated with elevated probability of UI are smoking or high body mass index. Comorbidities positively associated with UI include Alzheimer's disease, diabetes, hypertension, and obstructive sleep apnea (18–20).

Treatment and management options for UI vary between men and women and the type of incontinence. For women with uncomplicated stress incontinence for example, mid-urethral slings, bulking agents, colposuspension, or compression devices may be used. For men with stress urinary incontinence, urethral bulking agents, male slings, and artificial urinary sphincters are the most common treatments (21). Management includes lifestyle advice (such as reducing caffeine intake, or exercising), pelvic floor muscle training, or scheduling voiding bladder training (12). While treatment efficacy is measured by clinical observations and urodynamic tests, considering the patient's perspective on their quality of life is an important part of care planning and treatment (22), and is what gives rise to the use of PROs in clinical practice.

2.2 Quality of life for patients with UI

Based on existing literature, a number of domains of quality of life have been reported to be affected by UI: travel, social activities/recreation, emotional health, activities of daily living, and sexual function (23). In particular, those who experience shame and anxiety, limit their activities, and avoid social interaction, report lower quality of life (24–27). However, perceived

severity (28), age, type of incontinence, and social support systems often determine the degree to which these domains are affected. For example, patients with mixed and urge incontinence report lower quality of life, than those with stress UI (29). Across studies, age, type and severity of UI, body weight, psychological stress, and help seeking behaviour were consistently reported as statistically significant factors affecting quality of life. Ethnicity, economic status, symptoms, and perceived health status however were inconsistent across settings and studies (24).

2.3 PROs used for UI

PROs for UI serve many functions, from assisting diagnoses, measuring treatment outcomes, to assessing quality of life. For example, Vesia, the Alberta Bladder Centre, collects ICIQ-UI-SF for patients that present with lower urinary tract symptoms, and have used it for randomized controlled trials. In addition, the International Consortium for Health Outcomes Measurement (ICHOM) recommends the collection of various ICIQ modules for patients with overactive bladder. The ICIQ-UI-SF is also widely used in pre-post studies of UI interventions (30–32) and for determining the prevalence of UI across populations (33). See Appendix A for a list of other urinary incontinence PROs (12).

2.4 Overview of PROs

PROs can be generic or condition-specific. Generic PROs assess health broadly, and typically ask about an individual's general health, pain, mobility, or mental health status. The benefit of generic PROs is that they can be used to compare health-related quality of life across sectors and disease categories. Furthermore, patients' health states associated with a number of generic PROs can be equated with utility scores, the foundation for deriving Quality-Adjusted-Life-Years (QALYs) for cost-utility analysis. Examples of popular generic PROs are the Short

Form-36 (SF-36), Health Utilities Index (HUI), and the EuroQol Five-dimensions – Three-Level (EQ-5D-3L); the SF-36 and EQ-5D-3L are the most widely used internationally (34).

Condition-specific (or disease-specific) PROs are designed to assess outcomes unique to particular diseases or sectors of care, including the condition's severity, symptoms, and burdensomeness. While these generally do not produce utility scores[†] they are more sensitive to detecting changes in a patient's quality of life over time, or differences between groups of patients with the same condition (7).

2.4.1 Criteria for selecting PROs

Effectiveness, appropriateness, and feasibility (7) (see Table 2.1) are among factors to consider when selecting PROs. Effectiveness refers to whether evaluations of the PRO have been conducted using standard instrument validation criteria such as validity, reliability, and responsiveness. Validity is the extent to which an instrument measures the construct of interest, and supports the interpretations of scores for a given purpose (35,36). Reliability is a measure of consistency, i.e., a reliable measure would produce similar results across applications or time, if measuring the same construct (36). Responsiveness is the extent to which an instrument detects change in an outcome (35). A more detailed description of these validation criteria can be found in the subsequent sections. Appropriateness includes considering the type of instrument (generic or condition-specific), or ensuring it provides information salient to stakeholders (such as health-utility values). One may consider patient perspectives regarding appropriateness as well (such as careful wording of sensitive questions, clarity on purpose of collection, etc.). Feasibility

[†] While condition-specific instruments seldom produce utility values, in some cases it is possible to convert data from a condition-specific instrument to one that does produce utility values. This data conversion from one instrument to another is called 'mapping'.

considerations include cost (if they are proprietary), time requirements, and mode of administration (electronic or paper surveys) (7).

Table 2.1 Common questions considered when selecting PROs

Effectiveness
• Have psychometric properties (validity, reliability, and responsiveness) been assessed? • Has it been successfully implemented in a similar context?
Appropriateness
• Is a generic or condition-specific instrument more appropriate for the target population? • What languages are available? Does the reading level required match the target population? • What information are the stakeholders interested in? • If informing cost-utility analysis, are health utility values available for the instrument?
Feasibility
• Is the instrument proprietary or open-access? • How is it administered (telephone, paper, online)? • What are the time requirements for completion?

2.5 Development of the ICIQ-UI-SF instrument

The International Consultation on Incontinence Questionnaire Short-Form (ICIQ-UI-SF) was developed in 1998, sponsored by the World Health Organization. The instrument's development was iterative, involving International Consultation on Incontinence experts and 63 urology clinic attendees in the UK, to ensure the questionnaire was easily interpretable and reflected salient areas ('domains') of incontinence such as symptom severity and interference with daily life. This resulted in a developmental version of the instrument (*d*ICIQ), which assessed the following: frequency of leakage, how bothersome leakage is, frequency of protection use, usual amount of leakage, worst amount of leakage, interference with everyday life, social life, sex life, and overall quality of life. Principal factor analysis—an item reduction technique—and the analysis of validity, reliability, and responsiveness were used to devise the resultant instrument (37). Details on the factor analysis were not provided/published.

For a summary of the ICIQ-UI-SF instrument derived from the above process, see Table 2.2**,** and Appendix B. The instrument contains four items (questions) and can be generally completed in a couple of minutes. The first three items are scored and then summed to produce a total score, ranging from 0 to 21. A higher ICIQ-UI-SF score indicates higher frequency, severity, and impact of UI on UI-related quality of life. The first item has 6 response categories, and is scored from 0 to 5. The second item has 4 response categories, scored 0, 2, 4, or 6. The last item has 11 response categories and is scored from 0 to 10. All items have the same directionality, meaning that for each item, a higher score indicates higher symptom burden. The last item is a diagnostic item to assess the perceived cause of incontinence. This item was included upon the request of clinicians and is not used for scoring.

Due to the instrument's brevity, items have low levels of missing data. The instrument has been reportedly used for: screening for incontinence, summarizing perceived causes of UI, and facilitating discussions between patients and clinicians. Incontinence within the questionnaire is defined as "minimum leakage of 'about once a week or less often' in items assessing 'amount of leakage'" (37). Additional modules to the ICIQ-UI-SF exist, including: quality of life (ICIQ-UIqol), sexual matters for males (ICIQ-MLUTSsex) or females (ICIQ-FLUTSsex), or treatment satisfaction (ICIQ-S*) (38).

Table 2.2 Summary of the International Consultation on Incontinence Questionnaire Short-Form

Purpose	• Screening for incontinence • Summary of impact and perceived cause of symptoms • Facilitate patient-clinician discussions
Number of items	• 4
Question items	• Frequency of urinary incontinence • Amount of leakage • Overall impact of urinary incontinence • Self-diagnostic item
Scoring	• 0-21 (higher score, increased severity/burdensomeness)
Completion time	• Few minutes
Languages available	• Afrikaans; Arabic; Australian-English; Brazilian-Portuguese; Bulgarian; Czech; Danish; Dutch; Estonian; Finnish; French; German; Greek; Hungarian; Icelandic; Italian; Japanese; New Zealand-English; Norwegian; Polish; Romanian; Russian; Slovakian; South African-English; Spanish; Swedish; Turkish; Ukrainian; UK-English; US-English
Additional modules	• Quality of life (ICIQ-UIqol) • Sexual matters for males (ICIQ-MLUTSsex) • Sexual matters for females (ICIQ-FLUTSsex) • Treatment satisfaction (ICIQ-S*)

2.5.1 Psychometric analysis of the ICIQ-UI-SF

The International Continence Society's Consultation on Incontinence awarded the ICIQ-UI-SF instrument a 'Grade A' status (39) for assessing symptoms and quality of life, using the standard evaluation criteria (validity, reliability, and responsiveness). The most robust evaluation of the English language ICIQ-UI-SF was conducted in a study by Avery et al. (37), however a number of other studies have examined psychometric properties of translated ICIQ-UI-SF questionnaires. Table 2.3**,** Table 2.4**,** and Table 2.5 summarize the findings of these validation studies; Table 2.6 summarizes each study's sample characteristics.

Table 2.3 Summary of ICIQ-UI-SF validation studies – validity

Study	Content Validity	Construct Validity	Convergence Validity
Avery et al. (2004)	1-2% missing data	***Discriminates by***: Age ($p < 0.001$); Sex ($p < 0.001$); Type of UI ($p < 0.001$)	***Moderate/Strong agreement with:*** Bristol Female Lower Urinary Tract Symptoms (r_s= 0.53-0.86) ***Weak/moderate agreement with:*** ICSmale SF (r_s= 0.24-0.58)
Hashim et al. (2006)	< 1% missing data	***Discriminates by***: Type of UI by sex ($p < 0.0001$); Severity by type UI ($p < 0.0001$)	***Strong agreement with:*** Urodynamic test (r_p= 0.82)
Espuña et al. (2007)	2.59% missing data	-	***Moderately strong agreement with:*** Urodynamic test (r_p= 0.6)
Pereira et al. (2010)	-	***Discriminates by***: Type of UI ($p < 0.0001$); Level of education ($p < 0.0001$); Income ($p < 0.0001$)	***Moderate to strong agreement with:*** King's Health Questionnaire (r_s= 0.44-0.77)

Table 2.4 Summary of ICIQ-UI-SF validation studies – reliability

Study	Stability	Internal Consistency
Avery et al. (2004)	Items 1, 2, 4 κ = 0.68-0.90; Item 3: κ = 0.58	α = 0.95
Hashim et al. (2006)	All items κ = 0.85	α = 0.71
Pereira et al. (2010)	All items κ = 0.72-0.75; r_p= 0.89	α = 0.88

Table 2.5 Summary of ICIQ-UI-SF validation studies – responsiveness

Study	Responsiveness
Avery et al. (2004)	Mean patient scores improved on items 1, 2, 4 ($p < 0.001$) following conservative management and treatment for both males and females.
Espuña et al. (2005)	Post-treatment decrease in scores ($p < 0.0005$).
Hashim et al. (2006)	Decrease in percentage of patients reporting symptoms post-treatment ($p < 0.0001$ for all). Mean patient scores improved following post-treatment, from 12.6 to 6.8 ($p < 0.0001$).
Seckiner et al. (2007)	Differences in post-treatment parameters including first sensations of bladder filling, cystometric capacity, maximum detrusor pressure, and compliance ($p < 0.01$). Mean patient scores improved post-treatment, (13.9 to 9.4).

Table 2.6 Sample characteristics of ICIQ-UI-SF validation studies

Study	N	Sex		Age		Notes
		F	M	Median/ Mean	Range	
Avery et al. (2004)	469	324	145	57.2	23.4-101.3	
Sub-sample						Patients recruited through Bristol clinic, Leicester community, and Bristol community clinic.
Convergent val. (BFLUTS)	118	118	0	57.7	24.4-88.3	
Convergent val. (ICSmaleSF)	27	0	27	58.6	23.6-82.6	
Stability (test-retest)	144	121	23	58.1	24.5-90.9	
Espuña et al. (2005)	71	71				Women with stress UI treated with tension free vaginal tape
Hashim et al. (2006)						
Sub-sample						Patients attending urology clinics at 2 teaching hospitals (one in Egypt, other in Syria) with varying degrees of UI
Content and construct	131	87	44	37.8	18-73	
Stability/internal consist.	102	68	34	37.7	17-73	
Sensitivity	53	35	18	37.2	16-73	
Espuña et al. (2007)	116	116		54	13.99 (SD)	-
Seckiner et al. (2007)	60	42	18	49.8	28-70	Patients with varying degrees of UI referred to the Department of Urology at Zonguldak Karaelmas University, Turkey
Pereira et al. (2010)	123	94	29	53 (med)	16-86	Married (68.3%), working (41.5%), income equal to 4 monthly minimum wages (48%), illiterate (17.9%), UI for > 1 year (83.7%)

2.5.1.1 Validity

Validity concerns whether the construct of interest is indeed being measured, or more formally, whether the construct of interest is the source of item covariation of an instrument (40). Three commonly assessed forms of validity are content, construct, and convergence validity (Table 2.7); these were explored in four ICIQ-UI-SF studies (37,41–43). Content validity refers to how the elements of an instrument represent the construct one is attempting to measure (44). Content validity would be present if an instrument contained a random subset of the universe of applicable items (40). While exhaustive lists of these items are not often available, the assessment of content validity usually involves interviews with subject area experts who can speak to whether salient domains are represented in the instrument. In the development of the ICIQ-UI-SF instrument, clinicians and social scientists concluded the instrument was easily interpretable and covered relevant domains including frequency and amount of leakage, as well as impact on daily life (37). Many of the evaluation studies also reported the levels of missing data (ranging from 1.00% to 2.59% (37,41,42)) in their assessment of content validity as a proxy for the acceptability of items.

Construct validity refers to the relationships between items or instrument scores and other variables/underlying theories—it is the degree to which the instrument 'behaves' the way existing research would suggest. In this regard, the ICIQ-UI-SF has been reported to discriminate between types of incontinence in men and women (37,41,43), and severity (41).

