4-1-2016

Comparison of interactive voice response (IVR) with paper administration of instruments to assess functional status, sexual function, and quality of life in elderly men.

RC Rosen
New England Research Institutes, Inc., Watertown, MA

AJ Stephens-Shields
Department of Biostatistics and Epidemiology, Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA

GR Cunningham
Division of Diabetes, Endocrinology and Metabolism, Baylor College of Medicine and St. Luke's Episcopal Hospital, Houston, TX

D Cifelli
Department of Biostatistics and Epidemiology, Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA

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

See next page for additional authors

Published in final edited form as:

Qual Life Res. 2016 April ; 25(4): 811–821. doi:10.1007/s11136-015-1133-1.

Comparison of Interactive Voice Response (IVR) with Paper Administration of Instruments to Assess Functional Status, Sexual Function and Quality of Life in Elderly Men

Raymond C. Rosen, Ph.D.[1], **Alisa J. Stephens-Shields, Ph.D.**[2], **Glenn R. Cunningham, M.D.**[3], **Denise Cifelli, M.S.**[2], **David Cella, Ph.D.**[4], **John T. Farrar, M.D., Ph.D.**[2], **Elizabeth Barrett-Connor, M.D.**[5], **Cora E. Lewis, M.D., M.P.H.**[6], **Marco Pahor, M.D.**[7], **Xiaoling Hou, M.S.**[2], and **Peter J. Snyder, M.D.**[8]

[1]New England Research Institutes, Inc., Watertown, MA, USA

[2]Department of Biostatistics and Epidemiology, Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA, USA

[3]Division of Diabetes, Endocrinology and Metabolism, Baylor College of Medicine and St. Luke's Episcopal Hospital, Houston, TX, USA

[4]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[5]Department of Family and Preventative Medicine, University of California San Diego School of Medicine, La Jolla, CA, USA

[6]Division of Preventive Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

[7]Department of Aging and Geriatric Research, University of Florida, Gainesville, FL, USA

[8]Division of Endocrinology, Diabetes, and Metabolism, Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA, USA

Abstract

Purpose—Patient reported outcome (PRO) measures are essential for assessing subjective patient experiences. Interactive voice response (IVR) data collection provides advantages for clinical trial design by standardizing and centralizing the assessment. Prior to adoption of IVR as a

CORRESPONDING AUTHOR: Peter J. Snyder, MD, Perelman School of Medicine at the University of Pennsylvania, Room 12-135, Translational Research Building (Building 421), 3400 Civic Center Boulevard, Philadelphia, PA 19104-5160, P: (215) 898-0208, F: (215) 898-5408, pjs@mail.med.upenn.edu.

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: GRC has served as a consultant to AbbVie, Clarus Therapeutics, Endo Pharma, Ferring, Lilly, Repros Therapeutics, and he has served as an expert witness for Repros Therapeutics and Solvay. He has received research support from Ardana, Unimed and Abbvie; CEL was supported by the National Institute for Diabetes, Digestive and Kidney Diseases, National Institutes of Health (DK079626) to the UAB Diabetes Research and Training Center; PJS reports grants from NIH and AbbVie for the conduct of this study; Remaining authors report no conflict of interest.

mode of PRO administration in The Testosterone Trials (TTrials), we compared IVR to paper versions of the instruments to be used.

Methods—IVR versions of the FACIT-Fatigue scale and Psychosexual Daily Questionnaire, Question 4 (PDQ-Q4) were developed. In one pilot study, IVR versions of these scales were compared to paper versions in 25 men ≥65 years at each of two clinical sites. In another study, IVR versions of the SF-36 Vitality Scale (SF-36), Positive and Negative Affect Scale (PANAS), and Patient Health Questionnaire (PHQ-9) were evaluated in comparison to previously validated paper versions in 25 men at two clinical sites. Both paper and IVR versions of each instrument were administered in counterbalanced order, and test-retest reliability was evaluated by repeated administration of the test. Bland Altman plots were used to assess the degree of agreement. Test-retest correlations for each measure were also determined.

Results—Satisfactory agreement was observed between IVR and paper versions of each study measure. Specifically, linear and highly positive associations were observed consistently across the study for IVR and paper versions of all study measures. These ranged from r=0.91–0.99. Test-retest reliability for all measures was acceptable or better (r=0.70–0.90).

Conclusions—The IVR versions of TTrials endpoints in these two studies performed consistently well in comparison to paper versions.

