HHS Public Access

Author manuscript

Qual Life Res. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

Qual Life Res. 2017 November ; 26(11): 2973–2985. doi:10.1007/s11136-017-1640-3.

Evaluating Measurement Invariance across Assessment Modes of Phone Interview and Computer Self-Administered Survey for the PROMIS measures in a Population-based Cohort of Localized Prostate Cancer Survivors

Mian Wang[1], **Ronald C. Chen**[1,2,3], **Deborah S. Usinger**[1,2], and **Bryce B. Reeve**[1,4]

[1]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Department of Radiation Oncology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

Purpose—To evaluate measurement invariance (phone interview vs computer self-administered survey) of 15 PROMIS measures responded by a population-based cohort of localized prostate cancer survivors.

Methods—Participants were part of the North Carolina Prostate Cancer Comparative Effectiveness and Survivorship Study. Out of the 952 men who took the phone interview at 24 months post-treatment, 401 of them also completed the same survey online using a home computer. Unidimensionality of the PROMIS measures was examined using single-factor confirmatory factor analysis (CFA) models. Measurement invariance testing was conducted using longitudinal CFA via a model comparison approach. For strongly or partially strongly invariant measures, changes in the latent factors and factor autocorrelations were also estimated and tested.

Results—Six measures (Sleep Disturbance, Sleep Related Impairment, Diarrhea, Illness Impact – Negative, Illness Impact – Positive, and Global Satisfaction with Sex Life) had locally dependent items, and therefore model modifications had to be made on these domains prior to measurement invariance testing. Overall, seven measures achieved strong invariance (all items had equal

Send all correspondences to Mian Wang, 101 E Weaver St Room 214, Carrboro, NC 27510. Office Telephone: (919) 962-5369 mianwang@unc.edu.

Compliance with ethical standards

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee 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.

Conflict of interest: The authors declare that they have no conflict of interest.

loadings and thresholds), and four measures achieved partial strong invariance (each measure had one item with unequal loadings and thresholds). Three measures (Pain Interference, Interest in Sexual Activity, and Global Satisfaction with Sex Life) failed to establish configural invariance due to between-mode differences in factor patterns.

Conclusions—This study supports the use of phone-based live interviewers in lieu of PC-based assessment (when needed) for many of the PROMIS measures.

Keywords

PROMIS; measurement invariance; mode of administration; prostate cancer

Introduction

Launched in 2004, the Patient-Reported Outcomes Measurement Information System® (PROMIS®) has become internationally recognized in the health outcomes research field for the high standards of its patient-reported outcomes (PROs) measures [1]. The PROMIS measures were designed and evaluated using advanced qualitative and quantitative methods [2–3]. Evidence to date has supported the validity and reliability of the measures in different populations. There is international interest to expand the PROMIS measures in terms of translations and validation in different countries [4].

The original evaluation and calibration of the PROMIS measures was conducted with multiple populations who completed the questionnaires on a web-enabled device such as a laptop or desktop computer. However, it was recognized in clinical research studies that there is a great need to offer participants different options to complete PRO assessments. Paper-and-pencil is the traditional method and offers the convenience of collecting data without the need for an electronic device and access to the internet; however, paper-and-pencil-based assessment will not allow computerized-adaptive testing (CAT) and data entry errors may be more common than computer-based assessment. Phone-based assessment allows those without access to a web-enabled device to complete PROs, and is a viable option for those who are unable to read the questionnaire because of low literacy, non-native language, or visual handicaps.

Many of those unable to complete computer-based PRO measures because of low literacy are from vulnerable populations that are high priority for federal agencies like the USA's National Institutes of Health. Low literate patients are often excluded from PRO studies [5]. Therefore, the ability of a research study to allow participants to complete PRO measures by different assessment modes would benefit inclusion and participation rates from the vulnerable populations. Thus, the study's results will have improved generalizability.

A previous study evaluated measurement invariance of the PROMIS measures across computers, personal digital assistants (PDAs), paper-pencil, and interactive voice response (IVR) assessment modes. The study in adult populations found no statistically significant effects on mean score levels among the modes [6–7]. However, these modes did not include the option of a live phone interviewer reading the questionnaire to the participant. This assessment mode is very common in research studies as either a primary method for data

collection or back-up method of data collection when the participant may not complete the questionnaire on paper or via computer. Including another person in the process to collect PRO data, however, may have an impact on the participants' responses, especially for sensitive outcomes like depressive mood or sexual functioning. The goal of this study is to evaluate measurement invariance of the adult PROMIS measures between computer-based and phone-interview assessment modes in a population-based cohort of localized prostate cancer survivors. The PROMIS measures evaluated in this study reflect those domains relevant for assessing outcomes for men participating in a population-based prospective research study comparing alternative treatments for prostate cancer.

Method

Study/Participants

The North Carolina Prostate Cancer Comparative Effectiveness and Survivorship Study (NC ProCESS) is a population-based, observational, comparative effectiveness research study examining the impact of localized prostate cancer on the lives of men [8]. Using the rapid case ascertainment (RCA) mechanism of the North Carolina Central Cancer Registry, men with newly diagnosed prostate cancer were recruited from all 100 counties in North Carolina (NC). To participate, men had to speak English. Of the 2473 eligible men, 1419 of them enrolled in the study from January 2011 to June 2013. Additional details on NC ProCESS are described elsewhere [8]. The overall goal of this project was to prospectively evaluate cancer-specific and patient-reported outcomes of these men with newly diagnosed prostate cancer.

This study (#10-1483) was approved by the University of North Carolina Institutional Review Board.

Measures

Health-related quality of life was assessed with the NIH's PROMIS. PROMIS measures have undergone rigorous evaluation including validation in men with prostate cancer [1, 9]. PROMIS scores are normed to the US general population with mean of 50 and standard deviation of 10. Higher symptom scores reflect increased symptom burden and higher function scores reflect better functioning.

For the current study, 106 PROMIS items from the following domains/subdomains were administered (in the order they appeared on the surveys): Pain Interference, Fatigue, Depression, Anxiety, Sleep Disturbance, Sleep Related Impairment, Diarrhea, Bowel Incontinence, Physical Function, Illness Impact – Negative, Illness Impact – Positive, Interest in Sexual Activity, Therapeutic Aids, Erectile Function, Orgasm, and Global Satisfaction with Sex Life.

