3. RESEARCH STRATEGY
A. SIGNIFICANCE

A.1. Female urinary incontinence (UI): Urinary incontinence is a common condition, disproportionately affecting women over men.[6] In the U.S., the prevalence of bothersome UI is as high as 23-32% in women over 60 years.[7] UI is associated with embarrassment, social isolation, and functional decline. In 2005, over $16 billion was spent in direct costs alone for female UI.[8] It is projected that with the aging population, by year 2050, over 28.4 million women will suffer from UI, with a significant increase in the number undergoing UI treatment.[1]

A.2. Limitations on current patient-reported outcomes for UI: UI treatments are primarily aimed at improving a woman's symptoms, functioning and quality of life. Physical exam and objective findings often do not translate into important patient benefit; therefore, high-quality patient-reported outcome measures (PROs) are paramount to inform treatment progress. PROs measure patient perceptions at 4 levels of increasing complexity: symptoms, functioning, general health perceptions, and quality of life (HRQOL).[9] The inclusion of PROs at all 4 levels is now mandated by research funding bodies, ethics committees, and regulatory agencies.[2]

Monitoring PROs across multiple dimensions imposes a considerable time and effort burden upon patients. The field of UI has made significant strides in the development of disease-specific PROs[10] measuring symptoms, HRQOL and sexual function.[11-16] However, all current UI PROs are validated in accordance with Classical Test Theory (CTT) psychometrics, and are subject to the following limitations: 1) increasing reliability requires longer scales; 2) comparing scores requires patients to completely answer all items on the same questionnaire; 3) scale properties are sample dependent; 4) precision measures are fixed for all scores; 5) mixed item formats and constructs lead to an imbalance on the total score, which complicates interpretation.[17]

Due to these limitations, there are currently a plethora of UI PROs with variable items, domains, and content. UI clinical trials often use a battery of "one size fits all" surveys to assess multiple dimensions. This increases redundancy, respondent fatigue and burden which increase survey and measurement errors. Short forms can reduce burden but at the cost of reduced precision and breadth. Also, patients often differ in symptom severity and level of health, and fixed surveys may include irrelevant items for an individual. Additionally, PRO selection is often inconsistent between studies (particularly between clinical and industry researchers), reducing comparability between studies. Finally, the Food and Drug Administration has set scientific standards for PROs that include requirements for qualitative evidence that PROs cover all outcomes important from the patient perspective.[2] To date, few UI PROs are driven by a patient-based conceptual framework and there has been little documentation of content validity from the patient perspective.[18]

In summary, current UI PROs are limited by their inflexibility, scope, burden, and lack of patient-centeredness. There is a pressing need for a UI PRO assessment method that allows precise, efficient, multi-dimensional assessment that is feasible for researchers, industry, and clinicians. This advance is critical for providing evidence for the benefits of new and existing treatment interventions from the patient perspective.

A.3. Modern test theory and PRO measurement: Item response theory (IRT) psychometrics offer an alternative to inflexible questionnaires and allow adaptive testing in which individuals receive different scale items targeted to their specific symptom or impairment level. In brief, IRT includes a collection of models that provide psychometric information (including item difficulty and discrimination) to allow scaling (termed "calibration") of items for specific symptoms and health domains within a conceptual framework. This leads to the formation of "item banks", which are a collection of items that have been IRT-calibrated on a common metric and cover all aspects of a single construct (unidimensional)-(please see "Approach" for details on IRT and item banks). Item banks are content valid and have enough items to attain high measurement precision throughout the measurement range.[19] Once items are calibrated, the calibrations can be used to guide computer adaptive testing (CAT). In CAT, a seed item is delivered to the patient and based on her response, the most relevant (informative) items from that bank are selected for further administration before moving to the next dimension. Because CAT selects items that maximize information about a person's likely score, fewer items are needed to obtain a precise estimate for each domain. Two individuals taking a CAT may receive different items, but because items have been calibrated along a common metric, the scores are comparable without patients needing to answer irrelevant items. Studies comparing traditional PROs to IRT-CATs report reduced measurement error with CAT, greater precision and reduced patient burden.[20]

The NIH Patient-Reported Outcomes Measurement Information System (PROMIS) was developed to revolutionize PRO collection and reporting using IRT.[21] PROMIS objectives include: 1) develop and test generic PRO item banks, 2) create a CAT system, 3) create a public system for a common repository of items and CATs.[3] PROMIS has developed a generic conceptual framework for health and 12 generic item banks. To date, there has not been any item bank development addressing the needs of women with UI.

