Published in final edited form as:

Qual Life Res. 2016 July ; 25(7): 1645–1654. doi:10.1007/s11136-015-1222-1.

Item bank development, calibration and validation for patient-reported outcomes in female urinary incontinence

Vivian W. Sung, MD, MPH¹, James W. Griffith, PhD², Rebecca G. Rogers, MD³, Christina A. Raker, ScD⁴, and Melissa A. Clark, PhD⁵

¹The Division of Urogynecology and Reconstructive Pelvic Surgery, Department of Obstetrics and Gynecology, Alpert Medical School of Brown University, Providence, RI

²Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago IL

³The Division of Urogynecology, Departments of Obstetrics and Gynecology and Surgery, University of New Mexico, Albuquerque, NM

⁴The Division of Research, Department of Obstetrics and Gynecology, Alpert Medical School of Brown University, Providence, RI

⁵Departments of Epidemiology and Obstetrics and Gynecology, Alpert Medical School and School of Public Health of Brown University, Providence, RI

Abstract

Purpose—Current patient-reported outcomes for female urinary incontinence (UI) are limited by their inability to be tailored. Our objective is to describe the development and field-testing of 7 item banks designed to measure domains identified as important UI in females (UIf). We also describe the calibration and validation properties of the UIf-item banks, which allow for more efficient computerized-adaptive testing (CAT) in the future.

METHODS—The UIf-measures included 168 items covering 7 domains: Stress UI (SUI), Overactive Bladder (OAB), Urinary Frequency, Physical, Social and Emotional Health Impact, and Adaptation. Items underwent rigorous qualitative development and psychometric testing across 2 sites. Items were calibrated using item response theory and evaluated for internal consistency, construct validity and responsiveness.

RESULTS—750 women (249 SUI, 249 OAB, and 252 mixed UI) participated. Mean age was 55±14 years, 23% were Hispanic, 80% white. In addition to face and content validity, the

Correspondence: Vivian W. Sung, MD, MPH, Division of Urogynecology and Reconstructive Pelvic Surgery, Women and Infants' Hospital / Warren Alpert Medical School at Brown University, 695 Eddy Street, Lower Level, Providence, RI 02903, Phone: 401-274-1122 × 8740 Fax: 401-453-7992, ; Email: vsung@wihri.org

Conflict of Interest: Dr. Vivian Sung declares that she has no conflicts of interest. Dr. James Griffith declares that he has no conflicts of interest. Dr. Rebecca Rogers declares that she has no conflicts of interest. Her disclosures include: Health publishing: McGraw Hill – Royalties from textbook. Dr. Christina Raker declares that she has no conflicts of interest. Dr. Melissa Clark declares that she has no conflicts of interest.

Compliance with Ethical Standards

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

measures demonstrated good internal consistency (coefficient alpha 0.92-0.98) and unidimensionality. There was evidence for construct validity with moderate to strong correlations with the UDI (r 's = 0.6) and IIQ (r 's = 0.6) scales. The measures were responsive to change for SUI treatment (paired t -test $p < .001$, ES range=1.3 to 2.9; SRM range=1.3 to 2.5) and OAB treatment (paired t -test $p < .05$ for all domains except Social Health Impact and Adaptation, ES range=.3 to 1.5, SRM range=0.4 to 1.0). The measures were responsive based on concurrent changes with the UDI and IIQ ($p < 0.05$). CAT versions were developed and pilot tested.

CONCLUSIONS—The UIf-item banks demonstrate good psychometric characteristics and are a sufficiently valid set of customizable tools for measuring UI symptoms and life impact.

Keywords

urinary incontinence; female; quality of life; patient-centered outcomes research; patient outcome assessments; psychometrics; item response theory; computerized adaptive test

Introduction

Female urinary incontinence (UI) is a prevalent condition affecting up to 50% of middle-aged and older women.^{1,2} It is associated with embarrassment, as well as social and functional decline. Patient report is extremely important for determining the severity of UI. It is generally accepted that physical exam and objective findings for UI may not translate into important patient impact and benefit.^{3,4} The societal and personal burden of UI lie in the major impact on a woman's function and quality of life (HRQL); therefore, high-quality, meaningful patient-reported outcomes (PROs) are paramount for patient-centered outcomes research (PCOR), comparative effectiveness research (CER) and to inform healthcare delivery and policy in this field.