Design

Participants completed surveys via phone at baseline (prior to treatment) and at 3-, 12-, and 24-months post treatment initiation. This mode of administration equivalence study focused on the 24-month assessment period. At the end of the phone-based survey, participants were

asked if they had access to a web-based device and were willing to complete the same questionnaire. If yes, participants provided an email address over the phone and subsequently received an email with link to the Qualtrics (a survey platform developed by a private research software company in the United States) survey within 24 hours. They had up to 5 days post phone interview to complete the questionnaire. If they opened the survey, they had 3 days to complete the survey. Participants received a $30 gift card for each completed survey.

Analysis

Missing data handling—Missing data patterns were examined and reported. The missing completely at random (MCAR) assumption was tested using the nonparametric test of homoscedasticity [10], and multiple imputation techniques were performed except for items that had ‘not applicable’ missing responses.

Invariance tests—The invariance tests (and related procedures) for each domain were performed in the following sequence: 1) Item categories were matched to identify categories with no observed responses, and category collapsing was performed if needed; 2) Unidimensionality of the domain was evaluated using single-factor confirmatory factor analysis (CFA) within each survey mode, and necessary model modifications were made to achieve acceptable model fit; and 3) Between-mode measurement and structural invariance was examined using a series of longitudinal CFA models.

To ensure consistency with the officially published PROMIS instruments and to improve generalizability of our findings, the invariance tests included only 72 items that could be found in PROMIS short forms published online for public access (see Appendix A for exceptions). Interested readers are referred to the Health Measures website for those short forms being evaluated in the current study and their corresponding instructions (http://www.healthmeasures.net/explore-measurement-systems/promis/obtain-administer-measures). All reverse-worded items (see Appendix A) were backward reordered so that higher categories on every item indicated higher levels on the construct being measured. Also, note that only the subset of 401 men who participated in both the phone and PC surveys were analyzed using CFA.

Matching item categories: A category collapsing procedure was performed, where needed, so that an item would have the same number of categories (and thus the same number of estimated thresholds) under both survey modes. Specifically, all responses greater than or equal to any empty categories (be it under phone or PC) were collapsed downward until item responses under both surveys could be tabulated in the same consecutive ordinal pattern. For example, suppose that an item received no response in categories ‘4’ and ‘5’ over the phone, and no response in category ‘5’ over the PC. Then, under the category collapsing procedure, responses ‘4’ and ‘5’ would be recoded to ‘3’ regardless of survey mode. Category collapsing was implemented solely to establish a fair comparison between the two modes, and we did not intend to alter interpretations of the PROMIS items and their original response categories.

Single-factor CFA models: For every domain under each survey mode, the ordinal response data were analyzed using the robust weighted least squares estimator with a diagonal weight matrix (i.e., the WLSMV estimator; [11]). Three fit indices were used to determine the overall model fit: residual mean square error of approximation (RMSEA), comparative fit index (CFI), and weighted root mean square residuals (WRMR). If acceptable fit indices (RMSEA < 0.08, CFI > 0.90, and WRMR < 1; [12–14]) were obtained under one mode or both, the single-factor model would be retained as the basis for the ensuing longitudinal CFA models. Otherwise, the initial models would be further examined for violations of unidimensionality. Given that most of the studied PROMIS measures had been validated in past research, only minimal modifications were made (same toward both survey modes) to ensure acceptable model fit under at least one survey mode. The modified model then served as the basis for the subsequent invariance tests. Note, however, domains with only three items (Erectile Function and Orgasm) were directly fit using longitudinal CFA models, because fit indices were unavailable/meaningless for their just-identified single-factor models with zero degrees of freedom.

Measurement invariance: According to Chapter 5 of Little's book regarding model specifications [15] and Chapter 14 of Mplus User's Guide regarding the special treatment for ordinal items (i.e., omission of the metric invariance test when the scale is set by fixing a factor variance to one; [16], page 544), a series of longitudinal CFA models were fit (also using the WLSMV estimator) to investigate whether psychometric properties of the measures were invariant between surveys. For each domain, three measurement invariance tests were sequentially carried out:

1. Configural invariance (i.e., equivalent factor patterns):

   The same factor structure (adopted from the single-factor CFA stage) was simultaneously fit under both survey modes. The two latent factors (respectively for phone and PC) were fixed as standard normal for identification purposes, while all loadings and thresholds were freely estimated. Item/factor autocorrelations (in pairs) were freely estimated as an integral part of longitudinal CFA. Using the same cutoff criteria previously mentioned for the single-factor CFA models, a configural model with unacceptable fit would fail, indicating that the underlying factor patterns were different between surveys. Poorly fit domains would skip all following tests and be further examined for model misspecifications.

2. Strong/scalar invariance (i.e., equivalent factor patterns, loadings, and thresholds):

   In contrast to the configural model, ceteris paribus, the strong invariance model freely estimated the factor mean and variance under the PC mode, while constraining loadings and thresholds equal between survey modes. The configural and strong invariance models were then compared. To establish strong invariance, a nonsignificant scaled-χ^2 difference test statistic [17] should be observed along with a negligible decrease in the CFI index (i.e., $\Delta CFI > -0.002$, where $\Delta CFI = CFI_{strong} - CFI_{configural}$; [18]).

3. Partial-strong invariance (i.e., equivalent factor patterns with equated loadings and thresholds on selected items):

   For domains that failed the strong invariance test, partial-strong invariance models were explored by freeing some of the item equality constraints. Specifically, model results from the previous two steps were compared, and items with large discrepancies in their loadings and thresholds were flagged as candidate items. A series of partial-strong invariance models were fit by equating all items but the candidate. This process repeated until a partial-strong invariance model achieved the same (or better) model fit (i.e., nonsignificant scaled- χ^2 difference and $CFI_{partial.strong} - CFI_{configural} > -0.002$; [17–18]) in comparisons to the configural model.

Latent distributional properties and structural invariance: For domains with a strongly (or partially-strongly) invariant measure, latent factor means and variances under the PC mode were estimated, along with the between-mode factor autocorrelations. Since the two (almost identical) surveys were administered at most five days apart, we should expect minimal changes in the means and variances, but high factor autocorrelations.

Structural invariance (of means and variances) and significance of the factor autocorrelations were also tested. The freely estimated PC-mode factor mean was tested against zero (i.e., the fixed mean under the phone mode) using a z statistic. As to latent variances, we fit an equal-variance model that constrained the latent variances to one under both survey modes, and then a scaled- χ^2 difference test ($df = 1$) was conducted to compare the equal-variance model to the strong/partial-strong invariance model (which only fixed the phone-mode variance to one). Factor autocorrelation estimates were obtained from the equal-variance model and then tested against nil using a z statistic.