A.4. Applying IRT to advance PRO assessment in female UI: Although PROMIS is comprehensive, our data indicate that it does not fully address the needs of women with UI. PROMIS is testing its item banks in diverse chronic conditions, but has not focused on conditions unique to women including pelvic floor disorders. This new generation of IRT-based PRO measurement can readily address many of the limitations of our current UI PROs. The overarching goal of our research is to develop item banks and a dynamic CAT to measure patient-important UI outcomes precisely and efficiently. Continued improvement in UI health outcomes requires the evolution and advancement of high quality PRO assessment tools. The application of IRT and CAT has the potential to significantly improve scientific knowledge, PRO assessment, and have a high impact on the care of women with UI.

B. INNOVATION: Our proposed research represents a major advance in how UI PRO measures are developed, administered and collected. This is directly in line with this funding announcement, and will result in a methodological advance in standardized outcome measurement and novel research tools and technologies for UI, a chronic, debilitating condition that disproportionately affects women over men. Furthermore, the Institute of Medicine's recent 2010 consensus report, "Women's Health Research-Progress, Pitfalls and Promise," highlighted the need for research advancing PRO measurement specifically for diseases and their associated treatments that are non-fatal but result in major morbidity for women.[5]

At the end of this project, we will have new, high-quality UI PRO assessment tools that can monitor treatment outcomes within multiple patient-important dimensions. This is a major shift from traditional paper-pencil questionnaires that can be inflexible, burdensome and duplicative. We will also explore the possibility of modeling the constructs using novel hierarchical multidimensional IRT models in lieu of creating separate uni-dimensional banks to further enhance measurement efficiency. After pilot-testing our UI-CAT prototype, we will be poised to submit an RO1 to compare its performance with traditional static questionnaires and use it as a means to tailor our treatments to specific outcomes valued by individual patients.

This project is innovative in several ways. First, it is the initial study using advanced IRT methodologies to modernize PRO measures for female UI and pelvic floor disorders, ensuring that UI PROs advance to the next generation. Second, it addresses a major limitation in the inflexibility and burden of current UI measures by developing dynamic, efficient, precise, and personalized CAT measures, a distinct shift from current practice. Third, a highly precise and efficient system is likely to be accepted by patients, researchers, industry and clinical practice, which can streamline PRO measurement and increase comparability between studies. Fourth, increased precision can lead to decreased sample size requirements in clinical trials, thus also increasing efficiency of our current trial methods. Fifth, it addresses outcomes important from the patient perspective, often overlooked in traditional UI PROs. Finally, we will gain a better understanding of the multidimensional structure of our constructs with possible development of more sophisticated future banks.

C. APPROACH:

C.1. Preliminary qualitative work: Our work thus far in UI item bank development has been step-wise and consistent with the rigorous PROMIS Qualitative Item Review methodology[22] which includes: identification of extant items, item classification and selection, item revision, focus groups, cognitive interviews, and final item revision. Consistent with PROMIS, we first conducted 4 semi-structured interactive focus groups of 35 women with UI to address the lack of a conceptual framework for UI PROs (Appendix 1). All sessions were audiotaped, transcribed, coded and analyzed. Our UI focus groups confirmed the relevance of many of the generic PROMIS domains, but identified additional themes that improve the comprehensiveness of this framework. Our working conceptual framework for UI is provided in Figure 1 (Appendix 1) and includes 6 overarching domains: physical health, social health, mental health, daily life function, external mediators and UI symptoms. Exemplar emerging themes within these domains included: the distinction between ability, participation and satisfaction for physical and social function, "new possibilities", and "coping/adaptation". Other than external mediators, women ranked the remaining 5 domains as highly important outcomes for UI treatment. Although external mediators played a role in exacerbating UI impact, women did not feel that changing mediators was a primary goal or outcome of UI treatment.