All current UI PROs have been developed and analyzed in accordance with classical test theory psychometrics and are therefore subject to several limitations including 1) increasing reliability and precision requires creating longer scales; 2) comparing scores between patients and studies often requires valid scores from the exact same measure; 3) precision (a.k.a. reliability) is fixed for all scores; 4) mixed item formats and constructs complicate interpretation;⁵ and 5) they lack sensitivity to detect change in individual symptoms because they cannot be tailored. As a result, there are currently a plethora of UI PROs that are subtly different but not directly comparable.⁶ Although there have been significant strides in the development of UI PROs for symptoms and HRQL,^{3,7,9} comprehensive assessment imposes a considerable time and effort burden on patients. Researchers are often challenged between asking "enough questions" to be comprehensive versus asking "too many questions" which can be burdensome.

The NIH Patient-Reported Outcomes Measurement Information System (PROMIS®) was created to develop standardized, precise and valid measures and to advance PRO assessment using item response theory (IRT) psychometrics and computerized adaptive testing (CAT).¹⁰ CAT is an efficient survey administration technique that can improve the precision and efficiency of PRO measurement by selecting items that are most relevant to an individual patient. Questions for CATs are drawn from an "item bank", which contains questions

focused on a single specific content area and are built on the principles of IRT, which links items to target concepts using calibrated mathematical relationships. In brief, IRT includes a collection of mathematical models that provide psychometric information (e.g., item difficulty and discrimination) to allow scaling (a.k.a. “calibration”) of items for specific health domains.¹¹ Data from PROMIS® supports reduced measurement error, greater precision, reduced burden, and reduced sample size requirements with IRT-calibrated measures compared to traditional paper-and-pencil measures.¹² Although PROMIS® is comprehensive, it does not fully address the needs of women with UI.¹³ Therefore, the overarching goal of our research was to utilize this innovation for improving UI measurement and to develop CATs that could measure UI symptoms and life impact efficiently. However, in order to develop a CAT, part of the process is the development and calibration of large banks of items for each domain. The objective of this current paper is to describe the development and field-testing of 7 item banks that were designed to measure domains identified as important to women with UI (UI in females, or from here on referred to as “UIf”-item banks). We also describe the calibration process and validation properties of the UIf-item banks.

Material and Methods

Our multi-disciplinary team followed a rigorous protocol consistent with methodology used in development of PROMIS® instruments. All participants provided informed consent. The Institutional Review Boards for Women and Infants Hospital (WIH) of Rhode Island and the University of New Mexico School of Medicine (UNM) approved this study. Table I presents an overview of our methods.

Development of the items

Item binning and winnowing—Item development for the UIf-item banks started with an extensive literature review of published questionnaires for UI yielding ~360 items. This review of extant items and patient focus groups helped to inform a previously published conceptual framework of patient-important outcomes in women seeking care for UI.¹³ We developed item banks through a binning and winnowing process that included item classification and selection, item revision and generation, item testing in patient cognitive interviews, item review by a range of experts (psychometricians, clinical researchers, clinicians who take care of UI and a qualitative methodologist), and final item revision.

In brief, the items were categorized (“binned”) into conceptually similar groups, based on our framework. Next, highly redundant items within bins were sorted out (“winnowed”) by the members of the author panel, selecting the items with the clearest expression of the content. Study and discussion of bins and items sometimes led to item modification for clarity, to ensure consistency in recall periods and response options, and to ensure items would be appropriate for CAT. Several new items were written to fill conceptual gaps. Based on our previous work, all items were written to assess current symptoms.¹³ Five-point response scales were used to increase sensitivity and reduce ceiling effects.¹² All UIf-scales are scored such that a higher score represents a higher level of symptom or concept of interest. For example, a higher score on the stress UI scale means more stress UI symptoms.

After a complete round of winnowing we selected the top seven outcome domains that ranked the highest in importance by patients participating in our previous focus groups: 1) Stress UI symptoms; 2) Overactive bladder (OAB) symptoms; 3) Urinary frequency symptoms; 4) Physical Health Impact; 5) Social Health Impact; 6) Emotional Health Impact; and 7) Adaptation. This resulted in a total of 171 items to undergo additional testing (Table II).

Face and content validation: Expert review and cognitive interviews

The items were reviewed by an external expert panel consisting of 2 experts in Urology/Urogynecology and 1 expert in Behavioral Medicine to assess face and content validity, comprehensiveness, item wording, and the wording of response options. Items were revised as needed based on expert input.

Each item was then assessed through 5 or more “think aloud” cognitive interviews with an experienced interviewer. The interview participants included English-speaking women who were seeking care for bothersome UI at WIH. Participants were reimbursed \$25. A cognitive interview script was developed to evaluate 1) the ability of patients to comprehend each question, 2) processes used by respondent to retrieve relevant information from memory, 3) decision processes, such as motivation and social desirability, 4) response processes, 5) relevance and importance of each question, and 6) comprehensiveness of our framework.¹⁴ If an item was revised based on participant input, it underwent additional 3-5 interviews assessing the qualities listed above.