Software

Analyses for the current study were conducted in R version 3.3.2 [19]. Missing data mechanism was examined using the nonparametric test of homoscedasticity from the *MissMech* package version 1.0.2 [20]. Missing data imputation was performed using the multivariate imputation by chained equations (*mice*) package version 2.25 [21]. CFA models were fit using the *lavaan* package version 0.5–22 [22], and results from the analyses of multiply imputed datasets were pooled by *semTools* version 0.4–14 [23].

Results

Descriptive Statistics

Table 1 provides demographic and clinical characteristics for men who completed the 24-month assessment over the phone with the assistance of an interviewer (n = 952) and the subset of 401 men who subsequently also completed the same survey online. Within the entire cohort, about 72% were White men and about 68% had more than high school education. In contrast, the subset who completed both surveys were about 85% White men, and about 82% of them had more than high school education.

Missing Data Handling

PROMIS items were checked for missing data patterns within the subset of 401 patients who completed both surveys. The majority of PROMIS items had less than 1% true missing responses (excluding 'not applicable' responses), and sex-related PROMIS domains in general showed the highest percentages of missing. For the most skipped items, about 3.49% of patients did not respond to item *SFSAT101* on Global Satisfaction with Sex Life when asked over the phone (compared to 1.50% over the PC), and about 2.49% did not respond to item *SFORG151* on Orgasm when asked over the PC (compared to 1% over the phone). Listwise, the full completion (i.e., no skipped PROMIS items) rates were 91.27% and 75.81% respectively for the phone and PC versions. In addition, the overall percentages of missing data points were very low under both modes (0.31% with phone and 0.91% with PC).

The nonparametric test of homoscedasticity suggested that the missing mechanism within the phone dataset was not MCAR ($p < .001$). Therefore, we decided to proceed with multiple imputation under the weaker assumption of missing at random (i.e., missingness on a variable was unrelated to the missing values after controlling for the other observed variables; [24]). Separately for each survey mode, missing values on the 72 analyzed items (see Appendix A) were imputed 90 times with possible values informed by other PROMIS items. Nonetheless, items with 'not applicable' responses (all items on Erectile Function, all items on Orgasm, and two on Global Satisfaction with Sex Life) were not imputed, and they were listwise deleted during CFA.

Prior to invariance tests, incomplete cases on the two domains that skipped multiple imputation were listwise deleted, resulting in reduced sample sizes (244 for Erectile Function and 228 for Orgasm). For Global Satisfaction with Sex Life, incomplete cases within each imputed dataset were also listwise deleted based on the two variables whose 'not applicable' responses were not imputed, resulting in 219 complete cases. For the remaining 12 domains, the CFA procedures utilized the (multiply imputed) full sample of 401 patients. Results from analyses of imputed domains were pulled according to Rubin's rules [25].

Invariance Tests

Matching item categories—Frequency tables revealed that category collapsing was necessary for five items under the phone mode and six items under the PC mode (with three overlaps), across the Anxiety, Bowel Incontinence, Depression, and Diarrhea domains. All of these items received no response in their highest category, except for an Anxiety item (*EDANX40*) whose second highest category was empty under the PC mode. In general, such a floor effect could be attributed to having either high item thresholds or a positively skewed sample on the assessed domains. Empty categories were collapsed across all imputed datasets before implementing CFA.

Single-factor CFA models—During the initial stage of single-factor CFA modeling, several domains were plagued by locally dependent items (i.e., substantial residual covariances existed after item relationships had been accounted for by the latent factor). Therefore, model modifications were made on these domains, and results based on the

modified models were provided in Table 2 (with corresponding changes detailed in the table notes), along with results from domains that required no modifications. Please note that only necessary changes were made to ensure acceptable model fit under at least one survey mode. Hence, it could be possible that the same factor structure fit worse under one mode than under the other, which might be indicative of configural non-invariance.

Measurement invariance—Results from measurement invariance tests were also included in Table 2. Strong invariance was established on seven domains: Fatigue, Depression, Anxiety, Diarrhea, Bowel Incontinence, Physical Function, and Illness Impact – Negative. For these domains, their items could be fully equated between surveys without introducing bias into the estimation of the latent constructs. In other words, one survey mode could be used (theoretically speaking) in lieu of the other without distorting metric of the construct being measured.

Furthermore, measures of four domains held partial strong invariance with the following adjustments (in addition to any modifications previously made to the underlying single-factor models):

i. Sleep Disturbance: item *SLEEP109* (with backward reordered categories) was not equated. Its standardized factor loadings were $\lambda_{phone} = 0.897$ and $\lambda_{PC} = 0.976$, and its thresholds were $\boldsymbol{\tau}_{phone} = [-0.676, 0.470, 1.327, 2.055]$ and $\boldsymbol{\tau}_{PC} = [-0.945, 0.337, 1.187, 1.982]$.

ii. Sleep Related Impairment: item *SLEEP119* (with backward reordered categories) was not equated. Its standardized factor loadings were $\lambda_{phone} = 0.635$ and $\lambda_{PC} = 0.466$, and its thresholds were $\boldsymbol{\tau}_{phone} = [-0.368, 0.295, 0.808, 1.599]$ and $\boldsymbol{\tau}_{PC} = [-0.716, 0.098, 0.589, 1.079]$.

iii. Illness Impact – Positive: item *II2.a* was not equated. Its standardized factor loadings were $\lambda_{phone} = 0.624$ and $\lambda_{PC} = 0.396$, and its thresholds were $\boldsymbol{\tau}_{phone} = [-1.646, -1.423, -1.093, -0.519]$ and $\boldsymbol{\tau}_{PC} = [-1.076, -0.861, -0.661, -0.147]$.

iv. Erectile Function: item *SFEFN202* (with backward reordered categories) was not equated between modes. Its standardized factor loadings were $\lambda_{phone} = 0.930$ and $\lambda_{PC} = 0.997$, and its thresholds were $\boldsymbol{\tau}_{phone} = [-1.268, -0.714, -0.186, 0.587]$ and $\boldsymbol{\tau}_{PC} = [-0.914, -0.357, 0.176, 0.914]$.

