Published in final edited form as:

Qual Life Res. 2016 May ; 25(5): 1123–1130. doi:10.1007/s11136-015-1162-9.

Evaluation of mode equivalence of the MSKCC Bowel Function Instrument, LASA Quality of Life, and Subjective Significance Questionnaire items administered by Web, interactive voice response system (IVRS), and paper

Antonia V. Bennett[1], **Kathleen Keenoy**[2], **Marwan Shouery**[2], **Ethan Basch**[1,2], and **Larissa K. Temple**[2]

Larissa K. Temple: templel@mskcc.org

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

[2]Memorial Sloan Kettering Cancer Center, 1233 York Avenue, New York, NY 10065, USA

Abstract

Purpose—To assess the equivalence of patient-reported outcome (PRO) survey responses across Web, interactive voice response system (IVRS), and paper modes of administration.

Methods—Postoperative colorectal cancer patients with home Web/e-mail and phone were randomly assigned to one of the eight study groups: Groups 1–6 completed the survey via Web, IVRS, and paper, in one of the six possible orders; Groups 7–8 completed the survey twice, either by Web or by IVRS. The 20-item survey, including the MSKCC Bowel Function Instrument (BFI), the LASA Quality of Life (QOL) scale, and the Subjective Significance Questionnaire (SSQ) adapted to bowel function, was completed from home on consecutive days. Mode equivalence was assessed by comparison of mean scores across modes and intraclass correlation coefficients (ICCs) and was compared to the test–retest reliability of Web and IVRS.

Results—Of 170 patients, 157 completed at least one survey and were included in analysis. Patients had mean age 56 (SD = 11), 53 % were male, 81 % white, 53 % colon, and 47 % rectal cancer; 78 % completed all assigned surveys. Mean scores for BFI total score, BFI subscale scores, LASA QOL, and adapted SSQ varied by mode by less than one-third of a score point. ICCs across mode were: BFI total score (Web–paper = 0.96, Web–IVRS = 0.97, paper–IVRS = 0.97); BFI subscales (range = 0.88–0.98); LASA QOL (Web–paper = 0.98, Web–IVRS = 0.78, paper–IVRS = 0.80); and SSQ (Web–paper = 0.92, Web–IVRS = 0.86, paper–IVRS = 0.79).

Conclusions—Mode equivalence was demonstrated for the BFI total score, BFI subscales, LASA QOL, and adapted SSQ, supporting the use of multiple modes of PRO data capture in clinical trials.

Correspondence to: Larissa K. Temple, templel@mskcc.org.

Conflict of interest All authors declared that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Keywords

Mode equivalence; Electronic PRO capture; Rectal cancer; Quality of life

Introduction

Multiple modes of survey administration are available for patient-reported outcome (PRO) data capture, including paper surveys, automated phone surveys using an interactive voice response system (IVRS), and Web-based surveys accessible via PC, tablet, or smartphone. Because of patient variation in preferences and ability to complete surveys via Web or telephone, providing multiple modes of survey administration within a clinical trial may increase eligibility and reduce missing data.

Measurement equivalence between modes is required to use multiple modes in a single study, or to compare data across studies that used different modes for data collection. Differences between modes, if evident in survey data, add measurement error and reduce statistical power. A meta-analytic review of 46 mode-equivalence studies comparing Web (full screen) and handheld (small screen) devices with paper has shown a high level of measurement reliability, with correlations on average of 0.90 [1]. A particular concern in comparison with Web and IVRS is that differences between Web (i.e., visual) and IVRS (i.e., aural) formats may affect survey responses; in this case, mode-equivalence testing is recommended [2].

In planning a national multicenter clinical trial to evaluate a new rectal cancer treatment paradigm, the cancer clinical trial consortium Cancer and Leukemia Group B (CALGB), found that there was a lack of knowledge as to whether PRO data collected via electronic modes of data capture differed from that collected via the traditional method of paper surveys. Therefore, the aim of this study was to assess the equivalence of paper, Web, and IVRS administration of PRO instruments commonly used to measure functional outcomes of patients with colorectal cancer. These instruments include the Memorial Sloan Kettering Cancer Center Bowel Function Instrument (MSKCC BFI) [3], originally validated as a paper survey and adapted to Web and IVRS formats; the MSKCC BFI is widely used and recognized as having strong psychometric properties [4, 5]. In addition, we evaluated the Linear Analog Scale Assessment overall Quality of Life (LASA QOL) item [6] and a single item adapted from the Subjective Significance Questionnaire (adapted SSQ) [7] measuring perceived change in bowel function over the past four weeks. These are frequently used in CALBG and Alliance for Clinical Trials in Oncology cooperative group studies.