Psychometric evaluation

After determining face and content validity, 168 items underwent psychometric testing. Two primary goals of this testing was to help distill item banks (remove poorly functioning items), and to calibrate the item banks. Calibration is a process in which items are graded by “difficulty” or “severity” and is necessary for the development of CAT algorithms. The process of calibration requires a large sample of women to respond to the items. The goal of calibration is to establish an item bank from which a CAT can draw upon to efficiently assess the individual patient using as few items as possible. Computer adaptive testing is very brief, requiring between 4-12 items per domain. So, for example, overactive bladder symptoms can be tested using 4-12 items, with items being individually tailored to the patient based on their responses to previous items. The creation of these CATs is the overarching goal of the project.

Study sites, subjects and sample size—A sample of 750 women seeking care for bothersome UI was recruited from 2 sites: WIH and UNM. Adult women (age ≥ 21) reporting bothersome UI and who were able to read English were eligible. We decided a priori to include one third women with stress UI (SUI), one third with OAB, and one third with mixed UI to ensure good representation from all 3 clinical conditions.

For item bank development, 500-1,000 subjects is generally sufficient,¹¹ and a general guideline is that 4-5 subjects per item (between 672-840 people to test 168 items) are needed

to obtain stable estimates. Therefore, we estimated that 750 women would be adequate for estimating stable model parameters needed for calibration.

Data collection from UI patients

Patients completed questionnaires detailed below. A subset of 40 women undergoing UI treatment were asked to repeat the measures 8 weeks post-treatment to assess responsiveness to change.

Preliminary Ulf-item banks—We developed a web-based questionnaire for the Ulf-banks using the PROMIS® Assessment Center (www.assessmentcenter.net), a public online research management tool that allows investigators to set up a study-specific website to collect data using either non-PROMIS or PROMIS instruments. To maintain a reasonable level of burden, each woman completed 6 out of the 7 banks for this testing (Table II). 249 women with stress UI, 249 with OAB, and 252 with mixed UI were included; therefore, each item was tested on 500-750 women. Once the testing phase is complete and CATs and short forms are developed, women would respond to all 7 domains in a clinical or research setting. It should be noted that these preliminary questionnaires included all items in the item banks which is required for calibration, but that based on this study, CATs have been calibrated to enable much shorter testing in the future.

Legacy UI questionnaires—Legacy measures refer to widely-used UI measures, which again are all static, fixed measures. For this study, legacy measures included the Urogenital Distress Inventory (UDI)⁸, Incontinence Impact Questionnaire (IIQ)⁸, the Incontinence Severity Index (ISI),¹⁶ Patient Global Impression of Severity (PGI-S)¹⁷ and Improvement (PGI-I)¹⁷ on paper and all items in these scales were included in our binning, winnowing, and revision process. The UDI has 3 subscales with 19 items: stress UI, irritative, and obstructive. The IIQ has 4 subscales with 30 items covering physical activity, social relationships, travel, and emotional health. The UDI and IIQ are typically administered together to cover UI-specific symptom and quality of life questions. UI legacy measures were selected based on their psychometric properties and common use in the field.³

Demographic and clinical data—We obtained information on patient demographics and clinical UI diagnosis.

Statistical methods

Data were analyzed using a mixed-methods approach combining classical test theory, IRT psychometrics, and confirmatory factor analysis. We also used exploratory factor analysis in some cases to explore possible dimensionality (see below).

Dimensionality—In all, we examined seven separate constructs. All items were written with a single construct in mind; we refer to these groups of items as *item banks*. The goal was to select items from these banks for CAT or the creation of short scales that could be administered to patients. To conduct IRT analyses, we assume that each item bank reflects a single dimension. For example, the items grouped together as “emotional function” are reflective of a single construct, as opposed to more than one construct. To test whether each

individual item bank measured a single construct (a.k.a., a single dimension or a single factor), we subjected each item bank to a parallel analysis^{18,20}, in which observed eigenvalues are compared to eigenvalues based on randomly-simulated data. Even in the case of multidimensionality, one can assess whether items are sufficiently unidimensional for IRT. To accomplish this, we conducted exploratory factor analysis (EFA) using a bifactor rotation criterion^{21,23}, which assumes one common factor (or one common dimension) among the items. Based on this EFA with bifactor rotation, we then calculated *expected common variance* (ECV) as a measure of unidimensionality, which is the proportion of common variance that is due to the general factor. In this case, the general factor would be the target dimension of interest (e.g., emotional function). Simulation research supports the use of ECV as an index of unidimensionality whereas traditional fit indices often yield misleading results for this purpose.²⁴ After a final list of items was chosen we calculated Cronbach's alpha for each scale. A Cronbach's alpha of 0.70 was considered necessary to meet recommended standards. The urinary frequency scale was not examined for dimensionality because it is only four items long.