The focus groups allowed us to more clearly define constructs with careful specification of subdomains that are comprehensive and important for UI. Next, we performed a literature review to identify extant UI PRO measures and classified each existing item based on domain coverage in our framework. We identified 26 "validated" UI questionnaires in the literature, with 363 items measuring symptoms or HRQOL. Of these, only 5 were developed with patient focus group or cognitive-based interview input. The surveys were variable in content and domain coverage. There were no items covering social role functioning, new possibilities, external mediators. Seldom did questionnaires distinguish between participation and satisfaction or responsibility versus discretionary physical and social activities. We will begin item bank development using these extant items (See Appendix 2 for example of how existing items are categorized into the Physical Health item bank).

Finally, because women with UI are often older than the general population, we conducted a trial to explore the feasibility of a web-based CAT in our target population. Using the PROMIS Assessment Center, we designed and implemented a web-based study in a convenience sample of 40 women seeking care for UI (AUGS Foundation Grant). We included paper short forms and CAT versions for 7 generic PROMIS domains. Women were randomized to receive the paper or the CAT version first. Women were timed while completing both versions and then completed a feasibility questionnaire. We found that all women were able to complete the CAT version and the majority (90%) favored the CAT reporting it was easier and faster to complete. Despite having 15% of women over the age of 70 years without significant electronic experience, the older population also favored the CAT. This study supports that web-based CAT administration of PROs is feasible in women seeking care for UI, including older women. Therefore, we would expect that our development and implementation of a UI-specific CAT should be well accepted and utilized by this population.

C.2. Collaborations: We have built a powerful, multi-disciplinary team of expert researchers in female pelvic floor disorders and UI, qualitative research, modern psychometrics and behavioral sciences. Please see Biosketches for specific details and roles.

C.3. Methods: We will develop and calibrate item banks for the 5 patient-important UI domains (see Table 1):

Table 1: Domains to be developed / calibrated

Item bank/domain	Exemplar content
Physical health	Participation, satisfaction, new possibilities
Social health	Participation, satisfaction, new possibilities, social relationships, social roles
Mental health	Emotional distress, cognitive function, preoccupation
Daily life function	Lifestyle, adaptation
UI symptom distress	Symptom bother

The process of developing item banks is complex and intensive. In the next 8 months prior to the start of this proposed R21, we will complete extant item selection and revision, and write new items for inadequately covered domains. Our process will ensure consistency in the style, response options, recall time frames, ease of literacy, and that items can "stand alone" to be appropriate for CAT.

The final critical steps for item bank calibration and CAT will be fulfilled through this R21 including cognitive-based interviews, field testing of item banks, and CAT algorithm development. Building on our previous work, we will use a mixed-methods approach and a combination of Classical Test Theory (CTT) and IRT methods, which are complementary when applied together. Basic item properties will be examined using traditional CTT statistics and item properties, functions and calibrations will be analyzed using IRT modeling.

C.3.a. Aim 1: To confirm the content validity of our item banks through cognitive-based interviews and expert review. (Year 1, Months 1-3 of grant)

Study design: We will develop a cognitive-based interview protocol to elicit feedback on all individual items considered for our 5 UI item banks. The cognitive-based interviews will employ a retrospective verbal probing technique, in which a participant completes a paper and pencil version of the questionnaire and a trained interviewer asks specific information for each question, as well as "think aloud" techniques.

Our protocol will review: 1) comprehension of question (what does respondent believe the question is asking?); 2) processes used by respondent to retrieve relevant information from memory (what does the respondent need to recall to be able to answer the question); 3) decision processes, such as motivation and social desirability (is respondent sufficiently motivated to accurately and thoughtfully answer question); and 4) response processes (can respondent match her response to the question's response options).[23] Our goal will