Methods

Postoperative colorectal cancer patients at MSKCC were recruited from clinic to participate in this study. Eligible patients were at least 18 years of age, had stage I–III colon or rectal cancer prior to surgery, had undergone surgical resection of the primary tumor at MSKCC, and did not have metastatic disease or stoma. Additionally, eligible patients had access to Web and e-mail from home, had Web avidity (i.e., reported using e-mail at least twice a week), were able to read and write in English, and did not have any auditory, visual, or motor impairment that would preclude the ability to use a telephone or computer. Patients who met clinical eligibility criteria based on review of the medical record were approached by research staff, and interested patients were screened for Web avidity and home Web access. Patients who met all eligibility criteria and completed written informed consent were enrolled in the study. Patients were purposefully recruited such that the study sample would contain 40–60 % rectal cancer patients. This study was approved by the MSKCC Institutional Review Board. ClinicalTrials.gov Identifier: NCT01458509.

This study employed a randomized crossover design. Enrolled participants were randomly assigned to one of the eight study groups. Group assignment was determined when the patient was registered onto the study in an institutional research database by personnel of the protocol office, using the method of random permuted block. Groups 1–6 completed the survey via Web, IVRS, and paper, in one of the six possible orders. Groups 7–8 completed the survey twice, either by Web or by IVRS, to enable comparison of between-mode reliability to within-mode reliability. The order and mode of surveys completed by each group can be seen in Table 1. Surveys were completed from home, on consecutive days. Participants could elect which day to begin the surveys, provided it was within 7 days of the enrollment visit. A patient incentive of $20 was provided at the enrollment visit.

The Web and IVRS surveys were hosted by the MSKCC Webcore survey system [8]. During the enrollment visit, the research staff registered the participant into the survey system and taught the participant how to log in and use the Web and IVRS interface. The participant was given the paper survey and a postage-paid return mail envelope. A one-page paper handout containing brief instructions and the dates of the scheduled assessments was also provided. The survey system automatically sent a reminder e-mail to the participant in the early morning (approximately 12:01 AM) the day before the first survey, and on the days each survey was due. A subsequent reminder e-mail was sent to the participant if the Web survey was not completed by 5:00 PM on the due date, and an automated reminder call was made to the participant if the phone survey had not been completed by 5 PM on that day. The Web and IVRS surveys could only be completed by the participant on the days they were due, and the paper survey was accepted if the postmarked date was within 2 days of the due date.

The 20-item survey completed via each mode contained the MSKCC BFI [3] comprising 18 items, with Likert scale response options ranging from “always” to “never”, and with a recall period of the previous 4 weeks; a single item of the LASA QOL [6] via 11-point numeric rating scale, where 0 indicates “as bad as it can be” and 10 indicates “as good as it can be” (with an implied reference period of “now”); and a single item adapted from the adapted SSQ [7] measuring perceived change in bowel function over the past 4 weeks, with Likert scale response options ranging from “much better than 4 weeks ago” to “much worse than 4 weeks ago.” The MSKCC BFI is summarized as a total score (possible score range 18–90), a four-item Dietary subscale (4–20), a four-item Urgency subscale (4–20), and a 6-item Frequency subscale (6–30), in which higher scores indicate better bowel function [3].

Participant demographic characteristics were self-reported at enrollment; clinical characteristics were abstracted from the medical record and were summarized to describe the sample. The median times to complete a survey by Web or IVRS were calculated from the time of the first and last question completed in the Webcore survey system, excluding several respondents who exited the survey and returned later to finish it. Rates of missing surveys and missing item responses were estimated for each mode. Cronbach's alpha coefficients for each subscale of the BFI were compared across modes. Cronbach's alpha is assessed during instrument validation, and indicates the degree to which items within a subscale are correlated with that subscale; we sought to identify whether this coefficient is consistent across modes [9].