Based on the above parameter estimates, item characteristic curves (ICC) for the four items were obtained using equations 10 through 12 in Asparouhov and Muthén [26], and the corresponding expected response functions ($\mathrm{ER} = \sum_{j=1}^{J} \mathrm{ICC}_j * j$, where j is the index for the five categories on each item) were plotted. As shown in Figure 1, expected responses of these items would differ in different magnitudes/directions between survey modes depending on a respondent's score on a domain. Nevertheless, respondents at the same level of Sleep Disturbance would generally report having worse sleep quality under the PC mode (the upper-left plot), and respondents at the same level of Erectile Function would report having better "ability to have an erection or get hard" while responding to an interviewer over the phone (the bottom-right plot).

In addition, three domains failed configural invariance because their RMSEA values exceeded the 0.08 cutoff. Further investigations revealed that:

i. For Pain Interference, items *PAININ9* and *PAININ22* locally covaried only under the phone mode.

ii. For Interest in Sexual Activity, substantial improvement in model fit was achieved after removing item *SFINT102* from the PC mode.

iii. For Global Satisfaction with Sex Life, different items (*SFSAT102* under the phone mode and *SFSAT105* under the PC mode) covaried locally with item *SFSAT101*.

Lastly, results for Orgasm were not reported because all three items had poor psychometric properties (two items had low loadings, and the other one had a negative loading), even though the global model fit indices of its configural invariance model were satisfactory.

Latent distributional properties and structural invariance—As expected, our sample were very stable on the assessed domains with invariant factor mean and variance between modes except the Illness Impact – Positive domain (see Table 3 for details). The estimated factor autocorrelations were all close to one and statistically significant, which was an indication of high test-retest reliability.

Discussion

Summary

To our knowledge, this was the first study to investigate measurement equivalence of PROMIS measures under two different modes, interviewer-assisted phone interviews versus self-report PC surveys, using a longitudinal CFA modeling approach.

Descriptive statistics revealed that a higher percentage of White men with higher education levels (in contrast to non-White men with lower education levels) agreed to participate in the PC survey. This observation was not surprising, since having internet access and being literate are requisites for completing the PC survey. Past census showed that computer use and internet access increase with higher educational attainment [27], and the White population is more literate [28] and has better access to the internet [27].

In terms of missing data patterns, we found that patients skipped more questions when completing the survey online. This finding was as expected, given that it is generally much more effortless for a patient to miss questions when self-reporting, due to either carelessness or disinclination, as oppose to responding to an interviewer. In addition, sex-related questions had the highest skip rates among the 106 PROMIS items, possibly because of patients' reluctance to provide sensitive personal information.

As to measurement invariance testing, measures of Fatigue, Depression, Anxiety, Bowel Incontinence, Physical Function, and Illness Impact – Negative were strongly invariant, holding equal factor patterns, loadings, and thresholds between surveys. For these domains, responding to a live interviewer over the phone (as an alternative to self-report online

surveys) would not adversely impact the validity of PRO assessment. From a longitudinal perspective, these measures also withstood the impact of potential short-term longitudinal confounds. Given that the two surveys were completed with up to five days in between, measurement invariance would not be established if time-related extraneous variables were altering the factor pattern and/or item properties. Moreover, for measures that held partial strong invariance, they could still be administered interchangeably between phone and PC, as long as the items that function differently between surveys (one on each domain, as found in the current study) are left unconstrained. Removal of these items from their original scales is not advised, as each item is part of a complete measure and deleting an item would jeopardize psychometric properties of the measure in its original form.

Limitations

Regarding the sample, this study was limited to English-speaking men who received their prostate cancer care in the state of North Carolina. Future research could consider improving sample diversity and results generalizability by recruiting patients with other health conditions from multiple sites across the nation/globe. Also, specific to the subset of 401 cases used for measurement invariance testing, the vast majority were non-Hispanic White or Black, and more than 80% of them had more than high school education. In addition, our sample for invariance testing was limited to those who were willing to complete the same questionnaire within a few (up to five) days. Therefore, our findings regarding invariance of the PROMIS measures should be interpreted with such limitations in mind.

Due to practical limitations when operating under the parent NC ProCESS study, participants completed the phone-interview first and the PC mode second, rather than being randomly assigned to mode. With such a repeated measures design, the effect of mode of administration and the impact of longitudinal extraneous variables on the PROMIS measures were hardly distinguishable. Thus, for measures that failed the measurement invariance tests in the current study, we were unable to rule out the possibility that they were indeed invariant across modes but affected by short-term longitudinal factors, or vice versa. Further research on cross-mode measurement invariance of PROMIS measures is needed to disentangle these possible explanations.

Conclusions

There has been a rapid increase in use of PROMIS measures in research and healthcare delivery settings. The PROMIS provides a valuable perspective from the patient of the impact of disease or treatment in terms of symptom burden or functional impact. Allowing more than one mode of administration for participants to report their health will improve the inclusion of a greater number of participants who may prefer a mode or limited in their access to a mode. More importantly, allowing phone-interviewer assessment of PROMIS measures allow those from vulnerable populations who have poor literacy skills to participate. This study supports the use of phone-based live interviewers in lieu of PC-based assessment (when needed) for many of the PROMIS measures. These results allow the data from these modes to be combined and analyzed together; however, it is encouraged to continue to evaluate measurement invariance with additional datasets.

Acknowledgments

Funding: This research was supported by grants from the Agency for Healthcare Research and Quality (HHSA29020050040ITO6) and the National Cancer Institute (R01CA174453).