be for each preliminary item bank to undergo 5 cognitive-based interviews. If items in any bank require major revisions, the revised items will undergo an additional 3-5 cognitive-based interviews after revision.
Study population: The study population will include 25-30 women (depending on revisions) seeking care for UI at our tertiary care center. We will include 5% minority, 5% disadvantaged, and 30% older women >65 years. Women will be compensated $25 for their participation.
Analysis: Content analysis and descriptive summary statistics will be used to evaluate information gathered and to characterize the participant sample. Items will be revised based on participants' responses. Items that do not appear comprehensible or relevant after revision will be considered for elimination. Conceptual gaps will be identified and new items created to fill these gaps. This will be an iterative process. Dr. (Consultant) has specific expertise in qualitative research and cognitive-based interview methods and will be closely involved in all aspects of this study. Our cognitive-based interview findings and our final UI item banks will be reviewed by our Expert Advisory Panel. After review, discussion, and approval, the content validation process for our UI item banks will be complete and the item banks will be ready for additional field-testing.

C.3.b. Aim 2: To calibrate and field-test the item banks in 700 women with urinary incontinence recruited across two hospital settings. (Year 1-Year 2, Months 4-19 of grant)

Study design: A large study population is needed for item calibration of our 5 UI banks. We will perform a cross-sectional study of 700 women seeking care for UI across two tertiary care sites, Women and Infants Hospital of Rhode Island and University of New Mexico. Eligible women will be approached by a key personnel member, introduced to the study, and if they agree, given written informed consent. They will complete the 5 item bank questionnaires via a computer or laptop linked to a web-based questionnaire. Demographic characteristics will be collected. To determine type and severity of UI, participants will complete validated questionnaires that assess symptom bother and severity (Pelvic Floor Distress Inventory[8]) and incontinence type using the Medical, Epidemiological, and Social Aspects of Aging (MESA) questionnaire.[24]
Study population: Adult, English-speaking women (>18 years) seeking care for symptomatic UI at one of the recruitment sites will be eligible. Women who are unable to complete the questionnaires due to language or cognitive barriers will be excluded. Women treated and no longer bothered by UI symptoms will be excluded. Women will be compensated $30 for their participation in the study. No personal identifiers will be collected.
Data collection: We will develop a web-based questionnaire using the PROMIS Assessment Center, a public online research management tool that allows investigators to: 1) set up a study-specific website to collect data using either non-PROMIS or PROMIS instruments; 2) access the PROMIS library of instruments; and 3) export data at any point during accrual. Our item banks will be uploaded into the PROMIS Assessment Center. Our team already has experience with the Assessment Center functions through our previous feasibility trial (see Preliminary Data) and Dr. Choi will provide expertise in uploading our items and any trouble-shooting required.

Recruitment and data collection will occur at the two sites during a 15 month period. In FY 2009, the Women and Infants Hospital site evaluated 881 new patients for UI and the University of New Mexico evaluated 400 new patients for UI. Therefore, including follow-up patients, we anticipate that 15 months will be adequate time to meet our recruitment objectives between the 2 sites. Recruitment from both sites will increase the racial, ethnic, and geographic diversity of our sample (See Planned Enrollment Table).
Sample size requirements: There are no definitive answers regarding sample size requirements for item calibration since the ability to fit IRT models depends on the match between the items and the population. Sample size needs increase with the complexity of the IRT model and the intent of the calibration.[25] For item bank development and Graded Response Modeling (see IRT psychometric methods below), 500-1,000 subjects is generally sufficient,[19] and a general guideline is that 5 subjects per item are needed to obtain stable estimates. We estimate that each of our 5 item banks will have approximately 20 items per bank. Therefore, we estimate that 700 women will be adequate for estimating stable model parameters needed for calibration.
Overview of analysis plan: We will evaluate item and scale properties, perform factor analysis to examine underlying structures of measured constructs, evaluate the assumptions of the IRT model, examine items for differential item functioning, and finally calibrate items for CAT development.
a. Traditional descriptive statistics: Individual item analyses will include response frequency, means, standard deviation, range, and patterns and frequency of missing data. Any patterns suggesting systematic missing data will be evaluated and the content of items examined. Inter-item correlations, item-scale correlations, and coefficient alpha will be estimated. For scale analysis we will also evaluate internal consistency reliability (coefficient alpha) to describe performance of the item set.
b. IRT psychometric methods: Our step-wise analysis plan is as follows:

b.1. Evaluation of IRT model assumptions: *unidimensionality*, *local independence* and *monotonicity*.[26]

Unidimensionality assumes that a person's response to an item is accounted for by her level of that specific trait and not other factors. This will be evaluated using confirmatory factor analysis using polychoric correlations for ordinal data. Confirmatory factor analysis model fit will be assessed using multiple indices.