Keywords

Interactive voice response; sexual function; vitality; fatigue; quality of life

INTRODUCTION

Interactive Voice Response (IVR) methodology has been in widespread use for two decades for assessing patient reported outcomes across a variety of disease states, interventions, and clinical trial designs [1–4]. This methodology involves computer-assisted, telephone interviewing of research participants about their health status, quality of life or health-related symptoms. It has major advantages in the automation and standardization of data collection in clinical trials. IVR versions of widely used outcome measures in psychiatry include the Hamilton Depression Scale [5,1] and Pittsburgh Sleep Quality Index [6,7]. The IVR versions of these measures have been favored by the FDA for use in registration trials in psychiatry and sleep medicine. More recently, IVR versions of other well-known measures have been developed and validated for use in female sexual dysfunction [8], type 2 diabetes [9], HIV-AIDS [10–12], chronic lung disease [13], prostate cancer [14], and various other diseases and indications. Among the advantages by study sponsors and regulatory agencies, use of IVR for data collection has been shown to provide accuracy and reliability of recording, reduce recall errors and improve reliability versus paper measures, and provide firm documentation of time and date of data entry [1–6].

The Testosterone Trials (TTrials) are seven coordinated trials to determine the benefit of testosterone treatment of hypogonadal men ≥65 years [15]. Based on the considerations above, we decided to use IVR to collect patient reported outcome (PRO) data in pre-selected domains of outcome, but needed to show reproducibility, feasibility and validity of the IVR measures prior to initiation of the main trials. This paper presents results of two mode of

administration studies of IVR compared to standard paper testing prior to its use in the TTrials. We compared IVR and paper versions of five measures: (a) The Psychosexual Daily Questionnaire, Question 4 (PDQ-Q4) [16], which assesses sexual activity; (b) Functional Assessment of Chronic Illness Therapy - Fatigue (FACIT-F*),* a validated, standardized self-report measure of strength and fatigue. [17]; (c) Short Form-36 (SF-36), Vitality subscale, a widely used measure of vitality and vigor [18,19]; (d) Positive and Negative Affect Scale (PANAS), a bi-directional validated measure of mood and affect [20]; and (e) Patient Health Questionnaire (PHQ-9), a screening measure for depression. [21] We assessed inter-modality agreement, test-retest reliability and internal consistency of IVR and paper versions of each measure.

STUDY DESIGN AND METHODS

Two separate pilot studies were conducted to minimize participant burden and to ensure completion of the study instruments. In the first study we evaluated the PDQ-Q4 and FACIT-F. The PDQ-Q4 was tested a single time over 7 days; the FACIT-F scale [17] was assessed first on day 1, and a retest was administered on day 7. In the second study we evaluated the SF-36 [18,19], PHQ-9 [21], and PANAS [20] measures. The PHQ-9 was retested at one week; retesting of the other measures was performed one day after the initial administration. All questionnaires were administered to all participants both in written format and by telephone using IVR technology. To control for order effects, the order of testing, paper or IVR first, was randomized at the initial visit for all subsequent testing.

Sites and Participants

Both studies were performed at two clinical sites, the University of Florida and University of Alabama at Birmingham, 50 men in the first study and 51 in the second. The participants were chosen to be similar to those expected in the TTrials, men ≥65 years, who understood English well enough to complete all the forms, and who successfully completed a Mini Mental Status Examination (MMSE), demonstrating the absence of significant cognitive impairment. The demographics of both groups are shown in Table 1. All men gave written, informed consent as approved by their respective institutional review boards.

Data Collection Procedures

First Clinic Visit and At Home Instructions—Each participant completed the first day of assessment at the clinic. Participants were instructed in use of the IVR system and also completed the MMSE as required for eligibility. If participants qualified, they completed the FACIT-F and first day of the 7-day PDQ-Q4 scales in both paper and IVR format according to their randomized order. After they completed the assessments in the clinic, they were given a set of six envelopes and instructions to complete the assessments at home in both IVR and paper formats for the subsequent six days. In both formats, participants had available blank response booklets so that the IVR test context was as similar to the paper test context as possible. Participants were instructed to mail the completed paper assessment forms to the coordinating center each day after completion.

IVR Methodology

The IVR system was operated by an independent research organization, Criterium (Saratoga Springs, NY), which programmed the questionnaires for this purpose. Participants were given written instructions on how and when to call the IVR phone number. When they called, they were asked to provide their study identification number by their telephone keypad. They were then asked the prerecorded screening questions and asked to answer by telephone keypad. Only subjects who responded correctly to screening questions were entered into the study. More formal measures of hearing ability or comprehension were not included.