Convergence validity refers to the degree that measures of a construct, which are theoretically related, are in fact related. With no 'gold standard', the relationship between the ICIQ-UI-SF and other UI instruments was assessed using Spearman's rank correlation. Agreement with the Bristol Female Lower Urinary Tract Symptoms (BFLUTS) on items

measuring ‘frequency’ and ‘usual amount’ of leakage ranged from moderate (r_s= 0.53) to strong (r_s= 0.86) (37). Agreement with the ICS*male* short form (ICS*male*SF) assessing perceived cause of incontinence was weak (r_s= 0.24) to moderate (r_s= 0.58) (37). Agreement with King’s Health Questionnaire ranged from weak to moderate (r_s= 0.44-0.77) (43). Agreement with urodynamic tests was moderate (r_s= 0.6) (42).

Table 2.7 Types of validity

Type	Definition
Content	Refers to how the elements of an instrument represent the construct one is attempting to measure. The assessment usually involves interviews with subject area experts.
Construct	Refers to the relationships between items or instrument scores and other variables/underlying theories. For example if based on theory UI is positively related with variables A and B, then an instrument measuring UI should demonstrate the same relationships to measures of A and B.
Convergence	Refers to the degree that measures of a construct, which are theoretically related, are in fact related. Absent gold standard measures, it is often observed by correlating responses from one instrument to another one measuring the same construct.

2.5.1.2 Reliability

The conceptualization and operationalization of reliability varies based on the type of analysis one is conducting, however, in this thesis, reliability is the extent that an instrument’s items are without measurement error (35), thus yielding results in a reproducible and consistent manner (45). Within Classical Test Theory, there are four main measures of reliability: test-retest, parallel forms, inter-rater, and internal consistency (36), are summarized in Table 2.8.

Table 2.8 Approaches to reliability

Type	Definition
Test-retest	Typically computes correlation between two administrations of the same test over a period of time where one would not expect a change.
Parallel forms	Correlation between sets of items measuring the same construct
Inter-rater	Only relevant if raters are involved in the assessment (uncommon for PROs). Typically measured by Cohen's Kappa or intraclass correlation coefficient.
Internal consistency	Assesses whether test tries to measure same general construct—measured based on correlations between different questions of the same test. Typically measured by Cronbach's alpha or McDonald's coefficient.

Test-retest reliability refers to the stability of responses over a given period where one would not expect them to change (36). This was calculated by the percentage agreement between test/retest scores, and weighted Kappa statistics. In the primary study's development of the ICIQ-UI-SF, agreement was 'good' to 'very good' for all items (κ= 0.68-0.90) except 'overall quality of life', which had moderate agreement (κ= 0.58) (37). In other studies, reliability was reported as κ=0.85 (41), and κ= 0.72 to 0.75 (43). The internal consistency refers to the correlation between instrument items. Cronbach's alpha for all studies was moderate to very high, from α = 0.71 (41), α = 0.88 (43), to α = 0.95 (37), all indicating some redundancy and internal consistency.

2.5.1.3 Responsiveness

Responsiveness refers to an instrument's ability to detect changes in an outcome over time (46). A number of studies have reported statistically significant differences between patient's ICIQ-UI-SF pre- and post-intervention score (37,41). For example, women with stress UI who were treated with tension free vaginal tape observed post-treatment differences in their scores ($p < 0.005$) (47). Among a cohort of patients receiving antimuscarinic therapy, their mean scores decreased from 13.9 (±3.7) to 9.4 (±2.9) (48).

2.5.2 Other ICIQ-UI-SF studies and applications

In addition to the literature investigating the instrument's psychometric properties, other studies of the ICIQ-UI-SF were conducted that provide insight into the applicability of the instrument.

2.5.2.1 Mode of administration

One study examined whether the mode of questionnaire administration affects the results. It was found that there is no difference between patients completing the questionnaire alone, or through a physician interview. While only women were included in this study, it is a preliminary insight into the survey's reliability (27).

2.5.2.2 Severity thresholds

One study compared the ICIQ-UI-SF to the Incontinence Severity Index (ISI), a validated instrument that produces severity categorizations (slight, moderate, severe, and very severe) for incontinence; it does not assess quality of life (49). A cross-sectional online study of 1,812 Norwegian women was used. The four severity categories of the ISI were plotted against the ICIQ-UI-SF total scores (with and without the quality of life domain question) and evaluated by Spearman's rank correlation. Strong and statistically significant correlations ($p < 0.01$) were found between the ISI severity categories and the ICIQ-UI-SF scores both with the quality of life question ($r_s = 0.62$) and without ($r_s = 0.71$). The proposed severity categories for the ICIQ-UI-SF were: slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21). While this study's average responder was much younger than the average age of those with incontinence, and there was limited power in the 'very severe' category, this was the first study to try and develop this categorization, which may be useful for observing improvement/decline within patients.

2.5.2.3 Diagnostic item

Many studies, particularly pre/post design studies, use the diagnostic item to classify the study population (Table 2.9). The question asks: "when does urine leak", and respondents tick all that apply. One study found that the combination of stress test and UI had good predictive value and recommends the use of ICIQ-UI-SF for diagnoses in combination with urodynamic tests (50). Another recent study used the ICIQ-UI-SF to determine the prevalence of urinary incontinence across France, and analyzed how the estimates varied based on the survey design and definition of incontinence, and the diagnostic item of the ICIQ-UI-SF was used (33). However, the ICIQ Development group did not provide official guidelines for how to use the diagnostic item.

Table 2.9 Diagnostic item

1	☐	Never – urine does not leak
2	☐	Leaks before you can get to the toilet
☐	☒	Leaks when you cough or sneeze
4	☐	Leaks when you are asleep
☐	☒	Leaks when you are physically active/exercising
6	☐	Leaks when you have finished urinating and are dressed
7	☐	Leaks for no obvious reason
8	☐	Leaks all the time

2.5.2.4 The Minimally Clinically Important Difference (MCID)

The Minimally Clinically Important Difference (MCID) estimates the change in an instrument's score associated with a patient's subjective experience of improvement (51). One study used data from the Trial of Midurethral Slings, and was applied to women with stress urinary incontinence. The MCID was determined by calculating the difference between mean ICIQ-UI-SF scores for individuals with the smallest improvement, and those with no change. It

was found that for surgical patients with stress incontinence, a decrease in 5 points at 12 months, and 4 points at 24 months is considered clinically meaningful.

Chapter 3: Study Purpose and Rationale

The purpose of this study is to evaluate the ICIQ-UI-SF to provide recommendations on the use of this instrument in a surgical population, its limitations, and areas for future inquiry. The instruments' brevity, coupled with its popularity should not dismiss the need for evaluation, for two main reasons. The first reason is a practical one: precisely because it is a popular instrument, it is important that its psychometric properties are commensurate with the way the instrument is used. Facilitating patient-clinician decisions requires a much lower confidence in the ICIQ-UI-SF's validity/reliability than surgical triage for example. The second reason is rooted in the evolving nature of measurement science, and that validation is not a discrete event, but rather an ongoing process, where "one amasses an evidential and consequential basis for the two main functions of test interpretation and test use" (52). Some of the knowledge gaps, which are salient to both interpretation and use, are described below.

3.1 Knowledge gaps and areas for future research

The first gap is the scarcity of information regarding the instrument's dimensionality. Although the primary study (37) did undergo factor analysis and claimed to find an 'underlying factor' for the instrument, foundational publications on this instrument did not report what the underlying factor ('latent variable') is, nor information about the item reduction decisions (37,49). Testing whether one factor underlies the ICIQ-UI-SF, and if not, understanding what factor(s) underlie the instrument, is fundamental to both the interpretation of the psychometric analyses, but also the proper application of the instrument.

The second gap relates to the methods used to assess reliability at the instrument level. To date only Cronbach's alpha and kappa statistics have been used as measures of reliability. However reliability values derived from past studies have ranged dramatically which can signal a

number of issues: the assumptions required for calculating the reliability statistics were violated or the samples upon which the statistics were calculated are qualitatively different across studies.

The third gap relates to item level characteristics, which have not been investigated. Item response theory is one analytical tool that provides an avenue to assess item level characteristics, and is not sample dependent, offering a complement to the existing body of evidence.

The fourth gap relates to differential item functioning (DIF), a potential source of bias. DIF occurs when two groups (e.g. men and women) who are otherwise equal on a trait (e.g. 'incontinence burden') have different probabilities of endorsing an item (such as a score of '10' on the item which asks how much incontinence interferes with their life). The ICIQ-UI-SF has not been assessed for differential item functioning.

The fifth gap is presence of ceiling and/or floor effects. Evaluating items/instruments for ceiling effects is particularly salient in the surgical population, as their burden is much higher and so responses will be concentrated at the extreme. Ceiling effects will have implications on calculating responsiveness (sensitivity to change) and the minimally clinically important difference (53). A number of studies have evaluated the instrument for responsiveness (37,39,41,47,48); however these measures may not be meaningful without exploration of ceiling effects.

The sixth gap is the missing case definitions for types of incontinence based on the diagnostic item. While this does not fall under the psychometric evaluation, to ensure some uniformity in the application of this instrument, creating a taxonomy for types of incontinence may provide value to clinicians.

Finally, psychometric evaluations of this instrument were not conducted on such a clinically homogenous sample, with such high incontinence burden. As such, it is unclear

whether the ICIQ-UI-SF is appropriate in this context. Furthermore, while the instrument is described as intended for screening for incontinence, summarizing impact/perceived cause, *and* patient-clinician discussions, this may not be true for all types of incontinence, interventions, etc.

3.2 Analytical strategy

To address some of the knowledge gaps listed above, this study applied a number of analyses to examine the following:

- Response patterns (distributions, skew, floor and ceiling effects)
- Dimensionality
- Reliability using Classical Test Theory and Item Response Theory perspective
- Differential item functioning for high/low socioeconomic status
- Mapping of ICIQ-UI-SF severity categories to other instruments

The analyses will be organized into the following chapters: descriptive statistics, confirmatory factor analysis (with principal component analysis for comparison), measures of reliability (classical test theory and McDonald's coefficient), item response theory, and a supplemental investigation. Each will have an overview of the method, the application to this study sample, and discussion of the limitations and implications on the ICIQ-UI-SF instrument. Table 3.1 summarizes the analyses included in this study, and their purpose.

Table 3.1 Summary of analyses

Analysis	Purpose
Descriptive statistics	• Understand response pattern (e.g. skew), floor/ceiling effects
Confirmatory factor analysis, Principal component analysis	• Confirm whether ICIQ-UI-SF is unidimensional • Compare result with other method
Cronbach's alpha	• Calculate a measure of reliability
McDonald's coefficient	• Calculate an alternative measure of reliability
Item response theory	• Investigate item level characteristics
Differential item functioning	• Investigate response patterns by low/high SES
Mapping of severity categories	• Preliminary investigation of potential application

Chapter 4: Data

4.1 Data sources

This study uses primary data from the Value and Limitations in Hospital Utilization and Expenditures (VALHUE) project, which is a partnership between Vancouver Coastal Health (VCH), Providence Health Care (PHC), and the University of British Columbia (UBC). VALHUE collects and analyzes PROs from a sample of elective surgery patients in the VCH region, assessing whether health-related quality of life, pain, and mental health status change as patients wait for surgery, and following surgery. Data collection has been ongoing since September 2012. This project was approved by the UBC BREB (approval number: H12-02062).

4.2 Data collection

Patients in this study have consented to surgical treatment of their chronic UI. Recruitment of patients begins with a phone call by Vancouver Coastal Health. Two telephone recruitment attempts are made during regular working hours. Patients that are not successfully contacted are cold mailed a survey invitation. Patients who express willingness to complete PROs are given the option of completing surveys online or by mail. Mailed surveys are sent back to VCH where data entry clerks input the information (54). Two reminder emails are sent to those participating through the web-based system, and reminder calls are made to those who opt for mailed surveys.

Each email or survey package includes details about the study, and a survey package containing generic instruments, the EuroQol EQ-5D-3L with Visual Analogue Scale (VAS) survey, along with a condition-specific instrument. Patients waiting for urological surgery receive the ICIQ-UI-SF condition-specific instrument.

The EQ-5D-3L contains five items, and measures the dimensions of mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (55). Patients score each item on three

levels: no problem, some problems, and severe problems. Utility scores, based on a random sample of Canadians, are available for all of the instrument's possible health states. Utility scores represent preferences for different health states (56). The values range from -0.34 (worse than death) to 1 (perfect health) (57). The EQ-5D-3L also includes a visual analogue scale (VAS) ranging from 0 ("the worst health you can imagine") to 100 ("the best health you can imagine").

4.3 Study sample construction

Patients undergoing procedures for UI were identified by their diagnostic codes reported in the VCH's Operating Room Management Information System (ORMIS). The ORMIS system is used for the scheduling, documentation, and tracking of surgical cases. Patients in this study were waitlisted for two main types of surgeries: insertion/removal of an artificial sphincter, or insertion of urethral sling. The corresponding ORMIS codes for insertion of an artificial sphincter are '37230' or 'UR0093', and insertion of urethral sling are '39108', 'URO0092', or 'URO080'. Those undergoing *removal* of an artificial sphincter ('39109', 'URO044') were excluded as surgery is performed due to infection/complications, rather than treatment for UI. These procedures are typically performed on males with moderate to severe stress urinary incontinence. For patients who completed multiple surveys prior to surgery, only the first survey was used (baseline); most patients who completed multiple baselines were undergoing follow up surgeries. For clinical homogeneity, only those undergoing initial treatment for UI were included, namely, *insertion* of an artificial sphincter or urethral sling.

Demographic variables retained for this analysis included age, sex, and deprivation index. In this study of the ICIQ-UI-SF, only men received the instrument. Women who presented with stress incontinence, on average, were waitlisted for treatment of various pelvic floor disorders, and were administered a separate instrument that assesses symptom severity specific to this

condition, as well as incontinence related questions. The deprivation index is a neighborhood-level indicator of socioeconomic status (SES). It is determined by linking a patient's contact address with a deprivation index that was created at the level of Dissemination Areas in British Columba. It is a composite measure integrating community level information such as highest educational achievement, unemployment, income, and housing. It is presented as quintiles, with the first quintile representing the 20% with highest SES, and the fifth quintile representing the 20% with lowest SES (58). The clinical variable retained was the diagnosis, determined by the ORMIS system.