References

1. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays R. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology. 2010; 63(11): 1179–1194. DOI: 10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]
2. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(5):S12–S21. DOI: 10.1097/01.mlr.0000254567.79743.e2 [PubMed: 17443114]
3. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). Medical Care. 2007; 45(5):S22–S31. Retrieved from http://www.jstor.org/stable/40221455. [PubMed: 17443115]
4. Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, Forrest CB. The case for an international patient-reported outcomes measurement information system (PROMIS) initiative. Health and Quality of Life Outcomes. 2013; 11:210.doi: 10.1186/1477-7525-11-210 [PubMed: 24359143]
5. Hahn EA, Cella D, Dobrez D, Shiomoto G, Marcus E, Taylor SG, Webster K. The talking touchscreen: A new approach to outcomes assessment in low literacy. Psycho-Oncology. 2004; 13(2):86–95. DOI: 10.1002/pon.719 [PubMed: 14872527]
6. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE Jr. Difference in method of administration did not significantly impact item response: An IRT-based analysis from the patient-reported outcomes measurement information system (PROMIS) initiative. Quality of Life Research. 2014a; 23(1):217–227. DOI: 10.1007/s11136-013-0451-4 [PubMed: 23877585]
7. Bjorner JB, Rose M, Gandek B, Stone AA, Junghaenel DU, Ware JE Jr. Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. Journal of Clinical Epidemiology. 2014b; 67(1):108–113. DOI: 10.1016/j.jclinepi.2013.07.016 [PubMed: 24262772]
8. Chen RC, Carpenter WR, Kim M, Hendrix LH, Agans RP, Meyer A-M, Godley PA. Design of the North Carolina prostate cancer comparative effectiveness and survivorship study (NC ProCESS). Journal of Comparative Effectiveness Research. 2015; 4(1):3–9. DOI: 10.2217/cer.14.67 [PubMed: 25565065]
9. Quach CW, Langer MM, Chen RC, Thissen D, Usinger DS, Emerson MA, Reeve BB. Reliability and validity of PROMIS measures administered by telephone interview in a longitudinal localized prostate cancer study. Quality of Life Research. 2016; 25(11):2811–2823. DOI: 10.1007/s11136-016-1325-3 [PubMed: 27240448]
10. Jamshidian M, Jalal S. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. Psychometrika. 2010; 75(4):649–674. DOI: 10.1007/s11336-010-9175-3 [PubMed: 21720450]
11. Muthén, BO., du Toit, SHC., Spisic, D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished technical report. 1997. Retrieved from: https://www.statmodel.com/download/Article_075.pdf
12. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999; 6(1):1–55. DOI: 10.1080/10705519909540118
13. Sass DA. Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. Journal of Psychoeducational Assessment. 2011; 29(4):347–363. DOI: 10.1177/0734282911406661

14. Yu, C-Y. Doctoral dissertation. University of California Los Angeles: 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Retrieved from https://www.statmodel.com/download/Yudissertation.pdf
15. Little, TD. Longitudinal Structural Equation Modeling. New York, NY: Guilford Press; 2013.
16. Muthén, LK., Muthén, BO. Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén; 1998–2017.
17. Satorra A, Bentler PM. Ensuring positiveness of the scaled difference chi-square test statistic. Psychometrika. 2010; 75(2):243–248. DOI: 10.1007/s11336-009-9135-y [PubMed: 20640194]
18. Meade AW, Johnson EC, Braddy PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology. 2008; 93(3):568–592. DOI: 10.1037/0021-9010.93.3.568 [PubMed: 18457487]
19. R Core Team. R Foundation for Statistical Computing. Vienna, Austria: 2016. R: A language and environment for statistical computing. URL http://www.R-project.org/
20. Jamshidian M, Jalal S, Jansen C. MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). Journal of Statistical Software. 2014; 56(6):1–31. DOI: 10.18637/jss.v056.i06
21. van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45(3):1–67. DOI: 10.18637/jss.v045.i03
22. Rosseel Y. lavaan: An R package for structural equation modeling. Journal of Statistical Software. 2012; 48(2):1–36. DOI: 10.18637/jss.v048.i02
23. semTools Contributors. semTools: Useful tools for structural equation modeling. R package version 0.4–14. 2016. Retrieved from: https://CRAN.R-project.org/package=semTools
24. Allison PD. Missing data techniques for structural equation modeling. Journal of Abnormal Psychology. 2003; 112(4):545–557. DOI: 10.1037/0021-843X.112.4.545 [PubMed: 14674868]
25. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.
26. Asparouhov, T., Muthén, B. IRT in Mplus. Version 2. Technical report. 2016. Retrieved from: https://www.statmodel.com/download/MplusIRT.pdf
27. File, T. Current Population Survey Reports. U.S. Census Bureau; Washington, DC: 2013. Computer and internet use in the United States; p. 20-568.Retrieved from: http://www.census.gov/prod/2013pubs/p20-569.pdf
28. Kutner, M., Greenberg, E., Baer, J. A First Look at the Literacy of America's Adults in the 21st Century. NCES 2006-470. National Center for Education Statistics. 2006. Retrieved from: https://nces.ed.gov/naal/pdf/2006470_1.pdf
29. Weinfurt KP, Lin L, Bruner DW, Cyranowski JM, Dombeck CB, Hahn EA, Flynn KE. Development and initial validation of the PROMIS sexual function and satisfaction measures version 2.0. The Journal of Sexual Medicine. 2015; 12(9):1961–1974. DOI: 10.1111/jsm.12966 [PubMed: 26346418]

Appendix A

List of Items Included in Confirmatory Factor Analysis.

Item Bank Version	Domain	Form	Item Names (in the Order They Appear on the Published Forms)
PROMIS Item Bank v1.0	Anxiety	Short Form 4a	{EDANX01, EDANX40, EDANX41, EDANX53}
	Depression	Short Form 4a	{EDDEP04, EDDEP06, EDDEP29, EDDEP41}
	Fatigue	Short Form 4a	{HI7, AN3, FATEXP41, FATEXP40}
	Illness Impact – Negative	Short Form 8a	{II50.a, II59.a, II67.a, II81.a, II54.a, II58.a, II71.a, II80.a}

Item Bank Version	Domain	Form	Item Names (in the Order They Appear on the Published Forms)
	Illness Impact – Positive	Short Form 8a	{II2.a, II17.a, II27.a, II36.a, II7.a, II15.a, II32.a, II35.a}
	Pain Interference	Short Form 4a	{PAININ9, PAININ22, PAININ31, PAININ34}
	Sleep Disturbance	Short Form 4a	{SLEEP109 **, SLEEP116 **, SLEEP20, SLEEP44}
	Sleep Related Impairment	Short Form 8a	{SLEEP10, SLEEP119 **, SLEEP18, SLEEP25, SLEEP27, SLEEP30, SLEEP6, SLEEP7}
PROMIS Item Bank v2.0	Physical Function	Short Form 4a	{PFA11 **, PFA21 **, PFA23 **, PFA53 **}
PROMIS Scale v1.0	Gastrointestinal Bowel Incontinence	Form 4a	{GISX45, GISX46, GISX47, GISX48}
	Gastrointestinal Diarrhea	Form 6a	{GISX38, GISX40, GISX41, GISX42, GISX43, GISX44}
PROMIS Sex Function and Satisfaction (SexFS) v1.0	Interest in Sexual Activity	Four Items	{SFINT101, SFINT102, SFINT103, SFINT104}
	Orgasm	Three Items	{SFORG101 **, SFORG150 **, SFORG151 **}
PROMIS SexFS v2.0	Erectile Function	Three Items	{SFEFN201 **, SFEFN202 **, SFEFN204 **}
	Global Satisfaction with Sex Life	Four Items	{SFSAT101, SFSAT102, SFSAT201, SFSAT105}

Note: The abovementioned PROMIS measures and their scoring manuals (except for PROMIS SexFS v2.0 measures) are available at http://www.healthmeasures.net/search-view-measures. See Weinfurt et al. [29] for the development of PROMIS SexFS v2.0. Erectile Function domain (SexFS v2.0) has four items but only three were used in the current study because the omitted item only appeared under the phone mode.