To evaluate mode equivalence, our primary interest was the magnitude of the difference in mean scores by mode and the level of between-mode measurement reliability. To estimate the magnitude, and not simply the statistical significance, of the difference in mean scores by mode, the effect size (Cohen's *d*) was estimated for the mean pairwise difference in scores for each pair of modes. An effect size of less than 0.20 indicated equivalence of scores [10]. Differences in mean scores by mode were also assessed using mixed-effects model for the 3-mode crossover design to identify by mode and order effects [11]. The model included terms for mode, order (i.e., first, second, or third), mode by order interaction, and sequence (i.e., Groups 1–6). Variation in mode effects by patient characteristics was examined by including one of the following patient characteristics to the above model: gender, education (at least a college degree vs. no college degree), age in tertiles (low-54, 55–59, 70+), cancer type (colon vs. rectal), and BFI total score in quartiles.

Measurement reliability, specifically the between-mode reliability and test–retest reliability, was assessed via intraclass correlation coefficient (ICC). Mode equivalence was defined by standard criteria used to assess measurement reliability (i.e., 0.70 or greater is considered sufficient), [12] and by comparing the between-mode reliability to the test–retest reliability to determine whether the reliability between modes was any less than the reliability of multiple assessments in a single mode.

Sample size was determined via methods described by Walter [13] and Stratford [14], such that the equivalence of score distributions could be assessed by testing whether the ICCs of scores per each pair of modes were ≥0.70. For Groups 1–6, the target enrollment was 60 patients (10 per group), anticipating a 70 % (43/60) survey completion rate. This provided 80 % power and a 5 % type 1 error rate in the crossover design to determine that an observed ICC of 0.85 is larger than 0.70. For Groups 7–8, the target enrollment was 116 (58 per group), also anticipating a 70 % (80/116) survey completion rate. Assuming the z-transformed ICC statistics would have correlation ≥0.60 provided 80 % power with 5 % type 1 error rate for most 5- and 10-point differences in ICC when ICCs are ≥0.80.

Results

Four hundred thirty-five patients met clinical eligibility criteria, and of these 265 (61 %) reported having home Web access and Web avidity sufficient for eligibility. One hundred seventy patients enrolled in the study. One hundred fifty-seven of 170 (92 %) completed at least one survey and were included in the analysis. One hundred twenty-two of 157 (78 %) completed all assigned surveys. In Groups 1–6, 61/81 (75 %) completed all three surveys, and 74/81 (91 %) completed two of three surveys. In Groups 7–8, 62/76 (81 %) completed both surveys.

Demographic and clinical characteristics of the cohort are shown in Table 2. Patients had a mean age of 56 (SD = 11), 53 % were male, 82 % were Caucasian, and 78 % had at least a college degree. Cancer diagnoses were colon (54 %) and rectal (46 %). Surgery included left or right colectomy (47 %) or low anterior resection (43 %), with or without coloanal anastomosis.

The median time to complete the survey was 1 min 44 s by Web (inter-quartile range 1:18–2:16), and 8 min 29 s by IVRS (IQR 8:05–9:23). On average, 99 % of questionnaire items were answered on completed surveys. Across all modes, 94 % of completed surveys had no missing data (94 % Web, 91 % paper, and 98 % IVRS).

Cronbach's alpha was calculated for each subscale of the BFI and compared across modes (Table 3). For each subscale, the resulting indices from each mode were similar in value and would not be interpreted differently: Dietary (range 0.78–0.84), Urgency (range 0.85–0.87), and Frequency (range 0.63–0.69). This measurement property of the BFI did not vary meaningfully by mode.