Chapter 5: Descriptive Statistics

5.1 Demographics

Between September 2012 and August 2016, 196 ICIQ-UI-SF were returned from men waiting for either insertion/removal of an artificial sphincter, or insertion of a urethral sling. The response rate of patients waiting for the indicated surgeries was 64.5%. This reflects those that agreed to participate, as well as those who were cold mailed the instruments. In accordance with past studies, missing data was not an issue in this analysis of ICIQ-UI-SF data (37,41,42). The rate of missing data was 1.0%. Thus, a complete case analysis was used, and it was not expected excluding approximately one percent of patients would bias the findings. Applying the exclusion criteria above, this study was based on 177 patients.

The average age of patients was 68.8, and 31.1% were characterized by being in lowest or second lowest SES quintile (Table 5.1). 8.5% of SES indicators were missing, which could be attributable to: new housing developments not yet assigned an SES indicator, living on reserve, homeless, or out-of-province. Just over half (53.1%) of patients were waiting for insertion of urethral sling.

Table 5.1 Sample characteristics

Age		
Mean		68.86
Standard deviation		8.71
Range		21-87
SES	**N**	**%**
1 (Highest SES)	37	20.90
2	41	23.16
3	29	16.38
4	25	14.12
5 (Lowest SES)	30	16.95
Missing	15	8.47
Surgery Type	**N**	**%**
Urethral sling	94	53.11
Artificial sphincter	83	46.89
UI Severity	**N**	**%**
Slight	1	0.56
Moderate	36	20.34
Severe	85	48.02
Very severe	55	31.07

The average general health utility value was 0.8, and just the VAS was 73.3. Of the 177 patients in this sample, the rate of missing data for the EQ-5D-3L was 2.6%, and 3.4% for the VAS (Table 5.2).

Table 5.2 Responses to PROs

General Health: EQ-5D-3L Utility value	
Mean	0.79
Standard deviation	0.16
Range	0.24-1.00
General Health: VAS	
Mean	73.25
Standard deviation	17.51
Range	20.00-100.00

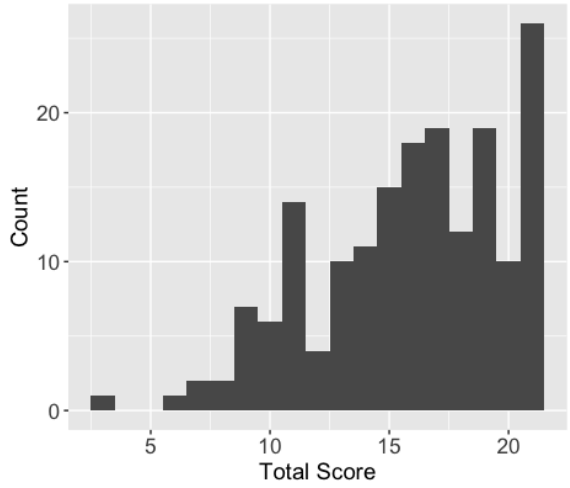
Given some missing responses, the samples for each reported outcome are: EQ-5D-3L (N=173), VAS (N=171)

5.2 ICIQ-UI-SF response summary statistics

The distribution of ICIQ-UI-SF total scores was left skewed. The mean score was 15.9 (SD: 3.9), with a range of 3 to 21 (Figure 5.1). Note that if an individual answered 0 for either of

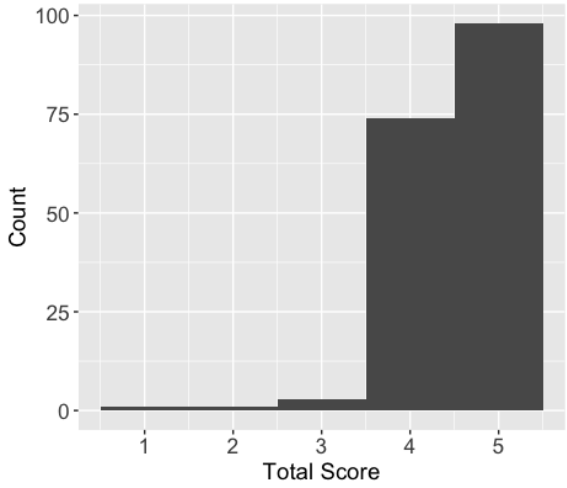
the first two questions, they were not classified incontinent; this did not occur in the sample not unexpectedly since these patients were proceeding with surgical treatment for UI. Distributions of item responses can be found in Figure 5.2, Figure 5.3, and Figure 5.4. Responses to items 1, 2, and 3 were also left skewed although there was variability in responses.

Figure 5.1 Distribution of total ICIQ-UI-SF scores



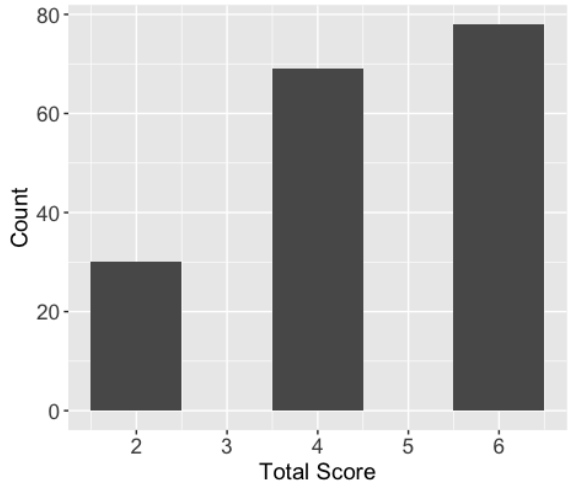
Mean, SD: 15.86, 3.94
Range: 3-21

Figure 5.2 Distribution of item 1 responses: "how often do you leak urine?"



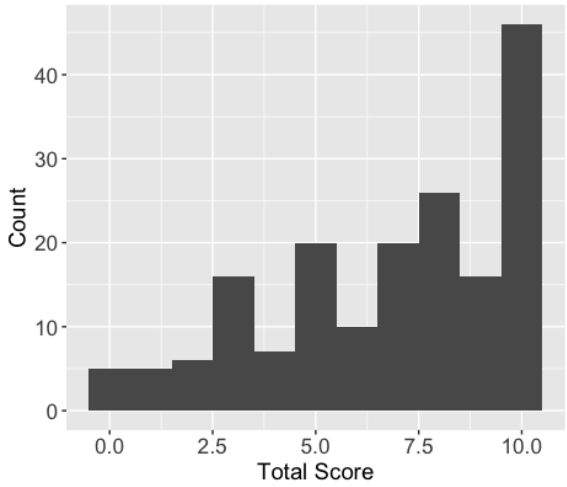
Mean, SD: 4.51, 0.62
Range: 1-5

Figure 5.3 Distribution of item 2 responses: "how much urine do you usually leak?"



Mean, SD: 4.54, 1.47
Range: 2-6

Figure 5.4 Distribution of item 3 responses: "overall, how much does leaking urine interfere with your everyday life?"



Mean, SD: 6.81, 2.90
Range: 0-10

5.3 Floor and ceiling effects

Instruments that are ‘too difficult’ or ‘too easy’ may exhibit floor or ceiling effects, respectively. If the items of an instrument are too difficult, respondents will score near the minimum. The analogue, a ceiling effect, occurs when items are ‘too easy’. If most respondents score near the maximum, there is little variance in responses, and so little information about respondents is revealed. For the ICIQ-UI-SF, if across levels of UI burden, most of the sample scored near the maximum, a ceiling effect would be present.

There are no defined quantitative tools for ascertaining floor or ceiling effects, however visually observing skewedness, or the percentage of extreme score values is a useful indication of the presence of either. Some studies of PROs have used the criterion that if greater than 15% of respondents achieved the lowest score (ICIQ-UI-SF: 0/21) or highest score (ICIQ-UI-SF: 21/21), this distribution could be evidence of a floor or ceiling effect, respectively (59). For the ICIQ-UI-SF, a score of 3/21 was used—since a score of 0/21 implies the individual does not have UI, a value which would be unexpected in this study’s sample. This score is associated with an individual leaking a small amount of urine about once a week or less often and this leakage is not reported to interfere with their lives. For ceiling effects, the percentage of respondents who scored 21/21 was reported.

Regarding floor effects, 0.56% scored 3/21. Furthermore, 10.73% of respondents scored 10/21 or below, which using this criterion suggests there is no floor effect. Regarding ceiling effects, 14.69% scored 21/21, which can be interpreted as a ceiling effect. This suggests individuals may have scored higher if the instrument allowed them to do so.

5.4 Correlations between items

For a cursory exploration of the relationships between items, both Pearson and Spearman correlations were calculated. This demonstrated that item pairs 1 and 2 (r_p= 0.50, r_s= 0.55) were much more strongly correlated than item pairs 1 and 3 (r_p= 0.28, r_s= 0.31), and item pairs 2 and 3 (r_p= 0.30, r_s= 0.30).

5.4.1 Summary

The response distributions demonstrated skewing, which when applicable, required application of methods robust to skewedness. Ceiling effects were detected at the instrument level, which hints that there may be less information revealed at high levels of UI burden. The computed correlations provided early evidence that item 3 would not be interchangeable with items 1 or 2, may be less related to the construct of UI Burden, or that more than one construct is being measured by the ICIQ-UI-SF.

Chapter 6: Confirmatory Factor Analysis

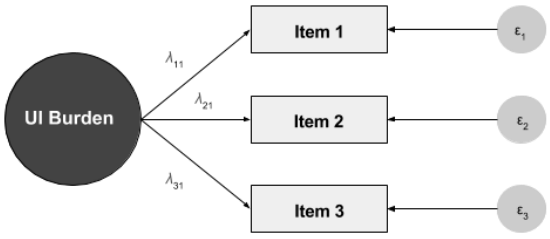
6.1 An overview of confirmatory factor analysis

Factor analysis encompasses a collection of methods that test how constructs influence responses on a number of measured items. The two types of factor analysis are: exploratory factor analysis (EFA)—which involves discovering the number and nature of constructs influencing responses, and confirmatory factor analysis (CFA)—which involves testing hypotheses about the relationships between the constructs and responses (40,60). EFA is hypothesis exploring; CFA is hypothesis testing. These hypotheses are informed by theory or previous study results (40). For example, if previous studies indicated that 'sadness', 'withdrawal', and 'lost sleep' were strong predictors of depression, there is evidence to support that an instrument containing questions about each of these symptoms, measures one underlying construct: depression.

The primary study reported that the ICIQ-UI-SF has one underlying construct. Because the items address severity and interference, both elements that contribute to disease burden and/or quality of life (26), the underlying construct was defined as "UI burden".

A one-factor model was applied to this data. It specifies that a single construct (UI burden) is the underlying cause of responses to the three ICIQ-UI-SF items. This model is summarized visually through the path diagram in Figure 6.1.

Figure 6.1 One-factor path diagram for ICIQ-UI-SF



The underlying factor, termed theta (θ) underlies the observed responses to the three ICIQ-UI-SF items (X_1, X_1, and X_3). Because they are related through the factor, UI burden, it assumes these three items are correlated. The lambdas (λ_{i1}) are factor loadings. They are a measure of association between each item and the θ. Items measuring a similar construct should exhibit similar factor loadings; this can be tested in CFA.

The purpose of this CFA was to confirm whether there is evidence supporting one underlying factor for the ICIQ-UI-SF, as suggested by the developers (37).

6.2 Methods – confirmatory factor analysis

The first assumption to proceed with a CFA includes a relatively large sample; it is recommended there are at least 15 participants per item. The second assumption is that there is indication that at least 1 factor explains the observed variation in the data; this can be investigated through inter-item correlations ($\geq$ 30) (61). The third assumption is multivariate normality. The last assumption, or requirement, is that at least 3 items are being modeled otherwise the model is under-identified (61). The parameters in the CFA can only be estimated when the number of freely estimated parameters (factor loadings and corresponding errors) does not exceed the input matrix. This is why at least 3 items are required for CFA. This will produce a 'just-identified' model where the parameter estimates will perfectly fit the data (61).

The Robust Maximum Likelihood method was chosen for this CFA is because it is robust to the violation of the normality assumption (62). In this model all item responses were treated as continuous. However, since only up to 10 response levels could be supported, the bottom two responses options (0 and 1) were combined. These had low counts and it was not expected this would bias findings. For sensitivity, the model was run treating all data as categorical, and also run one whereby the first two items were categorical and the last continuous (which is how the

instrument is structured). Wald Tests were applied to test whether factor loadings across the items are equal. Here the null hypothesis is that the parameters of item 1 = item 2 and that item 1 = item 3.

In just-identified models, goodness of fit statistics cannot be computed because they exhibit 'perfect fit'. This analysis was conducted in MPlus (63).

6.3 Results – confirmatory factor analysis

Assumptions to proceed with the analysis were met, and the Robust Maximum Likelihood was applied to address the violations to normality. All three iterations of analyses produced similar results; and indicated that factor loadings for items 1 and 2 were much more similar than item 3. For item 3, only 17% of the variance was attributable to the latent construct 'UI burden'. The Wald Test value was 17.4, with a p-value of 0.0002. This provides strong evidence against the null hypothesis that parameters of items 1 and 2 are equal to the parameters of item 3 (Table 6.1).

Table 6.1 CFA factor loadings – all continuous

	Factor Loading	R^2	Standard Error	P-Value
Item 1	0.689	0.475	0.104	0.0001
Item 2	0.732	0.536	0.105	0.0001
Item 3	0.412	0.170	0.092	0.0001

6.4 Interpretations – confirmatory factor analysis

While this CFA could not test model fit, the factor loadings revealed that two possible factors underlie the model. Items 1 and 2 are similar in that they measure symptom severity, and item 3 measures interference. However with so few items, it is not possible to test a two-factor

model. This analysis also indicates that the first two items explain most variance. This also provides evidence contrary to the primary study that one factor underlies the ICIQ-UI-SF.