Construct validity—Construct validity was based on comparisons with existing legacy PROs. Construct and concurrent validity were assessed using Pearson correlation coefficients for continuous scores and Spearman's rank correlation coefficients for ordinal scoring. We hypothesized that the UIf-symptom scores (SUI and OAB symptom banks) would be positively correlated with the UDI and SSI. We hypothesized that the UIf-function scores (Physical Health Impact, Social Health Impact, Emotional Health Impact, Adaptation) would be positively correlated with the IIQ.

Responsiveness—We assessed responsiveness properties in 20 women undergoing midurethral sling for SUI and 20 women undergoing OAB medical therapy. Responsiveness is the extent to which a PRO measure detects change in a patient's condition over time. We assessed UIf-scores from baseline and 8 weeks using within-subject *t*-tests for statistical significance. We estimated pre-to-post effect sizes using Cohen's²⁵ *d* and Guyatt's effect size (ES).²⁶ Effect sizes and standardized response means (SRMs) were interpreted as small (0.20), moderate (0.50) or large (>0.80) following guidelines proposed by Cohen.²⁷

Item response theory modeling—IRT is used as an analysis approach to determine the relationship between each item and the dimension of interest. Before calibrating items, corrected item-total correlations were examined to screen for any items that might not perform effectively in a one-dimensional IRT model. Item total correlations are defined as the correlation between the individual item and the total score of the remaining items. Once we confirmed the appropriate items, IRT models were fit to the item-level data. These analyses set the stage for CAT development. Items with corrected item-total correlations < .3 were not included in further IRT analyses.

Item response theory was used to calculate final calibration statistics using IRTPRO v. 2.1.²⁸ Additional dimensionality analyses were completed using Mplus software, version 7.11.²⁹ Each bank was examined for differential item functioning (DIF), using the lordif package in R^{30,31}, across the 3 UI conditions to ensure that these new scales would have consistent measurement properties across various urologic diagnoses. DIF occurs when one group

responds differently due to some reason other than true differences in that measured construct and can be a serious threat to validity. In addition to flagging items with potential DIF using lordif³¹, we also examined multiple group analyses in Mplus to inspect the magnitude of potential differences. Any items flagged for DIF were carefully reviewed by the authors and dropped if their content was deemed to be possibly problematic.

For six of the seven banks, a two-parameter graded response model was used to link items to the concepts of interest with the form:

$$T = \frac{1}{1 + \exp[-a(\theta - b)]}$$

Where T is a probability, a is slope parameter (a.k.a. discrimination) and b is a threshold parameter (a.k.a. threshold), and θ is the level of symptomatology on a standardized metric. Each item has 5 response options and therefore there are four thresholds per item. Maximum likelihood estimation was used to calculate all item parameters using the Bock-Aitkin Expectation-Maximization Algorithm.

For the urinary frequency bank, only four items were retained in the final scale, so it was decided to score this scale as a short form in which each item contributed equally to the score (summed and divided by 4). All of the banks were calibrated on a theta metric, for which the mean is set to 0 and the variance is set to 1. In Assessment Center, the scores are rescaled to what is termed a T metric, with a mean set to 50 and standard deviation set at 10, reflecting the average score of women seeking treatment for UI. This metric can provide a useful reference for future UI studies. A more detailed discussion on the assessment of item information and CAT can be found elsewhere.³² CATs were scored using the *expected a posteriori* (EAP) method, which is a commonly used to score questionnaire items that have been IRT-calibrated; technical details are beyond the scope of this manuscript, but are readily available in the literature.³³

Results

Face and content validity

Based on expert review and cognitive interviews 3 items were dropped due to redundancy leaving 168 total items that underwent additional field testing. Table II presents the number of items before and after testing.

Field testing

Participant characteristics—Table III presents the characteristics of the study population. A total of 475 women were enrolled from WIH and 275 were enrolled from UNM.