Mean scores for BFI total score, BFI subscale scores, LASA QOL, and adapted SSQ score varied by mode by less than one-third of a score point (Table 4). The effect size (Cohen's *d*) of the score differences was nearly zero for the BFI total score and subscale scores (range 0.00–0.05) and ranged from zero to less than "small" for the LASA QOL and adapted SSQ (range 0.00–0.17) (Table 5). The mixed-effects model controlling for mode, order, mode by order interaction, and sequence found no statistically significant differences in response by mode, except for the BFI Urgency subscale in which differences of 0.30 points were statistically significant; however, this difference is equivalent to an effect size of only 0.09. For the BFI total score, Dietary subscale, and Urgency subscale, very small (negligible in size) order effects were statistically significant. No variations in mode effects were evident by participant demographic and clinical characteristics.

For the BFI total score and subscale scores, the ICCs for between-mode measurement reliability (i.e., for Web vs. paper, Web vs. IVRS, and IVRS vs. paper) were high, and ranged from 0.88 to 0.98 (Table 6). The test–retest reliability of Web for the BFI total score and subscale scores ranged from 0.88 to 0.97. The between-mode measurement reliability of Web versus IVRS and Web versus paper was highly similar to the test–retest reliability of Web; the differences in ICCs ranged from 0.00 to 0.02. The between-mode measurement reliability comparisons including IVRS were also similar to the test–retest reliability of IVRS, except for the BFI Dietary scale, for which the test–retest reliability of IVRS was lower (0.73). The lower bounds of the 95 % confidence intervals (CIs) for estimates of the ICC were all greater than 0.70, except for the test–retest reliability of IVRS for the BFI Dietary scale.

For the LASA QOL item, the mode equivalence (and 95 % CI) of Web versus IVRS and paper versus IVRS was 0.79 (0.69–0.87) and 0.80 (0.71–0.99), respectively. The mode equivalence of Web versus paper and the test–retest reliability of Web was very similar, and quite high (0.98 and 0.99, respectively). The test–retest reliability of IVRS was low (ICC 0.49; 95 % CI 0.21–0.77), due to two patients with highly discordant responses (2 vs. 10, and 10 vs. 3); omitting these observations changes the ICC to 0.89. For the adapted SSQ item, the ICCs across mode were moderate to high, ranging from 0.79 to 0.86, with the lower bound of the 95 % CI all greater than or equal to 0.70. The test–retest reliability of Web was high (0.97); for IVRS it was moderate (0.74).

Discussion

This study demonstrated that the mode of survey administration did not affect the results in a meaningful manner, indicating that clinical trials can use any or all modes to administer the MSKCC BFI, LASA QOL, and the adapted SSQ. The measurement reliability between modes for multiple-item scales was high (≥0.87), and for single-item scales moderate or high (≥0.79). The observed between-mode reliability is similar to the test–retest reliability of each mode. Thus, these findings suggest there would be no or little loss of statistical power from using multiple modes instead of a single mode in measuring PRO. The effect size of the difference in scores between modes ranges from 0.00 to 0.05 for multiple-item scales, and 0.00 to 0.17 for single-item scales. In models controlling for order and sequence, the difference between modes on the Urgency subscale of 0.30 was statistically significant, but the effect size of the difference was 0.09 and can be considered negligible. The Cronbach's alpha coefficients did not vary by mode.

The BFI total score and subscale scores demonstrated a very high level of mode equivalence, encompassing measurement reliability, mean differences, and internal reliability. The single-item LASA QOL and adapted SSQ were not as robust as the multiple-item scales, but the measurement reliability and mean differences observed between modes were similar to those observed in repeated assessments by the same mode. For all scales, the equivalence of scores by mode was acceptable. These findings are consistent with the small but growing literature on mode equivalence.

Five prior studies have evaluated mode equivalence of paper and IVRS of various measures, using randomized crossover designs. Lundy et al. [15] found that paper and IVRS assessments of the EQ-5D index score had an ICC of 0.89 and mean difference of −0.009 (scale range 0.0–1.0), and the EQ visual analog scale (EQ VAS) had an ICC of 0.89 and mean difference of −2.96 (scale range 0–100). In a more recent study, Lundy et al. [16] found the ICCs for each subscale of the European Organization for Research and Treatment of Cancer (EORTC) quality-of-life questionnaire (QLQ-C30) ranged from 0.79 to 0.90 and had 95 % lower confidence interval greater than 0.70. Rush et al. [17] found a high level of correlation ($r = 0.89$), small differences in mean scores (Cohen's $d = 0.18$), and similarly high Cronbach's alpha coefficients (0.86 vs. 0.87) between paper and IVRS versions of the 16-item Quick Inventory of Depressive Symptomatology (QIDS). Dunn et al. [18] found very small mean differences and high correlations between paper and IVRS administration for the total score ($r = 0$ .96) and subscales (range 0.88–0.93) of the Changes in Sexual Functioning Questionnaire (CSFQ). Agel et al. [19] found only small mean differences between paper and IVRS administration for total scores and subscales of the Short Musculoskeletal Function Assessment (SMFA).