6.5 Follow-up investigation – principal component analysis

As a comparative exercise, principal component analysis (PCA) was included. Factor analysis is a measurement model of a latent variable (such as "UI burden") and it does so by accounting for (co)variances among the measured items. PCA conversely produces composite variables that account for a combination of common *and* random error. Thus when wanting to explain as much variance as possible, PCA should be favoured; if wanting to understand the source of common variance, factor analysis should be pursued (64). Given the CFA model was just-identified, and model fit could not be assessed, the PCA was included as a complementary assessment of dimensionality[‡]. Principal component analysis, while primarily a technique for data reduction, can also be used for identifying dimensions.

PCA is conducted on a matrix of Pearson correlation coefficients[§], and thus should meet the assumptions for that statistic. This includes: interval-level measurement, random sampling, linearity, and normality (65). The method is robust to violations of normality with large sample sizes. Running the PCA involved the following steps (65,66):

1. Checking assumptions/sample size adequacy
2. Computing the correlation matrix
3. Extracting a full components solution
4. Applying decision criteria to select the number of components

---

[‡] It is recognized that a measurement theoretical framework would suggest PCA should be used when data reduction is the goal; when exploring factor structure, factor analysis should be favoured (64).
[§] The computational difference between factor analysis and principal component analysis is minor. Factor analysis uses communality estimates along the main diagonal of the correlation matrix, rather than unities, which are used for PCA (40).

5. Rotating for a final solution
6. Interpreting results

Assumptions of normality and linearity were assessed with residual plots. The suitability of the sample size was assessed using Kaiser's Measure of Sampling Adequacy (KMO Index), which indicates whether partial correlations between items were reasonably small to conduct the analysis; a cut-off value of >0.5 was used for proceeding with the analysis.

Following this, the correlation matrix was computed and components were extracted. Upon extracting the full component solution, a decision was made regarding how many components to retain. Generally, only the first few components account for 'meaningful' amounts of variance. Commonly used criteria are: eigenvalue-one, proportion of variance accounted for, interpretability, and scree tests (65).

The eigenvalue-one criterion is one of the most commonly used because it limits subjectivity. It states that one retains all components with eigenvalues greater than 1.00. An eigenvalue of 1.00 corresponds to 1/*k* of the total variance among a set of items (40). In PCA, each item contributes 'one unit' of variance to the total, and thus any components with eigenvalues greater than one contribute more variance than one item, and is thus would be preferable to retain those components (65).

The next criterion is the proportion of variance accounted for. This criterion is more subjective—however it is typical that researchers retain components that account for at least 70% of the variance (67).

The third criterion is interpretability. This involves assessing: (i) the constructs being investigated, (ii) number of items loading on each component, and (iii) whether there is a simple factor structure. Regarding the first interpretability criterion, suppose three items loaded on

component one, and a different four items loaded on component two. The interpretability criterion would ask whether it is reasonable to assume they are measuring different constructs (such as language or math ability). The second interpretability criterion is to check how many items load on each component—the widely used threshold is three (65). The last interpretability criterion asks whether the factor pattern demonstrates a simple structure, meaning that items have high loadings on one component and not the others.

The last criterion is to visually examine a graph of the eigenvalues, and look for a 'break' in the graph. Components that are before the break are rendered meaningful.

6.6 Results – principal component analysis

The first two items in this instrument are ordinal, and the last is a rating scale however was treated as ordinal in this analysis. Observations were independent as only one instrument per patient was used. Assumptions of linearity and normality were sufficient to proceed, and the KMO index was 0.611. The full components solution (Table 6.2) shows all of the components, which accounted for the total variance.

Table 6.2 Full components solution

	Eigenvalue	Proportion	Cumulative
Component 1	1.739	0.580	0.580
Component 2	0.766	0.255	0.835
Component 3	0.495	0.165	1.000

The application of the decision criteria to the ICIQ-UI-SF instrument is captured in Table 6.3. In this analysis, only one component met the eigenvalue-one criterion with a value of 1.74. Regarding proportion of variance, retaining one component only accounted for 58% in this study, and thus this criterion suggested retention of the first two. Since there are only three items, widely regarded as the minimum number to extract a factor, the first part of the interpretability

criterion suggested retention of one. Checking for a simple structure was not applicable with one component. Upon visual inspection, the scree plot too indicated retention of one component (results not shown).

Table 6.3 Decision criteria for component retention

		Component 1	Component 2	Component 3
Eigenvalue-one	>1	✓		
Variance accounted for	>70%	✓	✓	
Interpretability	3 items loading on each component	✓		
Scree plots	Noticeable break in plotted eigenvalues	✓		

Weighing the decision criteria, one component was retained, and all three questions 'meaningfully' loaded onto that component (see Table 6.4). All factors ranged from 0.65 to 0.82. Although it is worth noting that the third item had a much lower loading and contributed to less variance than the first two items. Again, this result is unsurprising; although all three items assess UI burden, the first two items measure more closely related underlying construct than the third item. The next step in the analysis was matrix rotation, which is a technique that makes it easier to determine what each component measures. However rotation is not possible with one component.

Table 6.4 Truncated solution and factor loadings

	Item 1	Item 2	Item 3
Factor Pattern	0.81	0.82	0.65
Final Communality Estimates (Total = 1.74)	0.65	0.67	0.42

6.1 Interpretations and implications – principal component analysis

The outcome supports the original study's proposition that the ICIQ-UI-SF is a unidimensional instrument (37), although this slightly conflicts with the results of the CFA. While the CFA could not test model fit, it did provide evidence that the third item was fundamentally different from the first two. In this PCA this is too evident, however because PCA looks at both shared and error variance, it may overstate the meaningfulness of the loading of the third item on component 1. The outcome is that if this analysis were conducted in isolation it would support unidimensionality; if conducted along with CFA, it shows that this assumption may be strong for subsequent analyses.

Chapter 7: Measures of Reliability: CTT and McDonald's Coefficient

7.1 An overview of classical test theory

Classical Test Theory (CTT) is a widely used approach to psychological assessment. The distinct feature of CTT is that it assumes an observed instrument score is comprised of an individual's 'true score', which is unobserved, plus measurement error (46):

$$\underset{\{\text{Observed Score}\}}{\boldsymbol{X}} = \underset{\{\text{True Score}\}}{\boldsymbol{T}} + \underset{\{\text{Random Error}\}}{\boldsymbol{\varepsilon}}$$

The true score, is the measure of the latent variable, and is what 'causes' an item to take on a specific value. The three primary assumptions (40) within CTT, about the relationship between the latent variable and error term are as follows:

1. Error associated with individual items varies randomly. When aggregated across many observations, it has a mean of 0.
2. An item's error term is *not correlated* with another item's error term.
3. Errors are *not correlated* with the true score of the latent variable.

In its basic form, CTT is based on the notion of parallel tests, where each item is a 'test' of the value of the latent variable. From this, two additional assumptions emerge:

4. The influence from the latent variable is assumed to be the same for all items (i.e. factor loadings are the same for each item, they are *tau-equivalent*).
5. The influence of factors outside of the latent variable is equal for all items (each item has the same amount of error).

Consequently, with strictly parallel forms, each item is *as good* a measure of the latent variable as the other items[**].

CTT assess *instruments* rather than *individual items*, with focus on the reliability and validity of these instruments. Validity is the degree to which empirical and theoretical rationale support the interpretations of test scores (68). Reliability, while conceptually thought of as the degree to which an instrument score is free of measurement error, in CTT is more specifically the proportion of variance attributable to the true score of the latent variable. This analysis applied only the most common CTT approach to assessing reliability, internal consistency, and followed by calculation of McDonald's coefficient.

7.2 Methods – internal consistency

The most commonly used measure of internal-consistency (or homogeneity) for continuous data is Cronbach's Alpha. It assesses the degree to which items in an instrument measure the same construct (45), and is based on the correlations between different items on the same test. Since, under parallel tests, there is an assumption that items are linked through the latent variable rather than the error term, and so, if the items have a strong relationship to the latent variable, they will have a strong relationship to other items (i.e. 'highly consistent'). Cronbach's alpha is calculated as follows:

$$\propto = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}}$$

[**] Note there are other approaches within CTT that loosen some of these assumptions under certain conditions (e.g. sample size requirements), such as models based on tau-equivalency (40).

where k refers to the number of items (in this case 3), $\bar{c}$ refers to the average of item covariances, and $\bar{v}$ refers to average variance. The statistic has a value between 0 and 1. A value closer to 1 implies higher consistency/redundancy (69).

7.3 Results – internal consistency

This analysis was conducted using SAS version 9.4 (Cary, NC). Table 7.1 presents the findings including the raw and standardized coefficient alpha. Table 7.2 presents the standardized coefficient alpha with deleted items. This allows one to see how each item affects the alpha.

Table 7.1 Internal consistency of ICIQ-UI-SF instrument

	Cronbach's alpha
Raw	0.438
Standardized	0.632

Table 7.2 Standardized Cronbach's alpha with deleted item

Deleted Item	Correlation with total	Alpha
Item 1	0.490	0.463
Item 2	0.502	0.446
Item 3	0.339	0.670

The standardized coefficient indicated a low/moderate level of consistency. This result failed to meet the subjective, albeit widely used threshold for acceptability of greater than, or equal to, 0.70 (46). Although it should be noted, rules of thumb are often clumsily applied—in some cases a modest alpha of 0.70 is sufficient—in other cases, where inappropriate application of the scores poses risk to patients, an alpha of 0.90 may be too low. Interpreting whether Cronbach's alpha is indeed 'high' or 'low' is only meaningful when the intended use of the instrument is clearly articulated (70). This value is also lower than reported in previous studies

(37,41,43). Deletion of the third item also results in much higher internal consistency than deletion of items one or two.

7.4 Interpretations and implications – internal consistency

The range of values from analyses of Cronbach's alpha, from 0.63 in this study, to 0.95 (37), suggests that the ICIQ-UI-SF may not be very reliable, as it is very sensitive across populations. One possible explanation for these ranges is the characteristics of the sample completing the instrument. For example, if differential item functioning is an issue for the ICIQ-UI-SF, and this study sample contains distinct groups where differential item functioning occurs, then reliability estimates will be different than previous studies. Past studies have been conducted in various cultural settings, and so this too may explain some inconsistency across samples.

A second likely explanation is that the assumptions under parallel tests and tau-equivalency was violated, as suggested in the CFA. This means that some items are unrelated to the latent construct of incontinence burden, or two constructs are being measured. Looking at Table 7.2 there is some evidence that the first two items may not be measuring the same latent construct as the third item, since the third item's correlation with the other items is much weaker compared to the other items. Conceptually, one may expect more heterogeneity with the third item, as the first two questions are concerned with discrete phenomena 'how often do you leak urine/how much do you leak' and the last asks for a subjective evaluation of interference in daily life.

A third possible explanation is that there are other issues that were not present in past validation studies, such as floor/ceiling effects. Since this sample is more clinically homogenous

with higher symptom burden than past studies, the differences in symptom severity may account for why Cronbach's alpha is much lower compared to past studies.

Interpreting Cronbach's alpha should be coupled with a number of caveats. First, if an instrument is unidimensional, simply adding questions or scales of measurement may increase the alpha. However this may counteract the benefit of having a brief instrument (such as low levels of missing data). Second, since Cronbach's alpha can assume high values even if the constructs measured are unrelated, assessing dimensionality should precede the assessment of internal consistency. Thirdly, in classical measurement, the reliability of items cannot be isolated from the instrument itself. If issues with individual items are suspected, other analytical approaches must be employed.

7.5 Conclusions from classical test theory

This analysis highlighted issues for the ICIQ-UI-SF that required application of methods outside of CTT including: differential item functioning or item response theory. These other analyses helped explain why internal consistency in this study sample is much lower than reported in past studies.

7.6 McDonald's coefficient

One of the main limitations of Cronbach's alpha is the assumption of tau-equivalence, where item variances for *true scores* are the same across items, but error variances can vary (40). Some instrument developers argue the expectation that all items are equally influenced by the latent variable is too restrictive and unrealistic, and rather, favour congeneric models. Congeneric models have the least restrictive assumptions, whereby item and error means and variances can vary across items. The central assumption is that all items have a *shared common*

latent but they do not need to be equally influenced by that latent variable. However, the stronger the association, the more reliable the instrument will be (40).

When tau-equivalency is violated, Cronbach's alpha will generate a lower bound reliability estimate. However the magnitude of this underestimation can be difficult to assess, as it can depend on the samples from which alpha is calculated (71,72).

As a complement to Cronbach's alpha, a congeneric model reliability statistic—McDonald's coefficient—was calculated as another indicator of reliability. Since it does not require associations between the items and latent to be equal for all items, it uses the factor loadings generated by the CFA (71).

$$\text{McDonald's Coefficient} = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + (\sum \varepsilon_i)}$$

The λ_i are the factor loading for item *i* , and ε_i is its corresponding error. The item r-square (λ_i^2) is the percent of variance of item *i*, explained by the underlying variable (73) (UI burden). A cut-off value for McDonald's coefficient is ≥ 0.70 (74). Again, similar to Cronbach's alpha, if the instrument is used for direct comparisons, a much higher value is favoured.

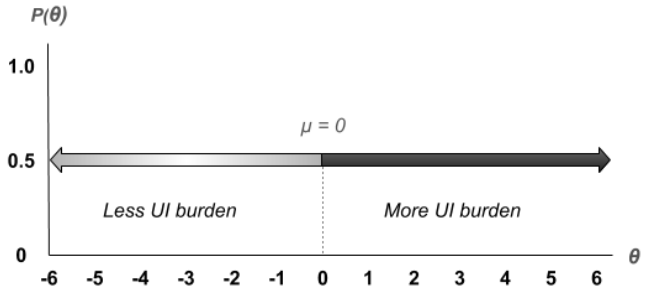
To calculate McDonald's coefficient, the factor loadings from the CFA above were used. This resulted in a value of 0.65. This value is comparable to Cronbach's alpha and lower than the recommended cut-off. This provided further evidence that the ICIQ-UI-SF has low/moderate reliability.