Validity properties of the item banks

i. Descriptive: There was a very low missing response rate (<2%) for all UIF-banks.

ii. Dimensionality analyses: The banks for physical health impact, social health impact, and emotional health impact were all unidimensional according to parallel analysis, supporting the use of unidimensional IRT. For the stress UI, overactive bladder, and adaptation banks, we conducted exploratory factor analysis with a bifactor rotation in Mplus²². Because the stress UI, overactive bladder, and adaptation banks showed evidence of possible multidimensionality, we calculated ECV to determine how much common variance in these measures was due to the general factor (the dimension of interest). For stress UI, overactive bladder, and adaptation, ECV was .77, .69, and .89 respectively. Because these values suggest a high level of general-factor variance, we concluded that each was sufficiently unidimensional. The results of the factor analyses were carefully inspected to ensure that any additional “factors” did not represent a new concept of interest, but this was not the case. The additional factors tended to be items that had similar wording (in the overactive bladder items “*I am bothered* because I am constantly in the bathroom urinating during the day” and “*I am bothered* because I have to urinate more often than I think I should”; emphasis added), rather than distinct concepts. Although it would have been possible to achieve even higher levels of ECV by removing certain items, this could come at the cost of dropping items with relevant content. Thus, we erred on the side of retaining relevant items. After fitting a one-dimensional model to each item bank, IRT statistics confirmed a strong relationship between each item and the concept of interest (see Supplementary Tables VI-XI).

iii. Reliability: Consistent with our unidimensional conceptualization of the 7 banks, the internal consistency (measured by coefficient alpha) ranged from 0.92-0.98, indicating that the items in each bank had a high level of construct homogeneity. Cronbach’s alpha also tends to be larger as there are more items, but for computer adaptive testing not all items are administered.

iv. Construct validity: Tables IV and V present correlation data for each of the 7 UIF-item banks (raw scores were summed for each bank) and the total UDI, IIQ and ISI scores as well as for the UDI-stress subscale score for SUI patients, and the UDI-irritative subscale score for OAB patients. We found moderate to strong correlations in the appropriate hypothesized direction between our 7 banks and UDI symptom and IIQ quality of life impact scales for stress UI, OAB, and mixed UI patients (data for mixed UI available upon request). There were large correlations between the UIF-SUI and UIF-OAB symptom banks and the total UDI (r ’s = 0.6) and total IIQ scores (r ’s = 0.6). The correlations between the UIF-SUI and OAB banks were stronger for each corresponding UDI-subscale score (stress subscale r =0.62 and irritative subscale r =0.74) than the total UDI score. Also the UIF-impact banks (Physical Health, Social Health, Emotional Health and Adaptation) were more strongly correlated with the IIQ scores which measure UI-HRQL than the UDI scores which measure symptoms.

v. Responsiveness: Our data support that UIF-item banks are highly responsive to change for women undergoing SUI surgical treatment (paired t -test p < .001 for all banks, ES range=1.3 to 3.0; SRM range=1.3 to 2.5) and OAB treatment (paired t -test p < .05 for all banks except Social health impact and Adaptation, ES range=.3 to 1.5, SRM range=0.4 to 1.1 except for Adaptation).

All UIf-banks were highly responsive based on concurrent changes with UI legacy measures ($p < 0.05$) and were at least moderately correlated ($r^2 \geq 0.4$) with patient global impression measures. These data support that the UIf-banks have strong validity properties and are responsive to change.

Item bank calibration analyses

All items were calibrated using the two-parameter logistic graded response model.²⁸

Our IRT parameters for each bank are provided in our supplementary tables (Tables VI-XI), for which larger values of the discrimination parameter indicate a larger strength of relationship between item and latent construct (dimensions of urinary symptoms in this case). Because our dimensionality analyses supported a sufficient unidimensionality, we next calibrated each item bank using IRT calibration. Some items were omitted from calibration analyses due to possible DIF, low corrected item-total correlations, or content-related concerns. The number of items retained for each bank is shown in Table II. In general most items were retained. The GRM fit well and item calibrations (slope and category threshold locations) enabled us to align the items in each of the 7 banks along a continuum from a low score (less severe UI symptoms or less severe impact of UI on life) to high score (more severe symptoms or impact).

Associated CAT algorithms were able to be developed for each item bank, except for Urinary Frequency Symptoms due to the low number of retained items (4). This will be administered as a 4-item short form. We piloted the UIf-CAT (which includes CAT versions for the remaining 6 domains) on 20 women, who received the six CATs and one short form as a series, administered seamlessly without any noticeable transition between domains for the responder through the Assessment Center. The mean number of items required for each domain ranged from 4-12 items and the time needed to complete the UIf-CAT in total ranged from 5-20 minutes. In other words, even though the largest bank (the OAB bank) has 39 items, the mean number of items required with CAT was 12 for that bank.

Discussion

The UIf-banks were developed using rigorous methods consistent with the NIH PROMIS® initiative. These provide a multidimensional assessment of UI applicable to the three most common female UI conditions. The banks satisfy the assumptions and requirements necessary for IRT modeling and item-bank application and demonstrate good face, content, and construct validity properties as well as evidence of responsiveness. Using these item banks we have developed associated CAT tools. This a paradigm shift in UI assessment: instead of administering a fixed number of items that provide limited information for any given individual, we can adaptively administer a varying number of items targeted to the individual's specific level of impairment with the intent of improving precision and efficiency of UI PRO measurement.