Two recent studies have compared paper, screen-based (e.g., Web), and IVRS using randomized crossover designs. Bjorner et al. [20] evaluated the mode equivalence of three PROMIS® scales measuring physical function, fatigue, and depression, comparing personal computer administration with IVRS, paper, and personal digital assistant in a randomized crossover design, and found ICCs ranging from 0.85 to 0.94, Cronbach's alpha ranging from 0.92 to 0.97, and no differences in score level. An evaluation of the NCI's patient-reported outcomes version of the Common Terminology Criteria for Adverse Events (NCI PRO-CTCAE) found that across an evaluation of 27 items, the median ICC comparing tablet versus IVRS was 0.78 (range 0.55–0.90); for tablet versus paper was 0.81 (0.62–0.96); for IVRS versus paper was 0.78 (0.60–0.91). Eighty-nine percent of ICCs were ≥0.70. The item-level mean differences by mode were small (less than 0.20 standard deviations) and were statistically nonsignificant (*manuscript under review*). Additionally, our findings support the equivalence of paper, Web, and IVRS.

This study of mode equivalence is the first study to compare paper, Web, and IVRS administration of the widely used LASA QOL and the SSQ. Furthermore, information on mode equivalence of the MSKCC BFI will provide insight into the potential mode equivalence of other disease-specific measures of symptom and function, many of which comprise multi-item subscales with Likert scale response options.

The rate of missing data for this 20-item survey was low and did not appear to vary by mode. For both Web and IVRS, in the repeated assessment (Groups 7 and 8), some patients did not complete the survey again on the second day. While the survey response rates varied between IVRS and Web, it is interesting that the number of items missing in a survey did not vary by mode, especially because the IVRS survey took longer to complete than the Web survey (median time of 8 min 29 s vs. 1 min 44 s, respectively).

The test–retest reliability of the BFI dietary subscale, the LASA QOL, and the adapted SSQ measured via IVRS was not statistically significantly greater than 0.70; i.e., the entire 95 % confidence interval was not above 0.70. Although evaluating the test–retest reliability of PRO instruments in each mode of administration was not of primary interest in this study, the lower rates of test–retest reliability via IVRS are a concern. We identified two pairs of highly discordant pairs of responses in the LASA QOL item which greatly reduced the ICC statistic. The type of response options and the directionality, i.e., whether a lower number indicates better or worse function, switched several times in the survey, most so between the MSKCC BFI, LASA QOL, and adapted SSQ, and this may have contributed to confusion in the aural (IVRS) administration of the survey. The small sample size in Groups 7 and 8 may have made the reliability statistic more sensitive to discordant responses.

This study employed a randomized crossover design with assessments completed on subsequent days. This enabled comparison of modes, while controlling for order effects and differences between individuals; the approximately 24-h period between each assessment would contribute to patients answering subsequent surveys based on their experience, rather than remembering the response provided on the previous day's survey. In designing this study, the time between assessments was carefully considered. To our knowledge, there are no empirical studies substantiating the time between assessments in evaluations of mode equivalence or test–retest reliability. However, the guiding principle has been to choose a time that is not so long that symptoms have changed, and not so short that respondents remember their previous answers. Clinically, bowel function can vary significantly over a short period of time. For this reason, it was important to use a relatively short time period between assessments. One limitation of this design was potential variability with respect to when the paper survey was completed—paper surveys included in the analysis were those postmarked no later than two days after the survey due date. However, this was intended to balance feasibility of the data collection with the desire to have surveys completed from home and on subsequent days. Eligibility for this study required participants to have home Web access, because the study was designed to evaluate within-patient differences in scores between all three modes, and because assessments were completed 1 day apart instead of during a single clinic visit. The demographics of the study sample may differ from those involving patients who do not have home Web access. While possible, it is unlikely that differences in demographics would significantly alter the measurement equivalence. However, the issue of differences in feasibility of assessment by each mode remains. Investigators should consider the feasibility of each mode of assessment in all clinical trial participants. The Pew Research Center on Internet, Science and Technology (pewinternet.org) publishes reliable information about computer use and internet access in the USA by demographic group.