Chapter 8: Item Response Theory

8.1 An overview of item response theory

Item Response Theory (IRT) refers to a collection of latent trait modeling techniques, used for evaluating instrument items. While in CTT, reliability is improved through redundancy, in IRT reliability is improved by selecting better items for an instrument (40). Broadly, IRT models describe the relationship between the level of a respondent's latent trait (such as 'UI burden'), and their propensity to select certain responses when completing items in an instrument. This can be graphically depicted using item characteristic curves (ICC) (dichotomous data) or category response functions (CRF) (polytomous data). In addition, one can measure where along a respondent's latent trait the item or instrument is most reliable, through item and test information functions, respectively. Latent traits in IRT are denoted by theta (θ). A unidimensional latent trait θ in these models is transformed to have a mean of 0 and standard deviation of 1 (Figure 8.1). For example, respondents with θ=3 have UI burden 3 standard deviations away from the sample average.

Figure 8.1 Illustration of the latent trait (θ) scale in IRT



A number of IRT models can be applied based on the instrument or item characteristics. In the case of the ICIQ-UI-SF, a graded response model (GRM) will be applied since the method accommodates ordered categorical responses, and items that have varying number of response

levels (75). The general equation for the GRM is based on the 2-parameter logistic model (2PL) as follows (76,77):

$$Pi_g^* = \frac{e^{a_i(\theta - bi_g)}}{1 + e^{a_i(\theta - bi_g)}}$$

The two parameters of interest in this model are: the location (bi_g) and discrimination (a_i). The location parameter indicates the level of the latent trait (θ) where respondents are indifferent between response levels (e.g. 'strongly agree' vs. 'agree'). If the item has g response levels, there will be $g - 1$ location parameters. The discrimination parameter (slope) indicates how well items discriminate between respondents along their latent trait scale (75). Each item has one discrimination parameter. When individuals have latent traits that are close together (e.g. $\theta = 3.1$ *vs.* $\theta = 3.2$), highly discriminating items will predict with greater accuracy whether respondents will provide different responses to adjacent response levels. Low discrimination means that respondents may not answer consistently between adjacent response levels.

Complementary IRT models may include a guessing parameter (c) and upper asymptote parameter (d), however these parameterizations are often more applicable to instruments measuring ability (such as numeracy), rather than behaviours/preferences and so are not pursued here.

CRFs can be transformed into item and test information functions (IIF/TIF), each of which is an index showing the item or test's ability to differentiate across individuals as a function of their latent trait. When interpreting these graphs, high discrimination will exhibit tall and narrow IIFs/TIFs (high precision, low range). Low discrimination will be characterized as short and wide IIFs/TIFs with low precision and wide range. Depending on the intended use of the instrument, various IIF/TIF patterns are desirable. For example, if the instrument is for

screening, precision may be most important at high levels of the latent trait (e.g. between severe/very severe depression), and so a TIF peaked around $\theta = 4$ may be desirable over a peak around the average ($\theta = 0$).

8.1 Methods – item response theory

A graded response model was applied to this data. All analyses were conducted using R Version 3.3.1 and the MIRT package (78).

8.1.1 Assumptions

To proceed with an IRT analysis, several assumptions must be investigated: (i) unidimensionality of the latent trait, (ii) local independence, and (iii) item invariance (79).

The first assumption is unidimensionality; this was examined through factor analysis/principal component analysis.

The second assumption is local independence (LI). LI means that after controlling for the latent trait (e.g. depression), items are *uncorrelated* (80)—alternatively put, items should be unrelated other than they measure the same latent trait (79). The first common reason for violations to LI are that item responses depend on a common source, such as an excerpt for a language test; the second reason is due to sequential presentation of questions (81). A number of methods exist for checking LI. In this analysis, the Pearson's X^2 statistic, often applied to polytomous data was be used. X^2 is defined as the correlation of deviation scores across all examinees (82). Deviation refers to the difference between observed responses and expected performance based on the IRT model. X^2 values exceeding 0.20 generally indicate LI may be violated (83).

The last assumption is item invariance. Since the IRT model applies to all members of that population, the population must be qualitatively homogenous. If this is violated, there is

evidence of differential item functioning (DIF). DIF occurs when respondents from qualitatively distinct groups, such as male/female, who are equal on the level of a latent trait (such as level of depression), have different probabilities of selecting certain responses to an item. When DIF is present, it causes latent trait estimates to be too high/low for one group relative to the other (84).

For this analysis, DIF was assessed using the Likelihood Ratio Test (LRT), which is an IRT based method (85), although non-IRT methods can be applied as well. With the LRT approach, two models are fit. In the first, item parameters are constrained across groups, and in the second, item parameters can vary. When conducting pairwise DIF analysis, anchor items (items assumed to be DIF-free) must be selected. While the selection of DIF-free anchor items is an evolving science, different methods exist based on the type of instrument and level of prior insight about candidate DIF items. One common approach is using 'all others as anchors' (AOAA). Here each item is tested one at a time in a separate analysis, using the other items as anchors ('constrained'). One extension to this method is selecting anchors based on items with the largest discrimination parameters—however it still remains unclear 'how many' anchors this would require. For this analysis since there are only three items, the AOAA approach was used (86). DIF was performed on patients' SES level. Studies have found that low SES groups not only report higher impairment (this is a true score difference), but also provide lower valuations of their health once impaired. It is the latter than can introduce bias at a group level, when SES differences are unaccounted (87). Another study cited that among a cohort of stress UI patients, a number of factors, including socioeconomic status, independently impacted scores on instruments valuing quality of life; this again provides a rationale for considering DIF on SES (88). Although the relationship between SES and health is a robust finding in research, the relationship is not perfectly linear (89), which makes selection of a cut-off value challenging. For

this reason, two grouping were tested. The two highest SES quintiles were compared to the three lowest SES quintiles. Then for sensitivity analyses, just the highest SES quintile was compared to the lowest quintile. DIF on age was investigated, but was abandoned since response levels were very unbalanced between categories in this sample of patients.

8.1.2 Assessing model fit

First, the CRFs were visually inspected for clear distinction across response levels, and peaks for each response level somewhere across the latent trait scale. This implied that at some point along the scale of 'UI burden', one response was the most likely. In addition, it was noted if a_i parameters demonstrated good discrimination (greater than 1.70) (90). CRFs that were not well defined, or were shallow and wide, were evidence of poor discrimination. When this was identified, the IRT analysis was re-run to investigate whether definition improves by collapsing response levels. Models with collapsed response levels were compared using the Akaike information criterion (AIC) and Bayesian information criterion (BIC) statistic, with lower values indicating better fit, as well as further visual inspection of CRFs.

Assessing model fit statistically was not possible with a three-item instrument due to low degrees of freedom. Otherwise it would have been examined using the M_2 statistic, whereby the null hypothesis is perfect model fit (91).

Item and test information functions were inspected visually to make conclusions about where the along the latent trait scale items and the instrument have the highest discrimination.

8.2 Results – item response theory

Although unidimensionality was a strong assumption for this analysis based on the evidence from the CFA and PCA, the degrees of freedom were too low to investigate a multidimensional model. This analysis proceeded under the assumption of unidimensionality.

Table 8.1 presents the results for LI. Item pairs 1 and 3 met the threshold for violations to local independence, although this value was not significant.

Table 8.1 Local independence

Item Pair	X^2	P-value
1,2	0.137	0.576
1,3	0.242	0.402
2,3	-0.166	0.973

For assessing DIF, first the 2 highest SES quintiles (48.15%) were compared to the 3 lowest SES quintiles (51.85%). Those with missing deprivation indices were excluded from the analysis. No items were flagged as exhibiting DIF. Subsequently, just the highest SES quintile (20.90%) and lowest SES quintiles (16.95%) were compared. Again, no DIF was detected. For illustrative purposes, expected test score plots for the two groups are included in Appendix C, and scores were overlapping for all items.

In the first iteration of the IRT the a_i parameters for the first two items were considered high (90), indicating good discrimination (Table 8.2). It should be noted however that the curves for the lower response levels did not have unique peaks. Only 4 responses were ever the most likely across the scale of UI burden. Nevertheless, for both items 1 and 2, the probability of selecting higher response levels increased as one moved along the latent trait scale (Figure 8.2-Figure 8.4). Item 1 had a bimodal IIF, with highest discrimination when θ *was* between *-4 and -2*, and at the mean. Item 2 had a peaked and very narrow IIF suggesting highest discrimination 2 standard deviations below and above the mean θ. With item 3, the a_i parameter indicated moderate discrimination (90), and the CRFs were flat and overlapping, suggesting poor discrimination. Item fit could improve by collapsing this item's response levels. The IIF was correspondingly wide and flat. Two subsequent models were fit to see if there would be

improvement through collapsing response levels of the third item. While there are no official guidelines for collapsing, item response levels that were adjacent were selected, particularly when responses were sparse.

Table 8.3 shows which response levels were collapsed. The third model that collapsed the original ten response levels to seven had the best fit assessed through the AIC and BIC statistic (Table 8.4) and CRFs were defined across the latent trait scale. Discrimination (a_i) also improved marginally (from 0.88 to 0.90) (Figure 8.5). Statistical assessment of model fit was not possible since the degrees of freedom were too small.

The TIF (Figure 8.6) showed that the most information is revealed when respondents are below the average UI burden ($-4<\theta<0$). Thus the ICIQ-UI-SF does not differentiate well when individuals have high UI burden. The TIF of the model with collapsed response levels for the third item (Figure 8.7) showed similar results, however with a tighter range ($-2<\theta<0.5$).

Table 8.2 IRT coefficients for iterations of analyses

Parameter	Item 1			Item 2			Item 3		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
A	2.55	2.47	2.06	2.44	2.53	3.3	0.88	0.90	0.90
B1	-3.11	-3.14	-3.42	-1.19	-1.17	-1.09	-4.41	-4.34	-4.33
B2	-2.75	-2.78	-3.02	0.17	0.17	0.16	-3.56	-2.91	-2.90
B3	-2.31	-2.33	-2.52				-2.95	-1.60	-1.60
B4	-0.18	-0.18	-0.19				-1.97	-0.85	-0.51
B5							-1.65	0.11	0.92
B6							-0.92	1.41	
B7							-0.6		
B8							0.01		
B9							0.81		
B10							1.36		

Table 8.3 Response levels collapsed for model iterations

Model 1 levels (original)	Number of respondents	Model 2 levels	Number of respondents	Model 3 levels	Number of respondents
0	5	0	5	0	5
1	5	1	11	1	11
2	6				
3	17	2	24	2	24
4	7				
5	21	3	21	3	32
6	11	4	32		
7	21			4	47
8	26	5	40		
9	14			5	58
10	44	6	44		

Table 8.4 IRT model comparisons

	Model 1 *No collapsed response levels*	**Model 2** *7 response levels*	**Model 3** *6 response levels*
AIC	1381.16	1238.87	1158.90
BIC	1158.90	1241.85	1161.49