We acknowledge the complexities of developing item banks and associated CATs. There is substantial effort and advanced psychometric skills and specialized software required. However, once developed, the CAT algorithm will select items and estimate scores without

any special requirements. Therefore, patients' experience in completing questions using a CAT is not different from responding to more traditional questionnaires, other than the requirement of completing them on the computer. Although some may argue that certain patient populations (such as older patients) may have difficulty completing questionnaires on the computer, our field testing does not support this. Our diverse population had a wide age range and it was rare for participants to have problems completing the questions on the computer.

Our correlations between the UIf- UI symptom banks (SUI symptom bank and OAB symptom bank) and the corresponding UDI-symptom specific domain (UDI-stress subscale and UDI- irritative subscale) were strong, while the UIf-impact banks were moderately correlated with the UDI, but strongly correlated with the IIQ. This is to be expected since the UDI measures symptom bother whereas the IIQ measures quality of life impact. A woman can be highly bothered by her symptoms, but may not let it affect her physical or social activities for example.

There are limitations to our study. We only included women seeking care for UI to develop our items and to calibrate our item banks which may limit external validity in less severely affected patients who are not seeking care. However, theoretically one of the most relevant advantages of CATs is that they can essentially eliminate floor and ceiling effects by applying items tailored to the test taker. In addition, computers must be available to administer a CAT and steps may remain before it can be adopted in routine clinical care and research. Although we calibrated 7 of the most highly ranked UI outcome domains based on patient input, there are still outcome domains which may be important to women that we were not able to calibrate due to feasibility. Also, the UIf-CATs are currently still undergoing additional patient testing to compare their efficiency and precision to UI legacy questionnaires.

The strengths of our study include that our calibration sample is based on a very large clinical sample from two specialty Urogynecology clinics in different geographic regions of the country with a diverse patient population. Also, this is the novel development of a CAT for female UI that can be the beginning of a paradigm shift in the way we measure PROs. Improved PRO instruments in UI can lead to improvement in CER and patient care by increasing patient-centeredness and the efficiency and statistical power of studies while decreasing patient burden.

Conclusion

In summary, we have presented detailed methods for migrating from fixed-format paper-and-pencil questionnaires to a promising adaptive method to measure patient-reported UI symptoms and impact. Further work continues on evaluating the performance of the UIf-CAT measures and establishing the responsiveness and minimum important differences for the UIf-item banks and associated CATs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors wish to acknowledge Drs. Seung Choi, Matthew D. Barber, Linda Brubaker, and David Williams for their contributions to this project.

Funding support: Dr. Sung and this work are supported by grant K23HD060665 and R21HD069962 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

This research was also supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University. Recurring funding for the SSCC is provided by Office of the President, Weinberg College of Arts and Sciences, Kellogg School of Management, the School of Continuing Studies, and Northwestern University Information Technology.

References

1. Nygaard I, Barber MD, Burgio KL, et al. Prevalence of symptomatic pelvic floor disorders in US women. *JAMA*. 2008; 300:1311–6. [PubMed: 18799443]
2. Dooley Y, Kenton K, Cao G, et al. Urinary incontinence prevalence: results from the National Health and Nutrition Examination Survey. *J Urol*. 2008; 179:656–61. [PubMed: 18082211]
3. Abrams P, Andersson KE, Birder L, et al. Fourth International Consultation on Incontinence Recommendations of the International Scientific Committee: Evaluation and treatment of urinary incontinence, pelvic organ prolapse, and fecal incontinence. *Neurourol Urodyn*. 29:213–40. [PubMed: 20025020]
4. Staskin D, Tubaro A, Norton PA, Ashton-Miller JA. Mechanisms of continence and surgical cure in female and male SUI: surgical research initiatives. *Neurourol Urodyn*. 30:704–7. [PubMed: 21661017]
5. de Ayala, RJ. *The Theory and Practice of Item Response Theory*. The Guilford Press; New York: 2009.
6. Cotterill N, Goldman H, Kelleher C, Kopp Z, Tubaro A, Brubaker L. What are the best outcome measures when assessing treatments for LUTD?--achieving the most out of outcome evaluation: ICI-RS 2011. *Neurourol Urodyn*. 31:400–3. [PubMed: 22415657]
7. Barber MD, Kuchibhatla MN, Pieper CF, Bump RC. Psychometric evaluation of 2 comprehensive condition-specific quality of life instruments for women with pelvic floor disorders. *Am J Obstet Gynecol*. 2001; 185:1388–95. [PubMed: 11744914]
8. Shumaker SA, Wyman JF, Uebersax JS, McClish D, Fantl JA. Health-related quality of life measures for women with urinary incontinence: the Incontinence Impact Questionnaire and the Urogenital Distress Inventory. Continence Program in Women (CPW) Research Group. *Qual Life Res*. 1994; 3:291–306. [PubMed: 7841963]
9. Rogers RG, Kammerer-Doak D, Villarreal A, Coates K, Qualls C. A new instrument to measure sexual function in women with urinary incontinence or pelvic organ prolapse. *Am J Obstet Gynecol*. 2001; 184:552–8. [PubMed: 11262452]
10. Cella D, Riley W, Stone A, et al. The Patient Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*.
11. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. *Qual Life Res*. 2007; 16(Suppl 1):95–108. [PubMed: 17530450]
12. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. *Arthritis Res Ther*. 13:R147. [PubMed: 21914216]
13. Sung VW, Marques F, Rogers RR, Williams DA, Myers DL, Clark MA. Content validation of the patient-reported outcomes measurement information system (PROMIS) framework in women with urinary incontinence. *Neurourol Urodyn*. 30:503–9. [PubMed: 21400574]