This research contributes to a growing body of evidence supporting the equivalence of PRO scale scores across paper, Web, and IVRS modes of administration, particularly for multiple-item scales. The single-item scales included in this survey, the LASA QOL and adapted SSQ, are subject to the well-known phenomenon of lower measurement reliability when compared to multi-item scales. In this survey, the response scales for the LASA QOL and adapted SSQ were very different from the items that preceded them, both in terms of the range of response options and the directionality of response options. Investigators are cautioned to implement surveys keeping in mind the potential for response error resulting from confusion on the part of the respondent. In addition, investigators should consider that the length of time necessary to complete a survey by IVRS is longer than for Web or paper. This study found a high level of mode equivalence for the BFI total score and subscales and acceptable results for the single-item LASA QOL and adapted SSQ. These findings support the use of multiple modes of data capture for these instruments in a clinical trial. The findings are consistent with other studies, and it is likely that this is also the case for similar PRO instruments. In the clinical care context, making all modes available to patients is encouraged, to facilitate the monitoring of their symptoms and function.

Acknowledgments

Funding This study was funded by Cancer and Leukemia Group B Foundation (US) (Grant Number: 2U10CA037447-26). This study was also funded in part by the Memorial Sloan Kettering Cancer Center Core Grant P30 CA008748. The core grant provides support to institutional cores, such as Biostatistics and Pathology, which were used in this study.

References

1. Gwaltney CJ, Shields AL, Shiffman S. Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. Value in Health. 2008; 11(2):322–333. [PubMed: 18380645]
2. Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value in Health. 2009; 12(4): 419–429. [PubMed: 19900250]
3. Temple LK, Bacik J, Savatta SG, et al. The development of a validated instrument to evaluate bowel function after sphincter-preserving surgery for rectal cancer. Diseases of the Colon and Rectum. 2005; 48(7):1353–1365. [PubMed: 15868235]
4. Chen TY, Emertsen KJ, Laurberg S. What are the best questionnaires to capture anorectal function after surgery in rectal cancer? Current Colorectal Cancer Reports. 2015; 11:37–43. [PubMed: 25663833]
5. Wong C, Chen J, Yu CL, Sham M, Lam CL. Systemic review recommends the European Organization for Research and Treatment of Cancer colorectal cancer-specific module for measuring quality of life in colorectal cancer patients. Journal of Clinical Epidemiology. 2015; 68(3):266–278. [PubMed: 25455838]
6. Locke DE, Decker PA, Sloan JA, et al. Validation of single-item linear analog scale assessment of quality of life in neuro-oncology patients. Journal of Pain and Symptom Management. 2007; 34(6): 628–638. [PubMed: 17703910]
7. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. Journal of Clinical Oncology. 1998; 16(1):139–144. [PubMed: 9440735]
8. Bennett AV, Jensen RE, Basch E. Electronic patient-reported outcome systems in oncology clinical practice. CA: Cancer Journal for Clinicians. 2012; 62(5):337–347.
9. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16:297–334.
10. Cohen, J. Statistical power analysis for the behavioral sciences. 2. London: Routledge; 1988.
11. Yarandi, H. Crossover designs and proc mixed in SAS. Paper SD04; The proceedings of the SouthEast SAS Users Group; Nashville, TN. 2004. http://analytics.ncsu.edu/sesug/2004/SD04-Yarandi.pdf
12. Nunnally, JC.; Bernstein, IH. Psychometric theory. 3. New York: McGraw-Hill; 1994.
13. Walters SJ, Campbell MJ, Paisley S. Methods for determining sample sizes for studies involving health-related quality of life measures: A tutorial. Health Services and Outcomes Research Methodology. 2001; 2:83–99.
14. Stratford PW, Binkley JM. Applying the results of self-report measures to individual patients: An example using the Roland–Morris Questionnaire. Journal of Orthopaedic and Sports Physical Therapy. 1999; 29(4):232–239. [PubMed: 10322596]
15. Lundy JJ, Coons SJ. Measurement equivalence of interactive voice response and paper versions of the EQ-5D in a cancer patient sample. Value in Health. 2011; 14(6):867–871. [PubMed: 21914508]
16. Lundy JJ, Coons SJ, Aaronson NK. Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. Quality of Life Research. 2014; 23(1): 229–237. [PubMed: 23765449]
17. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the quick inventory of depressive symptomatology and the Hamilton rating scale for depression: A sequenced treatment alternatives to relieve depression trial report. Biological Psychiatry. 2006; 59(6):493–501. [PubMed: 16199008]
18. Dunn JA, Arakawa R, Greist JH, Clayton AH. Assessing the onset of antidepressant-induced sexual dysfunction using interactive voice response technology. Journal of Clinical Psychiatry. 2007; 68(4):525–532. [PubMed: 17474807]