Figure 8.2 CRFs and IIFs for item 1

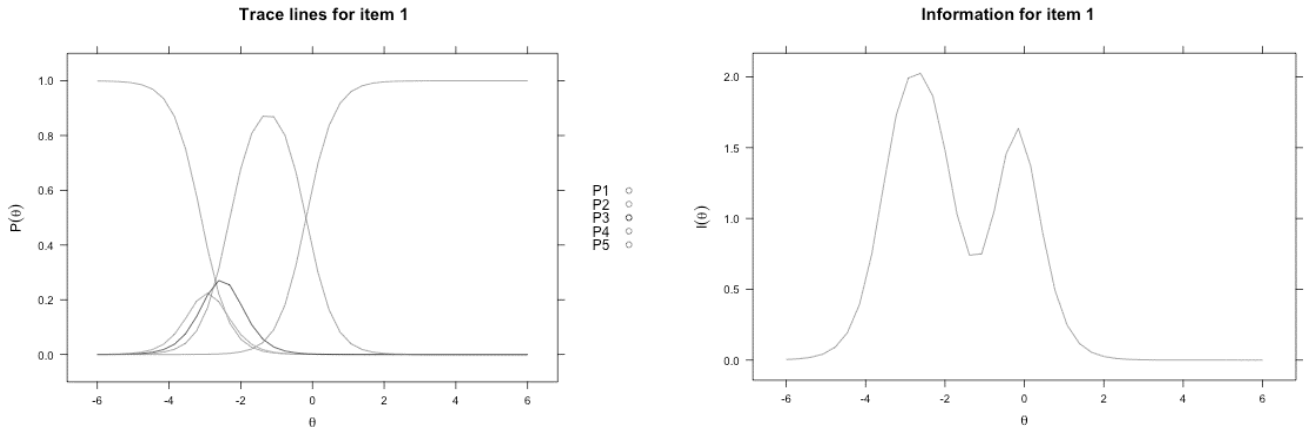


Figure 8.3 CRFs and IIFs for item 2

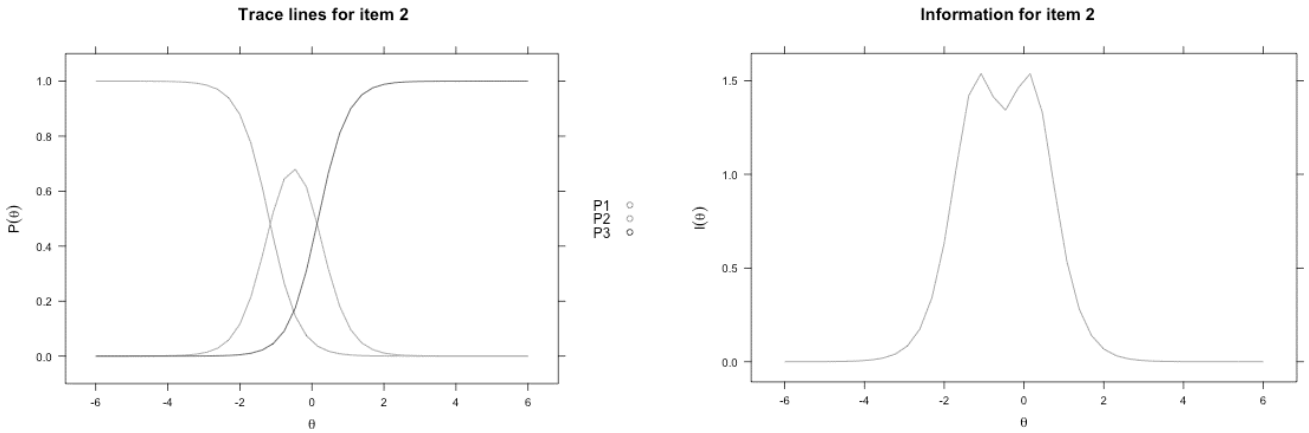


Figure 8.4 CRFs and IIFs for item 3

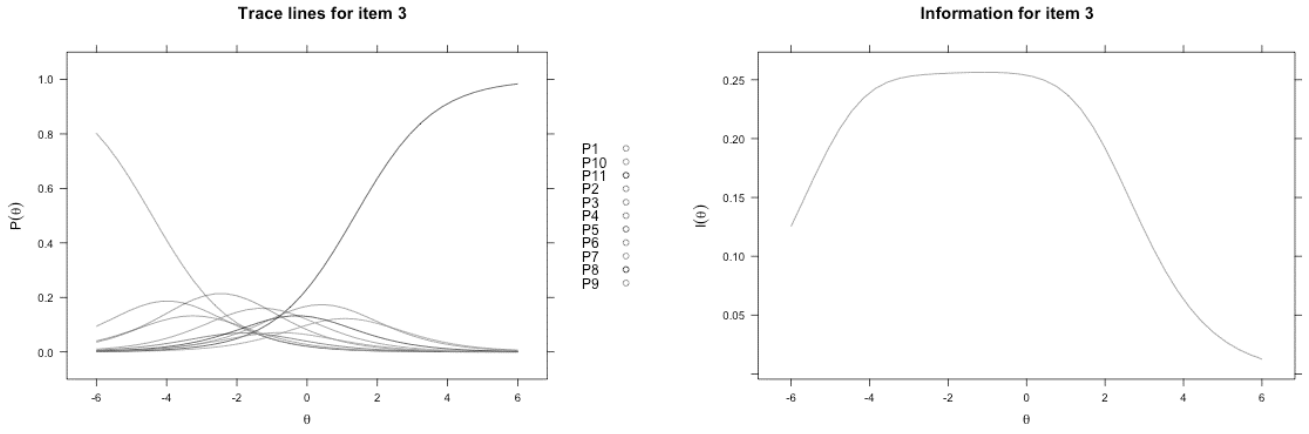
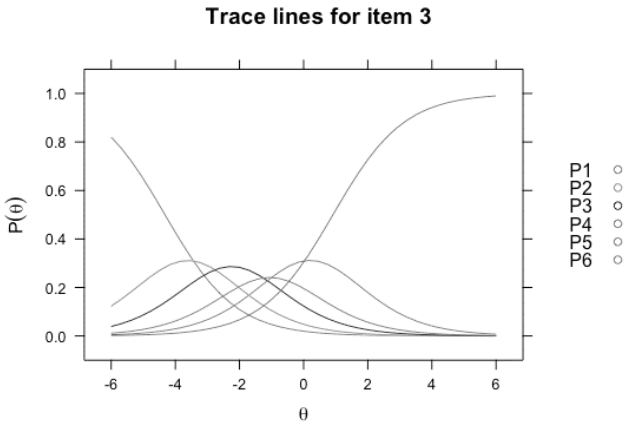


Figure 8.5 CRFs and IIFs for item 3 (collapsed model)



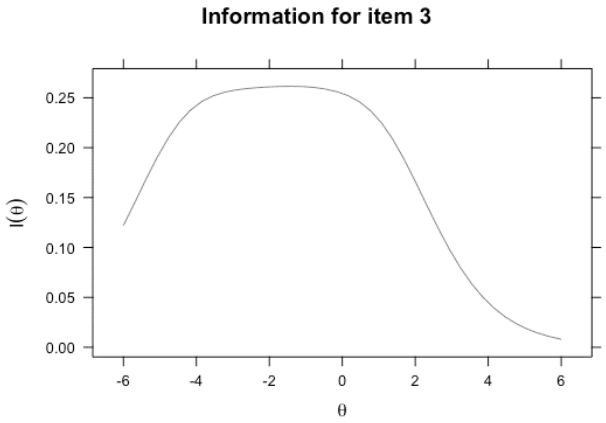


Figure 8.6 TIF - model 1

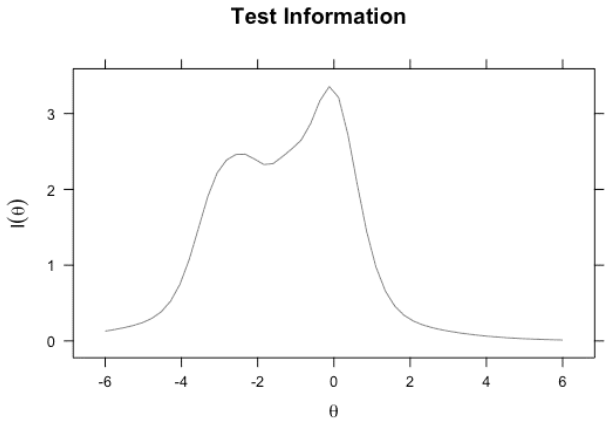
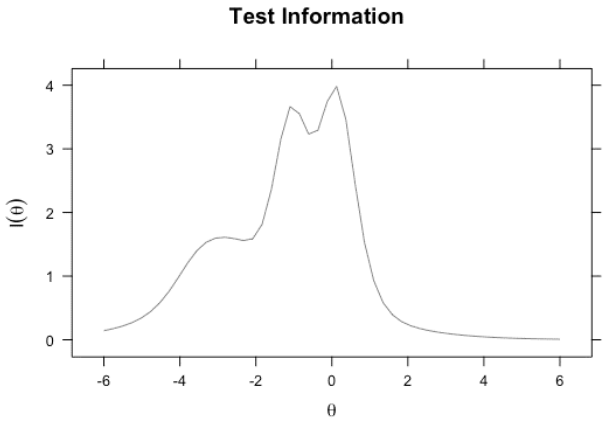


Figure 8.7 TIF - model 3



8.3 Interpretations and implications

This analysis found that the first two items have good discrimination just above and below average UI burden for this population. The third item could benefit from collapsing response levels since only 4 are ever the most likely; the low discrimination of the third item means respondents will not consistently choose between adjacent response categories (such as 7 versus 8 out of 10 when describing interference of UI), particularly at high levels of UI burden. The model fit was best when 11 response levels were collapsed to 6, and this echoes literature

investigating optimal response levels for highest reliability (92). While no DIF was found for SES, there still remain areas of investigation, particularly on sex, age, and cultural differences (88).

The primary contribution of the IRT analysis is that the ICIQ-UI-SF has low discrimination for patients with a high UI burden—this is of particular importance since most past studies have investigated populations with much lower UI burden. The degree to which this poses concern is a function of how clinicians and other health practioners intend to use the ICIQ-UI-SF instrument. For example if treatment trajectories vary greatly across high levels of UI burden, then reliability is needed on the higher end of the latent trait scale, and adjustments to the instrument should be made.

There were a number of limitations with this analysis. Model fit could only be assessed qualitatively because the degrees of freedom were too low. Since there were few respondents with low levels of UI burden, reflecting this study's sample of patients, model fit may have been undetectably poor. Fortunately this limitation does not necessarily reduce confidence in the DIF analysis, since LRT is susceptible to type I errors, rather than type II, in the event of poor model fit. Second, because the third item has so many response levels, DIF for age could not be investigated, since the two groups did not have balanced responses. Lastly, unidimensionality was a strong assumption for this analysis, although it the directionality of bias introduced is unclear.

This analysis also highlighted tension between brevity and reliability. While having low levels of missing data is desirable, many techniques for dealing with missing data exist. When confidence in the precision of scores is needed for proper application of the instrument, adding items to increase reliability whilst occasionally employing techniques for missing data may be

more desirable. Brief instruments, such as the 3-item ICIQ-UI-SF are *not necessarily* unreliable, however, the challenge they pose is quantitatively ascertaining reliability.

8.4 Conclusions from item response theory

The main finding from the IRT analysis were: the instrument does not discriminate well when respondents have a high level of UI burden, the reliability of the third item may improve by collapsing response levels, and there was no evidence of DIF between patients of high/low SES.

Chapter 9: Supplemental Investigation

Although most evaluations of instruments focus on the application of CTT and/or IRT, a number of other evaluations can be conducted that are indicators of reliability, validity, and general appropriateness for routine collection. The supplemental investigation included was a validation of the severity categories derived for the ICIQ-UI-SF against other PROs.

9.1 An overview of ICIQ-UI-SF severity categories

A past analysis mapped the ICIQ-UI-SF to the Incontinence Severity Index, which yielded four severity categories based on ones total score on the ICIQ-UI-SF: slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21) (49). The advantage of this categorization is it provides clinicians with some indicator of UI burden, which can be applied with more confidence than individual scores if there is concern about the score's reliability. Given the findings of the analyses above, the application of severity categories to patients' ICIQ-UI-SF scores may be more appropriate than individual scores. As such, similar to investigations of convergence validity in CTT, one may be interested in seeing whether these categories 'behave' in accordance to existing evidence about the effect of UI on quality of life. For example, one may want to check whether there is an inverse relationship between moving up the UI severity category, and ones self-reported general health.

The goals of this supplemental investigation were (1) assess how well the ICIQ-UI-SF severity categories concord with other commonly collected PROs, (2) determine whether these findings corroborate past analyses, and (3) comment on the appropriateness of using ICIQ-UI-SF severity categories over individual scores.

9.1.1 Methods – ICIQ-UI-SF severity categories

To conduct this analysis, instrument sum scores were assigned a severity categorization based on the previous literature: slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21). Note, given that there was only one person in the slight category, they were added to the moderate category. Thus the investigation only mapped the moderate, severe, and very severe categories. Then, responses to the PROs were compared using analysis of variance (ANOVA), with significance set at $p < 0.05$. The PROs were the EQ-5D-3L utility value and the VAS score. This is similar to assessing convergence reliability in CTT. However, rather than observing the correlations between ICIQ-UI-SF sum scores and scores of other PROs, the categories were compared to the mean scores of the other PROs.

9.1.2 Results – ICIQ-UI-SF severity categories

The results are shown in Table 9.1. As one increases ICIQ-UI-SF severity, mean EQ-5D-3L and VAS scores decreased. There was a pairwise significance between slight/moderate and very severe, (p-value: 0.02) for the EQ-5D-3L. Note there have been many MCIDs reported for the EQ-5D-3L, ranging from 0.03 to 0.54 (93)—suggesting differences in health states between the severity categories. This implies that regarding general health, there is a difference between those with slight/moderate incontinence and very severe incontinence. Regarding the VAS, the differences were not statistically significant and the differences between severe and very severe were negligible.

Table 9.1 Results from ANOVA

ICIQ-UI-SF Severity Category	EQ-5D-3L	VAS
	Mean, std. dev	Mean, std. dev
Slight/Moderate	0.86 (0.16)	78.89 (15.11)
Severe	0.79 (0.15)	72.68 (16.36)
Very severe	0.75 (0.17)	70.28 (20.01)
P-value	0.02	0.07

9.1.3 Interpretations and implications – ICIQ-UI-SF severity categories

This analysis helped to confirm that the severity categories 'behave' as literature on the impact of UI on quality of life would suggest, namely there is an inverse relationship between UI burden and general health. It is conceptually understandable that the EQ-5D-3Lwould pick up an effect over the VAS alone, as domains of mental health, self-care, and usual activities are more closely related to UI burden than just a broad question about one's health.

These findings offer the unique opportunity to 'sense check' some of the findings of the psychometric assessments above. For example, given there was only pairwise significance between slight/moderate and very severe, rather than severe and very severe, it provides additional evidence the ICIQ-UI-SF does not discriminate well across those with high incontinence burden.

Clinicians could dichotomize patients based on their ICIQ-UI-SF score as mild/moderate or severe/very severe, so that all else equal, those in the latter group would be prioritized for treatment first. Research on whether the EQ-5D-3L is a substitute for the ICIQ-UI-SF should be pursued in future research.

The limitation of this analysis is it was not possible to examine differences between slight and moderate incontinence because apart from one patient, no one classified as having slight

incontinence. The challenge is it would likely be assessed in only non-surgical groups, since those with slight incontinence would generally not be candidates for surgery.

Chapter 10: Summary of Analyses

The analyses above, in concert, revealed aspects of what the ICIQ-UI-SF measures, and how well the instrument and items do so. Table 10.1 summarizes the analyses conducted, their primary findings, and implications.