Items labeled with a '**' symbol are reverse-worded items whose higher categories indicate lower levels on the measured domain.

Appendix B

Table 1

Sample Characteristics.

Sample Characteristics at Baseline (unless noted otherwise)	All 24-month participants *N* (%)	Within the phone & PC subset *N* (%)
Total	952 (100%)	401 (100%)
Seniority (at 24-month)		
< Age 65	351 (36.87%)	148 (36.91%)
≥ Age 65	601 (63.13%)	253 (63.09%)
Race		
White	687 (72.16%)	341 (85.04%)
Black	232 (24.37%)	51 (12.72%)
Asian or Pacific Islander	2 (0.21%)	1 (0.25%)
American Indian or Alaskan Native	14 (1.47%)	2 (0.50%)
Other & Unknown	17 (1.79%)	6 (1.50%)
Highest level of education		

Sample Characteristics at Baseline (unless noted otherwise)	All 24-month participants *N* (%)	Within the phone & PC subset *N* (%)
Eighth grade or less	22 (2.31%)	1 (0.25%)
Some high school	65 (6.83%)	8 (2%)
High school graduate	201 (21.11%)	58 (14.46%)
Some college	266 (27.94%)	114 (28.43%)
College graduate	386 (40.55%)	216 (53.87%)
Unknown	12 (1.26%)	4 (1%)
Income		
Less than $10,000	41 (4.31%)	3 (0.75%)
$10,000 to $20,000	81 (8.51%)	11 (2.74%)
$20,001 to $40,000	200 (21.01%)	67 (16.71%)
$40,001 to $70,000	265 (27.84%)	135 (33.67%)
$70,001 to $90,000	119 (12.50%)	67 (16.71%)
More than $90,000	199 (20.90%)	107 (26.68%)
Unknown	47 (4.94%)	11 (2.74%)
Marital status		
Married	764 (80.25%)	340 (84.79%)
Divorced	74 (7.77%)	29 (7.23%)
Widowed	47 (4.94%)	12 (2.99%)
Never married	36 (3.78%)	9 (2.24%)
Separated	18 (1.89%)	6 (1.5%)
Other & Unknown	13 (1.37%)	5 (1.25%)
Employment status		
Employed full time	322 (33.82%)	152 (37.91%)
Employed part time	77 (8.09%)	37 (9.23%)
Unemployed	32 (3.36%)	13 (3.24%)
Retired	447 (46.95%)	180 (44.89%)
Disabled and not working	62 (6.51%)	15 (3.74%)
Unknown	12 (1.26%)	4 (1%)
Gleason grade 1 score		
< 7	542 (56.93%)	236 (58.85%)
= 7	326 (34.24%)	125 (31.17%)
> 7	84 (8.82%)	40 (9.98%)
Treatments (post baseline)		
Radiation	292 (30.67%)	104 (25.94%)
Hormone with radiation	79 (8.30%)	24 (5.99%)
Prostatectomy	384 (40.34%)	191 (47.63%)
Other treatments	30 (3.15%)	13 (3.24%)
No treatment (e.g., active surveillance, watchful waiting, and/or supplements)	254 (26.68%)	103 (25.69%)

Sample Characteristics at Baseline (unless noted otherwise)	All 24-month participants *N* (%)	Within the phone & PC subset *N* (%)
Unknown	7 (0.74%)	0 (0%)
How often do you use a computer to check email or browse the internet? (24-month PC survey only)		
Never	---	6 (1.50%)
Less than once a week	---	7 (1.75%)
Once a week	---	8 (2%)
Several times a week	---	72 (17.96%)
At least once a day	---	304 (75.81%)
Unknown	---	4 (1%)
Did you prefer answering the health questions on the phone or the computer? (24-month PC survey only)		
Prefer phone	---	43 (10.72%)
Prefer computer	---	233 (58.10%)
No preference	---	119 (29.68%)
Unknown	---	6 (1.50%)
If you were to participate in another study in the future, would you prefer answering questions on your health on the phone or the computer? (24-month PC survey only)		
Prefer phone	---	50 (12.47%)
Prefer computer	---	241 (60.10%)
No preference	---	104 (25.94%)
Unknown	---	6 (1.50%)
Which survey took longer to complete? (24-month PC survey only)		
Phone	---	343 (85.54%)
Computer	---	7 (1.75%)
No difference	---	42 (10.47%)
Unknown	---	9 (2.25%)
For the questions on sexual activities and function, did you prefer answering the questions on phone or computer? (24-month PC survey only)		
Prefer phone	---	29 (7.23%)
Prefer computer	---	181 (45.14%)
No preference	---	186 (46.38%)
Unknown	---	5 (1.25%)

Note: A small proportion of patients received multiple treatments.

Table 2

Confirmatory Factor Analysis and Measurement Invariance Testing Results.

Domain Name / CFA Model	χ^2	*df*	RMSEA	CFI	TLI	WRMR	χ^2 Test of Invariance	ΔCFI	Pass/ Fail
Pain Interference									