Local independence assumes that once the dominant factor influencing a person's response to an item is controlled, there should be no significant association among item responses. Uncontrolled local dependence among items in a CAT could result in invalid scores. Identification of local dependence includes examining the residual correlation matrix produced by the single factor confirmatory factor analysis. In addition, IRT tests of local dependence will be used. Any items that are flagged as having local dependence will be examined to evaluate their effect on IRT parameter estimates. If local dependence is apparent among clusters of items, we will closely examine items within each cluster for the presence of multidimensionality due to content overlap or relatively independent substantive dimensions. Following a content analysis of the clusters, exploratory factor analyses, including Schmid-Leiman transformations[27], will be conducted to better understand the nature of multidimensionality.[28] If significant problems with local dependence are identified, possible solutions will include item exclusion, splitting the item bank into 2 sub-banks, or using special item selection rules to ensure content balance[19] (also see Exploration of multi-dimensional IRT analyses, below).

Monotonicity assumes that the probability of selecting an item response indicating better health status should increase as the underlying (true) level of health increases. Our approach to studying monotonicity will include evaluation of graphs of item mean scores conditional on "rest-scores" and examining initial probability functions from nonparametric IRT models.

b.2. Fit IRT model to data: Once the assumptions are confirmed, we will fit appropriate IRT models to the data for both item and scale analysis and item calibration. This will set the stage for CAT development. This includes: a) estimating IRT model parameters using the Graded Response Model (GRM), a parametric, polytomous-response model that offers a flexible framework for modeling participant responses to examine item and scale properties; b) examining model fit; c) evaluating item properties using IRT category response curves and IRT item information curves; and d) evaluating scale properties using IRT scale information function curves. Multidimensional extensions of the GRM (both hierarchical and non-hierarchical) will be considered should the conceptual breadth and content heterogeneity of the constructs call for such models.

b.3. Evaluation of differential item functioning (DIF): DIF occurs when one group responds differently to an item than another group due to some reason other than true differences in that measured construct or trait.[29] DIF items are a serious threat to the validity of a scale because scores may be indicative of attributes other than those the scale is intended to measure. We will evaluate the presence, magnitude and impact of DIF. Of many DIF modeling techniques currently available, the ordinal logistic regression (OLR) provides a flexible model-based framework for detecting various types of DIF. Dr. team has further advanced this technique by creating an integrated platform to combine OLR DIF and IRT into a single software package that can evaluate statistical criteria using Monte Carlo simulations (a skill unique to Dr. team).[30] We will employ the OLR/IRT platform to examine both uniform and non-uniform DIF pertaining to age and race. In addition, because up to 50% of women with UI have another co-existing pelvic floor disorder (pelvic organ prolapse and/or anal incontinence), we will also evaluate the possible effect of co-existing pelvic floor disorders and types of UI (stress UI or urge UI) on DIF. Items essential to the measurement of the domain but displaying DIF can be recalibrated in appropriate subgroups to generate new item and trait estimates.

b.4. Exploration of multi-dimensional IRT analyses: As discussed, most IRT applications assume unidimensionality and local independence of items. There has been growing attention to multi-dimensional IRT models because unidimensionality can result in health constructs that are defined too narrowly, which may under-represent their true breadth and complexity. However, constructs that are too broadly defined may produce difficult to interpret scales due to changes in more than one dimension of the construct. Alternative modeling approaches based on hierarchical and non-hierarchical multidimensional frameworks allow for measuring multiple dimensions simultaneously. When subdomains are correlated moderately, applying a multi-dimensional IRT model can be preferable. There are three multidimensional framework options (bifactor, multi-unidimensional, and multi-dimensional IRT modeling).[28] We will explore these alternative frameworks in post-hoc simulation studies and compare them to the conventional unidimensional approach to assess measurement efficiency.

b.5. Item calibration for banking: After comprehensive review of item properties, the final selected item set will be calibrated along a metric using the GRM and CAT algorithms developed. We plan to establish one set of IRT item parameters for each item, unless we discover DIF on content-critical items that may require different calibrations for specific items based on patient characteristics. We will develop norm-based scoring

based on the study population, setting the "mean" to zero, and standard deviation (SD) to one. The resulting metric (and its T-score equivalence with mean=50 and SD=10) can provide a useful reference for future studies involving women with UI. We anticipate each UI item bank will include up to 20 calibrated items.