Table 10.1 Summary of analyses

Analysis	Purpose	Main Findings/Implications
Descriptive statistics	Understand response pattern (e.g. skew), floor/ceiling effects	• Responses to all questions are left skewed • Instrument exhibits ceiling effects • Past measures of responsiveness may be underestimated
Confirmatory factor analysis, Principal component analysis	Confirm whether ICIQ-UI-SF is unidimensional Compare result with other method	• CFA could not test model fit but showed factor loadings are different ($p < 0.0002$) between items 1 and 2, and item 3. • PCA provided some evidence for unidimensionality, but showed proportion of variance explained is lower than expected for a unidimensional instrument
Cronbach's alpha	Calculate a measure of reliability	• Reliability is low/moderate (0.63) • Reliability is too low for direct patient comparisons; better suited for group averages or as a complement to other measures
McDonald's coefficient	Calculate an alternative measure of reliability	• Reliability is low/moderate (0.65) • Reliability is too low for direct patient comparisons; better suited for group averages or as a complement to other measures
Item response theory	Investigate item level characteristics	• Does not discriminate among those with high incontinence burden • Item 3 has too many response levels • Most information is gathered when individuals have less UI burden than the mean (up to -4 SDs)
Mapping of severity categories	Preliminary investigation of potential application	• Severity categories are sensitive to differences between slight/moderate and very severe

Chapter 11: Discussion

11.1 Recommendations on the use of the ICIQ-UI-SF

Conclusions from psychometric evaluations must be tailored to the use of the instrument, as expectations for reliability and validity vary based on the application. This analysis, in combination with past studies generally shows that reliability of the ICIQ-UI-SF is low/moderate—as such, assigning a severity level or categorizing scores in some way may be appropriate for broad use. If the instrument is intended for facilitating patient-clinician discussions, where the scores are used more qualitatively, it may be used as is. If the goal is to use the ICIQ-UI-SF for direct comparisons, the instrument should be amended. This analysis also indicated that the severity questions and interference questions might not be measuring one underlying construct of UI burden. Thus, if a clinician has interest in understanding interference aspects of quality of life, other instruments may be favoured. The fact that the severity items explain much more variance than the third interference item may be desirable for clinicians, as they have much more control on this aspect through the intervention, than elements of interference. Table 11.1 summarizes some recommendations and considerations for applying the ICIQ-UI-SF in the surgical population.

Table 11.1 Recommendations for intended use

Intended Use	Recommendations/Notes
Directly comparing patient scores (e.g. triage)	• Reliability is too low for direct patient comparison • Risk of misclassification is high due to measurement error
Summarizing group averages (e.g. pre-post change of group)	• Ceiling effects may mute pre-post scores • Will primarily detect differences in symptom severity
Assessing impact on quality of life	• Current instrument mostly explains differences in severity in patients, not interference • If interested in 'interference' more so than 'symptom severity' aspects of quality of life, can consider using other instruments as a complement. After examination of their measurement properties, possible candidates are the Incontinence Impact Questionnaire (94) or Incontinence Quality of Life Questionnaire, although it is propriety (95).
Facilitating discussions	• Starting point for discussion, used as a complement to other measures

Reflecting on the intended use of the instrument is also salient in deciding what type of psychometric evaluation one conducts. If analyses are cursory and budget is limited, relying on CTT, some academics say is sufficient. However if the results are to be actionable and affect respondents at an individual level (triage, delivery of care, etc.) academics suggest augmenting evaluations with IRT, since it is more capable of detecting items that threaten validity/reliability (96). One example of where the bar for reliability is set very high is in FDA labeling claims (97).

11.2 Limitations and areas for future research

One of the pressing areas for future research is creating taxonomy for the use of the diagnostic item, which asks 'When does urine leak?' Although the diagnostic item is one of the more widely used components of the ICIQ-UI-SF, there is little standardization in its application. While this does not fall under psychometric assessment per se, it was added to this instrument at the request of clinicians, and so it is likely required to warrant collection. To illustrate

idiosyncrasies of the diagnostic item, consider this study sample where five combinations of responses comprise 49.1% of the sample (Table 11.2). While some combinations are likely to be interpreted with less ambiguity—for example ticking 'leaks urine when I cough/sneeze' along with 'leaks urine when I am physically active'. This combination of responses would widely be an indicator of some level of stress incontinence. However with other responses, there may be heterogeneity in the interpretation of results. Consider the following response patterns: some patients *only* tick 'leaks urine all the time', others tick every box *as well as* 'urine leaks all the time', and some tick half of the options *as well as* 'leaks urine all the time'. Ostensibly all of these patients have the same problem: 'urine leaks all the time'—however perhaps the additional ticking of responses may be a signal to clinicians that patients are more burdened. In this sample, using ANOVA, there were no statistical differences in the ICIQ-UI-SF, EQ-5D, or VAS scores between those that answered just 'leaks all the time' and those that ticked all responses. As such, it would be expected that patients would be triaged and treated equally, but whether this would happen in practice is unclear. Follow up discussions with clinicians may be a useful avenue for understanding the relative importance of how respondents complete the diagnostic item.

Table 11.2 Combination of diagnostic item responses

Combination of responses	N	% of study sample
8	32	18.1
2,3,4,5,6,7,8	20	11.3
2,3,5,6,7	14	7.9
3,5	11	6.2
3,5,7	10	5.6

The next area for research is further investigation of DIF, since this sample was unable to investigate DIF based on age, sex, and cultural background.

The third area for research is investigation of ordering effects of instrument items. Both theory and empirical work has shown that the ordering of questions can have consequences. For example, one study found that ordering response options from ‘poor’ to ‘excellent’ might reduce positive clustering. In particular, reducing the attractiveness of the first option is one of the reasons why one may put the ‘least desirable’ option first. Another study found that putting self-reported general health questions after domain-specific questions affected their general health responses (98) suggesting there may be a difference if item 3 (a more generic quality of life question) preceded items 1 and 2. While for the ICIQ-UI-SF this may be less pressing since it is a brief instrument, it should be noted that this phenomenon can affect reliability for PROs and should be considered in evaluations.

The next area for research is to do a formal crosswalk between a generic instrument, such as the EQ-5D-3L, and the ICIQ-UI-SF, to create a repository of possible instruments that can act as substitutes rather than complements. Since respondents have limited time and attention, PROs should be used judiciously—and so in some cases just collecting the EQ-5D-3L and the diagnostic item may be sufficient. Furthermore, many generic PROs have undergone very scrupulous psychometric assessment and may be a favourable option over a condition-specific instrument with moderate levels of reliability.

The final area for research is revisiting the MCID. While this research did not study the MCID—it did detect ceiling effects, which may impact this measure. Given the low/moderate reliability of the ICIQ-UI-SF detected in this study, the previously calculated MCID may not be applicable to this study’s population.

11.3 Conclusion

As PROs garner more attention, it is important to set standards for instruments' evaluation that correspond to their intended use. While the upfront investment of time may seem discouraging and is not within every analyst's toolkit, the high bar should not be reserved to psychometricians. Particularly in high-stake settings such as using PROs for triage or cost-utility analysis, applying methods outside of CTT may be required to maximize the use of PROs. One likely bottleneck to the uptake of proper evaluation is disconnect between academics that conduct analyses, and the setting they are applied in. Clearer articulation of how low/moderate/high reliability and validity affects interpretations of scores for example, may increase demand for higher quality instrument evaluations.

The movement toward patient-centered care offers a promising future for the routine collection of PROs, as they are a quick and cost-effective way of integrating patient level data into the timeliness and type of care patients receive. Regarding the ICIQ-UI-SF for routine collection, as it stands, both results from CTT and IRT reveal reliability is low/moderate, and is particularly threatened for patients with high UI burden.

References

1. Bettez M, Tu LM, Carlson K, Corcos J, Gajewski J, Jolivet M, et al. 2012 update: Guidelines for adult urinary incontinence collaborative consensus document for the Canadian Urological Association. J Can Urol Assoc. 2012;6(5):354–63.
2. Thom D. Variation in estimates of urinary incontinence prevalence in the community: effects of differences in definition, population characteristics, and study type. J Am Geriatr Soc. 1998;46(4):473–80.
3. Demaagd GA, Davenport TC. Management of urinary incontinence. P T. 2012;37(6):345–361H.
4. Incontinence : The Canadian Perspective. Can Cont Found. 2014;(December):1–31.
5. McKenna SP. Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. BMC Med [Internet]. 2011;9(1):86. Available from: http://doi.org/10.1186/1741-7015-9-86
6. McGrail K, Bryan S, Jennifer Davis. Let's all go to the PROM: The case for routine patient-reported outcome measurement in Canadian healthcare. Vol. 11, Healthcare Papers. 2012. p. 8–18.
7. Canadian Institute for Health Information. PROMs Background Document. 2009;
8. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. Patient reported outcome measures in practice. Bmj [Internet]. 2015;350(feb10 14):g7818–g7818. Available from: http://doi.org/10.1136/bmj.g7818
9. Lamers I, Kelchtermans S, Baert I, Feys P. Upper limb assessment in multiple sclerosis: A systematic review of outcome measures and their psychometric properties. Arch Phys Med Rehabil. 2014;95(6):1184–200.

10. Walmsley S, Williams AE, Ravey M, Graham A. The rheumatoid foot: A systematic literature review of patient-reported outcome measures. J Foot Ankle Res. 2010;3(1).
11. McKertich K. Urinary incontinence-assessment in women: stress, urge or both? Aust Fam Physician [Internet]. 2008;37(3):112–7. Available from: http://doi.org/10.1016/j.regg.2010.05.004
12. Lucas MG, Bedretdinova D, Berghmans LC, Bosch JLHR, Burkhard FC, Cruz F, et al. Guidelines on Urinary Incontinence. Eur Assoc Urol [Internet]. 2015; Available from: http://doi.org/10.1016/j.acuro.2011.03.012
13. Christine Khandelwal D, Kistler C. Diagnosis of urinary incontinence. Am Fam Physician. 2013;87(8):543–50.
14. Oh SJ, Ku JH, Hong SK, Kim SW, Paick JS, Son H. Factors influencing self-perceived disease severity in women with stress urinary incontinence combined with or without urge incontinence. Neurourol Urodyn. 2005;24(4):341–7.
15. Abrams P. Describing bladder storage function: Overactive bladder syndrome and detrusor overactivity. In: Urology. 2003. p. 28–37.
16. Nitti VW. The prevalence of urinary incontinence. Rev Urol [Internet]. 2001;3 Suppl 1:S2-6. Available from: http://doi.org/10.1034/j.1600-0412.2000.0790121056.x
17. Tanagho EA, McAninch JW. Smith's General Urology. Smith's General Urology. 2004. 193-218 p.
18. Thom DH, Van den Eeden SK, Brown JS. Evaluation of partutition and other reproductive variables as risk factor for urinary incontinence in later life. Vol. 90, Obstetrics and gynecology. 1997. p. 983–9.
19. Bump R, McClish D. Cigarette smoking and pure genuine stress incontinence of urine: a

comparison of risk factors and determinants between smokers and nonsmokers. Am J Obstet [Internet]. 1994;170:579–82. Available from: http://doi.org/10.1016/S0002-9378(94)70231-4

20. Kemmer H, Mathes AM, Dilk O, Gröschel A, Grass C, Stöckle M. Obstructive sleep apnea syndrome is associated with overactive bladder and urgency incontinence in men. Sleep. 2009;32(2):271–5.

21. Sandhu JS. Treatment options for male stress urinary incontinence. Nat Publ Gr [Internet]. 2010;7(4):222–8. Available from: http://dx.doi.org/10.1038/nrurol.2010.26

22. Teunissen D, van Weel C, Lagro-Janssen T. Urinary incontinence in older people living in the community: Examining help-seeking behaviour. Br J Gen Pract. 2005;55(519):776–82.

23. Mallah F, Montazeri A, Ghanbari Z, Tavoli A, Haghollahi F, Aziminekoo E. Effect of Urinary Incontinence on Quality of Life among Iranian Women. J Fam Reprod Heal. 2014;8(1):13–9.

24. Kwon BE, Kim GY, Son YJ, Roh YS, You MA. Quality of life of women with urinary incontinence: A systematic literature review. Int Neurourol J. 2010;14(3):133–8.

25. Hajjar RR. Psychosocial impact of urinary incontinence in the elderly population. Clin Geriatr Med. 2004;20(3):553–64.

26. Chiaffarino F, Parazzini F, Lavezzari M, Giambanco V. Impact of urinary incontinence and overactive bladder on quality of life. Eur Urol. 2003;43(5):535–8.

27. Hajebrahimi S, Corcos J, Lemieux MC. International consultation on incontinence questionnaire short form: Comparison of physician versus patient completion and immediate and delayed self-administration. Vol. 63, Urology. 2004. p. 1076–8.

28. Barentsen JA, Visser E, Hofstetter H, Maris AM, Dekker JH, de Bock GH. Severity, not type, is the main predictor of decreased quality of life in elderly women with urinary incontinence: a population-based study as part of a randomized controlled trial in primary care. Health Qual Life Outcomes [Internet]. 2012;10(1):153. Available from: http://doi.org/10.1186/1477-7525-10-153

29. Coyne KS, Zhou Z, Thompson C, Versi E. The impact on health-related quality of life of stress, urge and mixed urinary incontinence. BJU Int. 2003;92(7):731–5.

30. Kadono Y, Nohara T, Kadomoto S, Nakashima K, Iijima M, Shigehara K, et al. Investigating urinary conditions prior to robot-assisted radical prostatectomy in search of a desirable method for evaluating post-prostatectomy incontinence. Anticancer Res. 2016;36(8):4293–8.

31. Da Roza T, De Araujo MP, Viana R, Viana S, Jorge RN, Bø K, et al. Pelvic floor muscle training to improve urinary incontinence in young, nulliparous sport students: A pilot study. Int Urogynecol J Pelvic Floor Dysfunct. 2012;23(8):1069–73.

32. Patki P, Woodhouse JB, Patil K, Hamid R, Shah J. An effective day case treatment combination for refractory neurophatic mixed incontinence. Int Braz J Urol. 2008;34(1):63–72.

33. Bedretdinova D, Fritel X, Panjo H, Ringa V. Prevalence of Female Urinary Incontinence in the General Population According to Different Definitions and Study Designs. Eur Urol. 2016;69(2).

34. Stirling Bryan, James Broesch, Kacey Dalzell, Jennifer Davis, Kim McGrail MJM. What are the most effective ways to measure patient health outcomes of primary health care integration through PROM (Patient Reported Outcome. Vancouver, BC Cent ….

2013;(April 2013).

35. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Vol. 14, International Journal of Rehabilitation Research. 2008. 364 p.

36. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119(2).