Domain Name / CFA Model	χ^2	*df*	RMSEA	CFI	TLI	WRMR	χ^2 Test of Invariance	ΔCFI	Pass/ Fail
Phone Mode	18.927	2	0.145	0.999	0.997	0.521	---	---	---
PC Mode	4.329	2	0.054	1	1	0.210	---	---	---
Configural	69.574	15	0.095	0.999	0.998	0.673	---	---	
Fatigue									
Phone Mode	10.863	2	0.105	0.999	0.998	0.427	---	---	---
PC Mode	1.158	2	0	1	1	0.137	---	---	---
Configural	34.530	15	0.057	0.999	0.999	0.489	---	---	Pass
Strong	58.849	33	0.044	0.999	0.999	0.717	$p = 0.274$	0	Pass
Depression									
Phone Mode	20.001	2	0.150	0.998	0.994	0.626	---	---	---
PC Mode	1.852	2	0.010	1	1	0.180	---	---	---
Configural	33.150	15	0.055	0.999	0.999	0.467	---	---	Pass
Strong	53.474	32	0.041	0.999	0.999	0.670	$p = 0.403$	0	Pass
Anxiety									
Phone Mode	1.168	2	0	1	1.001	0.142	---	---	---
PC Mode	0.545	2	0	1	1	0.101	---	---	---
Configural	22.703	15	0.036	0.999	0.999	0.394	---	---	Pass
Strong	56.253	29	0.048	0.997	0.998	0.812	$p = 0.041$	−0.0018	Pass
Sleep Disturbance[1]									
Phone Mode	0.327	1	0	1	1.001	0.063	---	---	---
PC Mode	1.702	1	0.042	1	0.999	0.091	---	---	---
Configural	26.734	13	0.051	0.999	0.997	0.374	---	---	Pass
Strong	84.412	31	0.066	0.995	0.996	0.876	$p < 0.001$*	−0.004	
Partial Strong	40.321	26	0.037	0.999	0.999	0.568	$p = 0.325$	0	Pass
Sleep Related Impairment[2]									
Phone Mode	47.759	19	0.062	0.996	0.995	0.610	---	---	---
PC Mode	59.371	19	0.073	0.998	0.997	0.621	---	---	---
Configural	271.911	93	0.069	0.993	0.991	0.951	---	---	Pass
Strong	396.807	131	0.071	0.990	0.991	1.376	$p < 0.001$*	−0.003	
Partial Strong	284.918	126	0.056	0.994	0.994	1.092	$p = 0.513$	0.001	Pass
Diarrhea[3]									
Phone Mode	11.610	6	0.048	1	0.999	0.335	---	---	---
PC Mode	9.694	6	0.039	1	1	0.296	---	---	---
Configural	147.330	41	0.080	0.996	0.993	0.903	---	---	Pass
Strong	170.112	68	0.061	0.996	0.996	1	$p = 0.713$	0	Pass
Bowel Incontinence									
Phone Mode	4.154	2	0.052	1	0.999	0.281	---	---	---
PC Mode	6.001	2	0.071	1	0.999	0.341	---	---	---
Configural	25.389	16	0.038	0.999	0.999	0.471	---	---	Pass

Domain Name / CFA Model	χ^2	*df*	RMSEA	CFI	TLI	WRMR	χ^2 Test of Invariance	ΔCFI	Pass/ Fail
Strong	32.141	32	0.003	1	1	0.526	$p = 0.881$	0.001	Pass
Physical Function									
Phone Mode	1.861	2	0	1	1	0.140	---	---	---
PC Mode	0.854	2	0	1	1	0.109	---	---	---
Configural	9.988	15	0	1	1.001	0.195	---	---	Pass
Strong	29.871	33	0	1	1	0.392	$p = 0.773$	0	Pass
Illness Impact – Negative[4]									
Phone Mode	56.466	18	0.073	0.994	0.990	0.596	---	---	---
PC Mode	52.987	18	0.070	0.997	0.996	0.568	---	---	---
Configural	196.618	91	0.054	0.994	0.993	0.737	---	---	Pass
Strong	229.245	129	0.044	0.995	0.995	0.933	$p = 0.593$	0	Pass
Illness Impact – Positive[5]									
Phone Mode	20.893	8	0.063	0.998	0.992	0.304	---	---	---
PC Mode	16.456	8	0.051	0.998	0.995	0.280	---	---	---
Configural	186.692	71	0.064	0.989	0.981	0.729	---	---	Pass
Strong	314.963	109	0.069	0.980	0.978	1.183	$p < 0.001$*	−0.009	
Partial Strong	217.044	104	0.052	0.989	0.987	0.914	$p = 0.275$	0	Pass
Interest in Sexual Activity									
Phone Mode	3.002	2	0.035	1	1	0.201	---	---	---
PC Mode	15.065	2	0.128	0.999	0.997	0.447	---	---	---
Configural	94.260	15	0.115	0.996	0.993	0.761	---	---	
Erectile Function									
Configural	10.580	5	0.068	1	0.999	0.271	---	---	Pass
Strong	84.730	17	0.128	0.995	0.996	0.996	$p = <$ 0.001*	−0.005	
Partial Strong	18.272	12	0.046	1	0.999	0.405	$p = 0.423$	0	Pass
Global Satisfaction with Sex Life[6]									
Phone Mode	1.961	1	0.064	1	0.999	0.137	---	---	---
PC Mode	14.212	1	0.234	0.999	0.993	0.358	---	---	---
Configural	37.278	13	0.093	0.998	0.996	0.436	---	---	

Notes:

- χ^2 = scaled chi-square test of model fit.
- *df* = scaled degrees of freedom for the scaled chi-square statistic.
- RMSEA = scaled root mean square error of approximation.
- CFI = scaled comparative fit index.
- TLI = scaled Tucker-Lewis index.
- WRMR = weighted root mean square residuals.

- χ^2 Test of Invariance = scaled chi-square difference test comparing the strong (or partial strong) invariance model to the configural invariance model. Significant *p* values were marked with an *asterisk* after a Bonferroni correction (α = 0.05/50 = 0.001).
- $\Delta CFI = CFI_{strong} - CFI_{configural}$ (or $CFI_{partial.strong} - CFI_{configural}$)
- Model modifications (with possible explanations) applied toward both survey modes prior to invariance testing:
  1. Sleep Disturbance: freely estimated the residual covariance between *SLEEP20* and *SLEEP44* (both were worded similarly asking about problems with sleep).
  2. Sleep Related Impairment: freely estimated the residual covariance between *SLEEP6* and *SLEEP7* (both assessed sleepiness during the day).
  3. Diarrhea: freely estimated the residual covariances among *GISX38, GISX40*, and *GISX41* (all related to "having loose or watery stools").
  4. Illness Impact – Negative: freely estimated the residual covariance between *II58.a* and *II59.a* (both assessed social disconnectedness), and between *II71.a* and *II80.a* (both assessed uneasiness).
  5. Illness Impact – Positive: freely estimated the residual covariances among *II7.a, II32.a*, and *II35.a* (all assessed optimism), and among the other five items (all related to more profound understandings of life).
  6. Global Satisfaction with Sex Life: freely estimated the residual covariance between *SFSAT101* and *SFSAT102* (both used the phrase "sex life").
- For Bowel Incontinence domain, the autocorrelation of *GISX46* over time was not estimated due to model convergence issues.