19. Agel J, Rockwood T, Mundt JC, Greist JH, Swiontkowski M. Comparison of interactive voice response and written self-administered patient surveys for clinical research. Orthopedics. 2001; 24(12):1155–1157. [PubMed: 11770093]

20. 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. 2014; 67(1):108–113. [PubMed: 24262772]

Table 1

Numbers of completed surveys by group and mode

Groups 1–6	*N*	Web	Paper	IVRS	Total surveys
1. (P, I, W)	14	11	13	11	35
2. (P, W, I)	13	11	13	12	36
3. (I, P, W)	15	14	14	14	42
4. (I, W, P)	11	11	11	9	31
5. (W, P, I)	15	12	14	13	39
6. (W, I, P)	13	11	11	11	33
Total	81	70	76	70	216

Groups 7–8	*N*	Survey 1	Survey 2	Total surveys
7. (I, I)	37	37	29	66
8. (W, W)	39	39	33	72
Total	76	76	62	138

N number of participants who completed at least one survey, *P* paper, *I* interactive voice response system (IVRS), *W* Web

Table 2

Participant demographic and clinical characteristics

Demographic and clinical characteristics	***N* = 157**
Age, mean (SD)	56 (11)
Male	53 %
Married	77 %
Race	
White	82 %
Black or African-American	5 %
Asian	5 %
Hispanic or Latino/Latina	6 %
Education, highest level completed	
High school degree	8 %
Associate degree or some college	14 %
College degree or more	78 %
Diagnosis	
Colon	54 %
Rectal	46 %
Surgery	
Left or right colectomy	47 %
Total colectomy ± proctocolectomy	7 %
Low anterior resection ± coloanal anastomosis	43 %
Transanal excision	3 %

SD standard deviation

Table 3

Cronbach's alpha coefficient by mode for MSKCC BFI subscales

Cronbach's alpha	*N*	BFI dietary	BFI urgency	BFI frequency
Web	70	0.84	0.87	0.69
Paper	76	0.78	0.85	0.63
IVRS	70	0.80	0.86	0.68

Limited to Groups 1–6

MSKCC BFI Memorial Sloan Kettering Cancer Center Bowel Function Instrument, *IVRS* interactive voice response system, *N* number of participants who completed the survey in that mode

Table 4

Mean scores by mode for each scale

Mean (SD)	N^a	BFI total score	BFI dietary	BFI urgency	BFI frequency	LASA QOL	Adapted SSQ
Groups 1–6							
Web	68	69.5 (11.4)	14.4 (3.6)	17.2 (3.5)	23.4 (3.8)	8.5 (1.3)	2.6 (0.7)
Paper	76	69.5 (10.9)	14.4 (3.3)	17.2 (3.5)	23.1 (3.6)	8.4 (1.3)	2.6 (0.8)
IVRS	70	69.5 (11.7)	14.6 (3.7)	17.0 (3.6)	23.2 (3.9)	8.2 (1.7)	2.6 (0.7)
Groups 7–8							
Web	39	69.4 (12.3)	14.6 (2.9)	16.8 (3.9)	23.1 (4.2)	8.3 (1.6)	2.7 (0.7)
IVRS	37	66.4 (11.5)	14.0 (3.2)	16.4 (3.6)	22.4 (4.0)	7.1 (2.1)	2.3 (0.8)