14. Jabine, T.; Straf, JT.; Tourangeau, R. Cognitive aspects of survey design: Building a bridge between disciplines. National Academy Press; Washington, DC: 1984.
15. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res.* 2007; 16(Suppl 1):5–18. [PubMed: 17375372]
16. Sandvik H, Espuna M, Hunskaar S. Validity of the incontinence severity index: comparison with pad-weighing tests. *Int Urogynecol J Pelvic Floor Dysfunct.* 2006; 17:520–4. [PubMed: 16547687]
17. Yalcin I, Bump RC. Validation of two global impression questionnaires for incontinence. *Am J Obstet Gynecol.* 2003; 189:98–101. [PubMed: 12861145]
18. Hayton JC, Allen DG, Scarpello V. Factor retention decision in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods.* 20014; 7:191–205.
19. Horn JL. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika.* 1965; 30:179–85. [PubMed: 14306381]
20. O'Connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers.* 2000; 32:396–402.
21. Holzinger KJ, Swineford F. The bi-factor method. *Psychometrika.* 1937; 2:41–54.
22. Jennrich RI, Bentler PM. Exploratory bi-factor analysis. *Psychometrika.* 2011; 76:537–49. [PubMed: 22232562]
23. Jennrich RI, Bentler PM. Exploratory bi-factor analysis: The oblique case. *Psychometrika.* 2012; 77:442–54.
24. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement.* 2013; 73:5–26.
25. Cohen, J. *Statistical Power Analysis for the Behavioural Sciences.* 2 edition. Academic Press; New York: 1988.
26. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987; 40:171–8. [PubMed: 3818871]
27. Subak LL, Adamson GD, Boltz NL. Therapeutic donor insemination: a prospective randomized trial of fresh versus frozen sperm. *Am J Obstet Gynecol.* 1992; 166:1597–604. discussion 604–6. [PubMed: 1615966]
28. Cai, L.; Thissen, D.; du Toit, S. IRTPRO 2.0 for Windows. Scientific Software International; Chicago, IL: 2011.
29. Muthen, LK.; Muthen, BO. *Mplus User's Guide.* Seventh Edition. Muthen & Muthen Copyright; Los Angeles, CA: 1998-2012.
30. Computing RCTRFfS. R: A language and environment for statistical computing. Vienna, Austria: 2014. <http://www.R-project.org>
31. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. *J Stat Softw.* 39:1–30. [PubMed: 21572908]
32. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care.* 2000; 38:II28–42. [PubMed: 10982088]
33. Thissen, D.; Wainer, H. *Test Scoring.* L. Erlbaum Associates; Mahaway, N.J.: 2001.

Table I

Steps for developing UIf-item banks, CAT and short forms

Phase	Overarching goal	Tasks
Phase I	Identify concepts and develop patient-centered conceptual framework	<ul style="list-style-type: none"> • Patient focus groups ($n = 35$ patients) • Comprehensive literature search for extant surveys • Item classification, qualitative item review
Phase II	Development of item banks	<ul style="list-style-type: none"> • Binning and winnowing of items • Item revision and generation based on conceptual framework
Phase III	Qualitative review and testing of item banks	<ul style="list-style-type: none"> • Expert review & rating for clinical relevance, refinement ($n = 7$ experts) • Patient cognitive interviews and item rating ($n = 40$ patients)
Phase IV	Calibration testing of item banks	<ul style="list-style-type: none"> • Test UIf-item banks and validity measures • Included 750 women seeking clinical care for UI
Phase V	Creation of final item banks, short forms, and CAT algorithms	<ul style="list-style-type: none"> • Validity testing with UI legacy measures • Evaluation and modification of framework • IRT analyses
Phase VI	CAT Pilot testing	<ul style="list-style-type: none"> • CAT versions pilot-tested on 20 women