37. Avery K, Donovan J, Peters TJ, Shaw C, Gotoh M, Abrams P. ICIQ: A brief and robust measure for evaluating the symptoms and impact of urinary incontinence. Neurourol Urodyn. 2004;23(4):322–30.

38. Bristol Urological Institute. ICIQ Structure [Internet]. 2014 [cited 2016 Jul 26]. Available from: http://www.iciq.net/structure.html

39. Donovan J, Badia X, Corcos M, Gotoh M, Kelleher C, Naughton M, et al. Symptom and quality of life assessment. Incontinence Proc 2nd Int Consult Incontinence. 2003;519–84.

40. DeVellis RF. Scale Development: Theory and Applications. Vol. 26, Applied Social Research Methods Series. 2003. 1-171 p.

41. Hashim H, Avery K, Mourad MS, Chamssuddin A, Ghoniem G, Abrams P. The Arabic ICIQ-UI SF: An alternative language version of the English ICIQ-UI SF. Neurourol Urodyn. 2006;25(3):277–82.

42. Espuña Pons M, Castro Díaz D, Carbonell C, Dilla T. [Comparison between the "ICIQ-UI Short Form" Questionnaire and the "King's Health Questionnaire" as assessment tools of urinary incontinence among women]. Actas Urol españolas. 2007;31(5):502–10.

43. Pereira SB, Thiel RDRC, Riccetto C, Silva JM Da, Pereira LC, Herrmann V, et al. Validation of the International Consultation on Incontinence Questionnaire Overactive

Bladder (ICIQ-OAB) for Portuguese. Rev Bras Ginecol Obstet. 2010;32(6):273–8.

44. Haynes SN, Richard DCS, Kubany ES. Content Validity in Psychological Assessment : A Functional Approach to Concepts and Methods Introduction to Content Validity. Psychol Assess. 1995;7(3):238–47.

45. Brown A, Stochl J, Tim Croudace, (University of Cambridge D of P, Jan Boehnke, (University of Trier D of CP and P. Assessment , analysis and interpretation of Patient Reported Outcomes (PROs). 2011.

46. Elkin E. Are You in Need of Validation? Psychometric Evaluation of Questionnaires Using SAS. SAS Glob Forum. 2012;1–9.

47. Espuña Pons M, Montserrat Puig C, Rebollo P, Vanrell Díaz JA, Iglesias Guiu X. Evaluation of the results of surgery treatment for female stress urinary incontinence with the ICIQ-UI SF questionnaire [Internet]. Vol. 124, Medicina Clinica. 2005. p. 772–4. Available from: http://doi.org/10.1016/j.ejogrb.2014.06.020

48. Seckiner I, Yesilli C, Mungan NA, Aykanat A, Akduman B. Correlations between the ICIQ-SF score and urodynamic findings. Neurourol Urodyn. 2007;26(4):492–4.

49. Klovning A, K A, H S, S. H. Comparison of Two Questionnaires for Assessing the Severity of Urinary Incontinence: The ICIQ-UI SF Versus the Incontinence Severity Index. Neurourol Urodyn. 2009;28(5):411–5.

50. Espuna-Pons M, Dilla T, Castro D, Carbonell C, Casariego J, Puig-Clota M. Analysis of the Value of the ICIQ-UI SF Questionnaire and Stress Test in the Differential Diagnosis of the Type of Urinary Incontinence. Neurourol Urodyn. 2007;26:836–41.

51. Sirls LT, Tennstedt S, Brubaker L, Kim H-Y, Nygaard I, Rahn DD, et al. The minimum important difference for the International Consultation on Incontinence Questionnaire-

Urinary Incontinence Short Form in women with stress urinary incontinence. Neurourol Urodyn. 2015;34(2):183–7.

52. 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.

53. Hamilton DF, Giesinger JM, MacDonald DJ, Simpson AHRW, Howie CR, Giesinger K. Responsiveness and ceiling effects of the Forgotten Joint Score-12 following total hip arthroplasty. Bone Jt Res [Internet]. 2016;5(3):87–91. Available from: http://doi.org/10.1302/2046-3758.53.2000480

54. Sutherland JM, Crump RT, Chan AMPA, Liu G, Yue E, Bair M. Health of Patients on the Waiting List: Opportunity to Improve Health in Canada? Health Policy (New York) [Internet]. 2016;120(7):749–57. Available from: http://doi.org/10.1016/j.healthpol.2016.04.017

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

56. Weinstein MC, Torrance G, McGuire A. QALYs: The basics. In: Value in Health. 2009.

57. Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian valuation of EQ-5D health states: Preliminary value set and considerations for future valuation studies. PLoS One. 2012;7(2).

58. Vincent K, Sutherland JM. A Review of Methods for Deriving an Index for Socioeconomic Status in British Columbia. 2013;(April).

59. Lim CR, Harris K, Dawson J, Beard DJ, Fitzpatrick R, Price AJ. Floor and ceiling effects in the OHS: an analysis of the NHS PROMs data set. BMJ Open [Internet]. 2015;5(7):e007765. Available from: http://doi.org/10.1136/bmjopen-2015-007765

60. Jackson D, Gillaspy A, Purc-Stephenson R. Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendations. Psychol Methods [Internet]. 2009;14(1):6–23. Available from: http://doi.org/10.1037/a0014694

61. Brown TA. Confirmatory Factor Analysis for Applied Research. Methodology in the Social Sciences. 2006. 483 p.

62. Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods [Internet]. 2012;17(3):354–73. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0029315

63. Muthén L, Muthén B. MPlus (version 6.11). Los Angeles, California; 2010.

64. Preacher KJ, MacCallum RC. Repairing Tom Swift's Electric Factor Analysis Machine. Underst Stat [Internet]. 2003;2(1):13–43. Available from: http://doi.org/10.1207/S15328031US0201_02

65. SAS. Introduction : The Basics of Principal Component Analysis. In p. 1–56. Available from: https://doi.org/10.1002/ibd.21544

66. Shlens J. A Tutorial on Principal Component Analysis. 2003.

67. Kim J, Mueller CW. Review of factor analysis basics. In: Factor analysis statistical methods and practical issues. 1978. p. 8–87.

68. Messick S. Validity of Psychological Assessment. 1995;50(9):741–9.

69. Streiner DL. Being inconsistent about consistency: When coefficient apha does and doesn't matter. J Pers Assess. 2003;80(3):217–22.

70. Lance CE, Butts MM, Michels LC. The Sources of Four Commonly Reported Cutoff Criteria. Organ Res Methods [Internet]. 2006;9(2):202–20. Available from:

http://doi.org/10.1177/1094428105284919

71. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014;105(3):399–412.

72. Graham JM. Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability. Educ Psychol Meas [Internet]. 2006;66(6):930–44. Available from: http://doi.org/10.1177/0013164406288165

73. Raykov T. Coefficient Alpha and Composite Reliability With Interrelated Nonhomogeneous Items. Appl Psychol Meas [Internet]. 1998;22(4):375–85. Available from: http://doi.org/10.1177/014662169802200407

74. Fornell C, Larcker DF. Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. J Mark Res [Internet]. 1981;18(3):382. Available from: http://doi.org/10.2307/3150980

75. Ostini R, Nering ML. Polytomous Item Response Theory Models. Thousand Oaks: Sage Publications; 2006. 107 p.

76. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38:II28-II42.

77. An X, Yung Y. Item Response Theory : What It Is and How You Can Use the IRT Procedure to Apply It. SAS Inst Inc. 2014;1–14.

78. Chalmers RP. mirt : A Multidimensional Item Response Theory Package for the R Environment. J Stat Softw [Internet]. 2012;48(6):1–29. Available from: http://doi.org/10.18637/jss.v048.i06

79. Nguyen TH, Han H-R, Kim MT, Chan KS. An Introduction to Item Response Theory for Patient-Reported Outcome Measurement. Patient - Patient-Centered Outcomes Res.

2014;7(1):23–35.

80. McDonald RP. The dimensionality of tests and items. Br J Math Stat Psychol [Internet]. 1981;34(1):100–17. Available from: http://doi.org/10.1111/j.2044-8317.1981.tb00621.x

81. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. In: Quality of Life Research. 2007. p. 5–18.

82. Yen WM. Scaling Performance Assessments : Strategies for Managing Local Item Dependence. J Educ Meas. 1993;30(3):187–213.

83. Chen W-H, Thissen D. Local Dependence Indexes for Item Pairs Using Item Response Theory. J Educ Behav Stat [Internet]. 1997;22(3):265–89. Available from: http://doi.org/10.3102/10769986022003265

84. Holland PW, Wainer H. Differential item functioning. Differential item functioning. 1993.

85. Magis D, Béland S, Tuerlinckx 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.

86. Meade AW, Wright N a. Solving the measurement invariance anchor item problem in item response theory. J Appl Psychol. 2012;97(5):1016–31.

87. Mielck A, Vogelmann M, Leidl R. Health-related quality of life and socioeconomic status: inequalities among adults with a chronic disease. Health Qual Life Outcomes. 2014;12(1):58.

88. Tincello D, Sculpher M, Tunn R, Quail D, Van Der Vaart H, Falconer C, et al. Patient characteristics impacting health state index scores, measured by the EQ-5D of females with stress urinary incontinence symptoms. Value Heal. 2010;13(1):112–8.

89. Fink G. Stress consequences: Mental, neuropsychological and socioeconomic. In: Stress

consequences: Mental, neuropsychological and socioeconomic. 2010. p. xxiii,-756.

90. Frank B. Baker. The Basics of Item Response Theory. 2nd ed. Evaluation. USA: ERIC Clearinghouse on Assessment and Evaluation; 2001.

91. 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.

92. Preston C, Colman A. Optimal number of response categories in rating scales: reliability, validity, …. Acta Psychol (Amst) [Internet]. 2000; Available from: https://doi.org/10.1016/S0001-6918(99)00050-5

93. Briggs A, Lloyd A, Pickard S SJ. Minimal Clinically Important Difference in Eq-5D: We Can Calculate It, But Does That Mean We Should? [Internet]. ISPOR 22nd Annual International Meeting. 2017 [cited 2017 Sep 26]. p. IP15. Available from: https://myisporboston.zerista.com/event/member/360798?embedded=1

94. Handal VL, Massof RW. Measuring the Severity of Stress Urinary Incontinence Using the Incontinence Impact Questionnaire. Neurourol Urodyn. 2004;23(1):27–32.

95. Van De Vaart H, Falconer C, Quail D, Timlin L, Manning M, Tincello D, et al. Patient reported outcomes tools in an observational study of female stress urinary incontinence. Neurourol Urodyn. 2010;29(3):348–53.

96. Petrillo Stefan J ; McLeod, Lori D ; Coon, Cheryl D J; C. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. Value Health. 2015;18(1):25–34.

97. Rothrock NE, Kaiser KA, Cella D. Developing a Valid Patient-Reported Outcome Measure. Clin Pharmacol Ther [Internet]. 2011;90(5):737–42. Available from: http://doi.wiley.com/10.1038/clpt.2011.195

98. Garbarski D, Schaeffer NC, Dykema J. The effects of response option order and question order on self-rated health. Qual Life Res [Internet]. 2014;1443–53. Available from: http://dx.doi.org/10.1007/s11136-014-0861-y

Appendices

Appendix A PROs used for urinary incontinence

- International Consultation on Incontinence Questionnaire Short Form (ICIQ-UI-SF)
- Incontinence Quality of Life Questionnaire (I-QoL)
- Female/Male Lower Urinary Tract Symptoms (ICIQ-FLUTS, ICIQ-MLUTS)
- Incontinence Impact Questionnaire (IIQ)
- Urinary Incontinence-Specific Quality of Life Instrument (ICIQ-Uqol)
- Incontinence Symptom Severity Index (ISS)
- King's Health Questionnaire (KHQ)
- Leicester Urinary Symptom Questionnaire (LUSQ)
- Nocturia Quality of Life Questionnaires (N-QoL)
- Overactive Bladder Questionnaire (OAB-q)
- Pelvic Floor Distress Inventory (PFDI)
- Nocturia Quality of Life Questionnaires (N-QoL)
- Protection, Amount, Frequency, Adjustment, Body image (PRAFAB)
- Overactive Bladder Questionnaire (OAB-q)
- Protection, Amount, Frequency, Adjustment, Body image (PRAFAB)
- Pelvic Floor Distress Inventory (PFDI)
- Quality of Life Assessment Questionnaire Concerning Urinary Incontinence (Contilife)
- Urinary Incontinence Severity Score (UISS)
- Incontinence Outcome Questionnaire (IOQ)
- Actionable Bladder Symptom Screening (ABSST)
- Incontinence Severity Index (ISI)
- Incontinence Stress Index (ISQ)
- Urinary Incontinence Severity Score (UISS)
- Epidemiology of Prolapse and Incontinence Questionnaire (EPIQ)

Appendix B International Consultation on Incontinence Questionnaire Short-Form

1. **How often do you leak urine?** (Tick one box)
   - □ Never
   - □ About once a week or less often
   - □ Two or three times a week
   - □ About once a day
   - □ Several times a day
   - □ All the time

2. **We would like to know *how much* urine you think leaks? How much urine do you *usually* leak (whether you wear protection or not)?** (Tick one box)
   - □ None
   - □ A small amount
   - □ A moderate amount
   - □ A large amount

3. **Overall, how much does leaking urine interfere with your everyday life?** Please circle a number between 0 (not at all) and 10 (a great deal).

0 1 2 3 4 5 6 7 8 9 10

Not at all A great deal

4. **When does urine leak** (Please tick all that apply to you)
   - □ Never – urine does not leak
   - □ Leaks before you can get to the toilet
   - □ Leaks when you cough or sneeze
   - □ Leaks when you are asleep
   - □ Leaks when you are physically active/exercising
   - □ Leaks when you are finished urinating and are dressed
   - □ Leaks for no obvious reason
   - □ Leaks all the time

Appendix C Comparison of expected scores of low and high SES respondents