C.3.c. Aim 3: To develop and pilot-test a prototype web-based CAT for female urinary incontinence (Year 2, Months 20-22 of grant)

CAT integrates the advances in measurement theory and the power of computer technology to administer an optimized survey that selects questions on the basis of a person's response to previously administered questions. Highly informative questions are carefully selected, minimizing floor and ceiling effects without decreasing precision. Characteristics of an adaptive test include a calibrated item bank, an item selection procedure, a scoring method, and a criterion for terminating the test. Our previous feasibility trial using a generic PROMIS CAT supports the CAT mode is feasible and well accepted by women with UI.

We will develop a CAT specific for UI based on the following steps: Step 1 of a CAT algorithm is administration of an initial item that is informative for a person with average health (or average for that trait). The person's response to the first item is used in Step 2 to estimate the score and confidence interval. At Step 3, the computer algorithm determines if any stopping rules have been fulfilled. If not, then Step 2 is repeated for the next most informative item. The stopping rule is determined by the test administrator often based on test precision, when the confidence interval is within specified limits. The CAT would stop if a certain level of precision is achieved, or if the maximum number of items is used. Once the criterion is met, the algorithm ends assessment of that particular domain. Our UI-CAT algorithm will be developed in collaboration with Dr. who has specific expertise in CAT development. We will examine the suitability of the starting items (the first item to be selected by the CAT engine for each bank) for their content appropriateness and if needed override the selection by designating different starting items. We will also examine the optimal ordering of the five domains in the CAT administration, which may be determined on the basis of the content and expected CAT length. We will determine the optimal CAT length for each item bank considering the expected gain in precision and respondent burden.

Study design: We will pilot-test our newly developed UI-CAT on a convenience sample of 40 women with UI. Subjects will be recruited through the primary site at Women and Infants Hospital. We will take this opportunity to trouble-shoot any unexpected problems prior to larger scale administration.
Study population: Adult women (>18 years) seeking care as new or follow-up patients for symptomatic UI will be eligible. Women who are unable to complete the UI-CAT due to language or cognitive barriers will be excluded. Women will be compensated $25 for their participation in the study.
Data collection/analysis: Our UI-CAT will be developed through the PROMIS Assessment Center CAT engine. We will load the item parameters of our calibrated UI item banks and CAT algorithm into the CAT engine. We will explore feasibility, acceptability, patient attitudes, time to complete the assessment, and ease of administration of the UI-CAT. Clinical and demographic characteristics, UI-CAT scores, and patient attitudes and acceptability of the CAT will be described using descriptive statistics. Based on our previous experience for our generic CAT feasibility trial, we anticipate recruitment will be completed for this aim in 5 weeks.

C.4. Potential pitfalls: One might argue that CAT technology is too complex and the UI population will not readily accept it. However, our Preliminary Data support that women of all ages can easily complete a CAT and find it feasible and acceptable. Furthermore, it is precisely in older populations where long, burdensome questionnaires are often barriers to obtaining comprehensive PROs. In the unlikely event that we find that CAT is not feasible, we can still develop tailored short forms using our calibrated item banks, in which we select a set of items that are matched to the symptom severity level (or health status) of an individual. This will still allow increased precision, reduced burden and tailored assessments in a paper-pencil version. We can also use "Computer-assisted Personal Interviewing" or "Computer-assisted Telephone Interviewing" as alternatives. Another limitation is that it is beyond the scope of this grant to develop item banks for pelvic prolapse or fecal incontinence. We plan to build on the findings of this proposal and extend them to those populations in the future. Finally, our item banks will only be validated in English. Developing translated and culturally tailored item banks is intensive and we will consider this step in the future.