Table II

Calibrated UIf-item banks (N=750) *

Item bank	5-point response	Women who completed each bank	No. items before calibration	No. items after calibration *
UIf-Symptom banks				
Stress UI symptoms	Intensity, frequency	SUI, Mixed UI	33	29
Overactive bladder symptoms	Intensity, frequency	OAB, Mixed UI	46	39
Urinary frequency	Intensity, frequency	SUI, OAB, Mixed UI	7	4
UIf-Function banks				
Physical health impact	Intensity [†]	SUI, OAB, Mixed UI	19	16
Social health impact	Frequency	SUI, OAB, Mixed UI	22	19
Emotional health impact	Frequency	SUI, OAB, Mixed UI	27	27
Adaptation	Frequency	SUI, OAB	17	16
Total items			171	150

* The number of items retained in the final bank is the pool of items after dropping items with low corrected item-total correlations, or possible differential item functioning.

[†] Physical health impact items included response option "Did not do for other reasons"

249 women with SUI, 249 with OAB and 252 with Mixed UI

Table III

Demographic and clinical description of study population (N=750)

Variable	
Age (mean years, SD)	55.1 (13.9)
Range	21-91 years
Race *	
White	626 (83.8)
Black	15 (2.0)
American Indian/Alaskan	30 (4.0)
Asian	3 (0.4)
Native Hawaiian/Pacific	4 (0.5)
Other	92 (12.3)
Hispanic	163 (21.8)
Not Hispanic	587 (78.2)
Highest grade completed **	
<high school	27 (3.6)
High school	160 (21.4)
>high school, but no additional degree	303 (40.5)
Bachelors/graduate	259 (34.6)
Clinical diagnosis	
SUI only	249 (33.2)
OAB only	249 (33.2)
OAB-wet	193 (77.5) †
OAB-dry	56 (22.5) †
Mixed UI	252 (33.6)
Mixed, SUI and OAB-dry	28 (11.1) ‡
Mixed, SUI and OAB-wet	220 (87.3) ‡
Mixed, cannot tell	4 (1.6) ‡

* Some patients reported more than 1 race

** One missing response

† Proportion of OAB only patients

‡ Proportion of Mixed UI patients

Table IV

Correlation between UIf-item banks and UI legacy instruments for stress UI patients *

Stress Urinary Incontinence patients (N=249)				
	UDI	UDI-stress subscale	IIQ	ISI
SUI Symptom Bank				
Coefficient	0.59	0.62	0.56	0.59
p-value	<.0001	<.0001	<.0001	<.0001
Urinary Frequency				
Coefficient	0.57	0.19	0.52	0.21
p-value	<.0001	0.003	<.0001	<0.001
Physical Health Impact				
Coefficient	0.52	0.43	0.63	0.40
p-value	<.0001	<.0001	<.0001	<.0001
Social Health Impact				
Coefficient	0.52	0.29	0.78	0.36
p-value	<.0001	<.0001	<.0001	<.0001
Emotional Health Impact				
Coefficient	0.51	0.37	0.74	0.43
p-value	<.0001	<.0001	<.0001	<.0001
Adaptation				
Coefficient	0.51	0.35	0.65	0.41
p-value	<.0001	<.0001	<.0001	<.0001

* UDI = Urogenital Distress Inventory

IIQ = Incontinence Impact Questionnaire

ISI = Incontinence Severity Index

Table V

Correlation between UIf-item banks and UI legacy instruments for overactive bladder patients *

Overactive bladder patients (N=249)				
	UDI	UDI-irritative subscale	IIQ	ISI
OAB Symptom Bank				
Coefficient	0.60	0.74	0.62	0.49
p-value	<.0001	<.0001	<.0001	<.0001
Urinary Frequency				
Coefficient	0.44	0.49	0.45	0.22
p-value	<.0001	<.001	<.0001	0.0008
Physical Health Impact				
Coefficient	0.52	0.38	0.68	0.33
p-value	<.0001	<.0001	<.0001	<.0001
Social Health Impact				
Coefficient	0.49	0.40	0.78	0.30
p-value	<.0001	<.0001	<.0001	<.0001
Emotional Health Impact				
Coefficient	0.51	0.50	0.72	0.34
p-value	<.0001	<.0001	<.0001	<.0001
Adaptation				
Coefficient	0.58	0.61	0.69	0.55
p-value	<.0001	<.0001	<.0001	<.0001

* UDI = Urogenital Distress Inventory

IIQ = Incontinence Impact Questionnaire

ISI = Incontinence Severity Index