Mean scores for Groups 7–8 are calculated from the first (day 1) assessment

BFI Memorial Sloan Kettering Cancer Center Bowel Function Instrument, *LASA QOL* Linear Analog Scale Assessment Quality of Life, *SSQ* Subjective Significance Questionnaire, *IVRS* interactive voice response system

[a] Sample size (*N*) is indicated for the BFI total score

Table 5

Between-mode and within-mode mean differences for each scale

Mean difference (effect size)	N[a]	BFI total score	BFI dietary	BFI urgency	BFI frequency	LASA QOL	Adapted SSQ
Between-mode difference							
Web–IVRS	62	0.00 (0.00)	−0.19 (0.05)	0.10 (0.02)	0.08 (0.02)	0.25 (0.17)	0.01 (0.02)
Web–paper	64	−0.12 (0.02)	−0.17 (0.05)	−0.03 (0.01)	0.10 (0.03)	0.04 (0.05)	0.00 (0.00)
IVRS–paper	66	0.32 (0.03)	0.00 (0.00)	0.14 (0.04)	0.02 (0.01)	0.19 (0.12)	0.00 (0.00)
Within-mode difference							
Web–Web	33	−1.06 (0.09)	−0.51 (0.18)	−0.24 (0.06)	−0.03 (0.01)	−0.03 (0.02)	0.03 (0.05)
IVRS–IVRS	28	0.57 (0.05)	0.46 (0.14)	0.18 (0.05)	−0.29 (0.07)	−0.18 (0.08)	−0.12 (0.13)

Effect size is Cohen’s d, reported as absolute values. Sample is limited to available pair-wise comparisons

BFI Memorial Sloan Kettering Cancer Center Bowel Function Instrument, *LASA QOL* Linear Analog Scale Assessment Quality of Life, *SSQ* Subjective Significance Questionnaire, *IVRS* interactive voice response system

[a] Sample size (*N*) is indicated for the BFI total score

Table 6

Between-mode and within-mode measurement reliability for each scale

ICC	N^a	BFI total score	BFI dietary	BFI urgency	BFI frequency	LASA QOL	Adapted SSQ
Between-mode reliability							
Web–IVRS	62	0.97 (0.95–0.98)	0.88 (0.83–0.94)	0.98 (0.97–0.99)	0.93 (0.90–0.96)	0.79 (0.69–0.87)	0.86 (0.79–0.92)
Web–paper	64	0.97 (0.95–0.98)	0.88 (0.83–0.93)	0.98 (0.97–0.99)	0.94 (0.91–0.97)	0.98 (0.97–0.99)	0.92 (0.88–0.95)
IVRS–paper	66	0.97 (0.95–0.98)	0.87 (0.83–0.92)	0.98 (0.96–0.99)	0.91 (0.86–0.95)	0.80 (0.71–0.89)	0.79 (0.70–0.88)
Within-mode reliability							
Web–Web	33	0.97 (0.95–0.99)	0.88 (0.80–0.96)	0.97 (0.95–0.99)	0.96 (0.93–0.99)	0.99 (0.99–1.00)	0.97 (0.94–0.99)
IVRS–IVRS	28	0.95 (0.92–0.98)	0.73 (0.55–0.90)	0.93 (0.88–0.98)	0.95 (0.91–0.98)	0.49 (0.21–0.77)[b]	0.74 (0.57–0.90)

BFI Memorial Sloan Kettering Cancer Center Bowel Function Instrument, *LASA QOL* Linear Analog Scale Assessment Quality of Life, *SSQ* Subjective Significance Questionnaire, *IVRS* interactive voice response system

[a] Sample size (*N*) is indicated for the BFI total score

[b] The low ICC for IVRS–IVRS is driven by two highly discordant pairs; without these two pairs the ICC is 0.89