Table 3

Estimated Latent Factor Means, Variances, and Autocorrelations for Strongly (or Partially Strongly) Invariant Measures.

Domain Name	Tests of Latent Means (H_0: $\bar{\theta}_{PC} = \bar{\theta}_{Phone} = 0$)			Tests of Latent Variances ($H_0: \sigma^2_{PC} = \sigma^2_{Phone} = 1$)			Tests of Latent Autocorrelations (H_0: $\phi_{PC \cdot Phone} = 0$)		
	$\bar{\theta}_{PC}$	*z*-statistic	*p* value	σ^2_{PC}	$\chi^2_{df=1}$	*p* value	$\phi_{PC \cdot Phone}$	*z*-statistic	*p* value
Fatigue	−0.077	−2.068	0.039	1.062	10.775	0.001	0.844	49.410	< 0.001*
Depression	−0.011	−0.263	0.793	1.050	4.311	0.038	0.888	42.461	< 0.001*
Anxiety	0.074	1.390	0.164	1.085	8	0.005	0.808	26.242	< 0.001*
Sleep Disturbance	0.093	2.276	0.023	0.935	1.665	0.197	0.907	61.277	< 0.001*
Sleep Related Impairment	−0.080	−1.981	0.048	1.130	6.283	0.012	0.869	55.711	< 0.001*
Diarrhea	−0.031	−0.677	0.498	1.013	0.245	0.621	0.894	38.254	< 0.001*
Bowel Incontinence	−0.011	−0.226	0.822	1.026	3.112	0.078	0.954	60.285	< 0.001*
Physical Function	−0.047	−1.534	0.125	1.017	0.588	0.443	0.953	86.934	< 0.001*
Illness Impact – Negative	0.109	2.968	0.003	1.083	3.930	0.047	0.921	60.794	< 0.001*
Illness Impact – Positive	−0.199	−3.935	< 0.001*	0.939	1.016	0.313	0.798	26.536	< 0.001*
Erectile Function	0.086	2.149	0.032	1.064	2.613	0.106	0.935	84.183	< 0.001*

Notes: All latent factors under the referenced phone mode were standardized with a mean ($\bar{\theta}_{Phone}$) of zero and a variance (σ^2_{Phone}) of one. The PC-mode mean and variance estimates were respectively tested against the reference values. Factor autocorrelation estimates ($\phi_{PC \cdot Phone}$) were tested against zero. Significant *p* values were marked with an *asterisk* after a Bonferroni correction (α = 0.05/50 = 0.001).

Appendix C

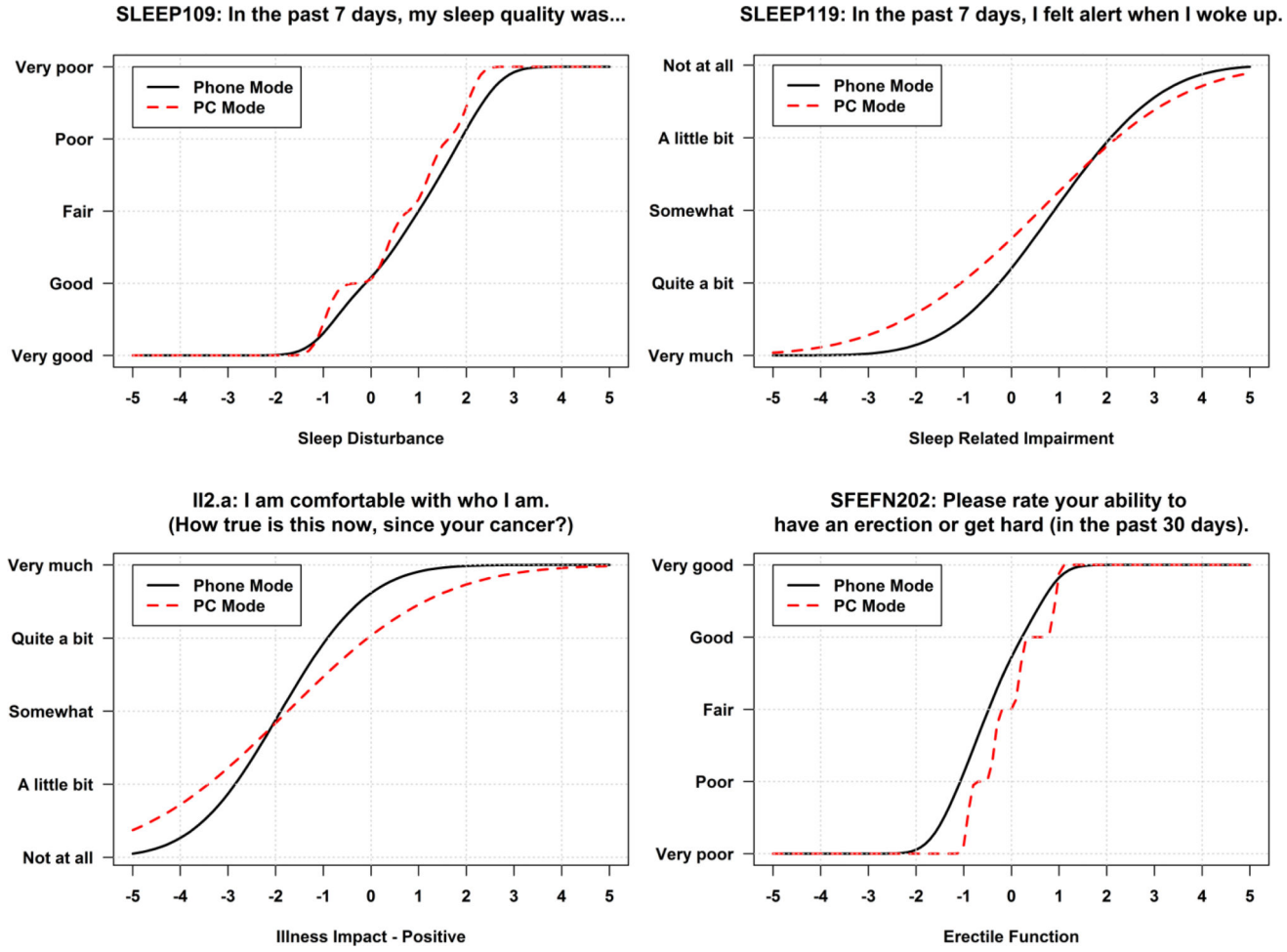

Fig. 1.

Expected Response as a Function of Domain Scores for Items that Behave Differently under Different Survey Modes.