Statistical Analyses

Sample size analysis for Bland-Altman Agreement Analyses—Sample size determination was based on the primary analyses of agreement between IVR and paper versions of each measure based on 95% Bland-Altman limits of agreement. [22] For this study, the primary parameters of interest are the lower and upper bounds of the 95% limits of agreement. We required that the size of the interval around the agreement bounds be no larger than one-tenth of the difference in standard deviation. Under these assumptions, 34 participants were required. Fifty participants were enrolled in the first pilot study and 51 in the second study to account for possible missing data.

The first objective of our analyses was to assess the level of agreement between the paper and IVR versions of the study measures. Our second and third objectives were to evaluate the test-retest reliability and internal consistency of each administration mode. The following analyses were completed on each measure following each objective:

IVR vs Paper Agreement—Means and standard deviations of each measure by mode and day of administration were calculated. Systematic differences between paper and IVR versions administered on the same day were tested by the Wilcoxon Signed Rank Test and further examined by Spearman correlation coefficients between IVR and paper responses and Bland and Altman plots and 95% limits of agreement. [22] Limits of agreement were calculated as the mean difference of IVR compared to paper $\pm t_{(.025)}$ times the standard deviation of the differences where $t_{(.025)}$ is the Student t-statistic value with upper tail probability equal to 0.025 and degrees of freedom equal to one less than the sample size. These limits and the size and nature of systemic bias were represented in Bland-Altman plots to assess acceptability of the IVR approach. Generally, a conclusion of good agreement is reached when the Bland-Altman limits of agreement exclude clinically meaningful values. Non-constant bias of IVR compared to paper versions was evaluated by the Spearman correlation of the difference and mean response by IVR and paper methods.

Test-retest reliability—Test-retest reliability was evaluated by Spearman correlation between day 1 and day 2 (or day 7 for the FACIT-F and PHQ-9) assessments for IVR and paper administrations of the FACIT-F, SF-36, PANAS, and PHQ-9, since there were two assessments available for each subject. Test-retest reliability was not evaluated for the PDQ-Q4 since for each participant a single PDQ-Q4 assessment was completed over the course of 7 days. A positive correlation indicates that values trend well between the test and retest

such that participants with high test responses also have high retest responses relative to other participants.

Internal consistency—Cronbach's alpha was computed on each day to assess internal consistency of both IVR and paper versions of all assessments measured. Cronbach's alpha ≥ 0.9 indicates excellent reliability such that items across a measure are well correlated and reflect a single construct such as sexual activity or vitality. Cronbach's alpha of 0.7 is generally thought to be acceptable in psychometric research [23].

Analyses were based on all available data and thus valid under the assumption of missing completely at random. Adjustments for missing data such as imputation or inverse weighting were not conducted. For most measures, ≥90% of participants provided complete data such that the effect of adjustments for missingness would likely have little impact on conclusions.

RESULTS

Sample Characteristics

Participant characteristics for Study 1 and Study 2 are shown in Table 1. The mean ± SD age was 75.9 ± 5.4 in Study 1 and 75.5 ± 4.9 in Study 2. Both samples were predominantly white and moderately well-educated, as over 60% had at least a college education (Table 1). Approximately 90% were married or in a long-term relationship. The men had generally similar demographic characteristics.

Distribution of Scores by Day and Mode of Test Administration

Table 2 shows the agreement analysis, including the means and standard deviations for the PDQ-Q4 and FACIT-F in Study 1 and SF-36, PANAS, and PHQ-9 in Study 2 by mode and day of administration. The mean IVR scores closely matched the corresponding paper responses in Study 1 for the PDQ-Q4 (1.49 ± 1.55 vs 1.44 ± 1.57) and FACIT-F (40.33 ± 7.89 vs 40.85 ± 8.35). Mean scores and standard deviations for the variables in Study 2 (SF-36; PHQ-9 Depression, PANAS) were similarly well-matched.

Concordance of IVR and Paper Versions

Complete concordance results are shown in Table 2 and Figures 1–5. There were no significant differences in the paper and IVR modes of administration for the PDQ-Q4, FACIT-F, SF-36, or PANAS positive. Although the Wilcoxon test for differences between IVR and paper responses on the PDQ-Q4 was nearly significant ($p=0.06$), the magnitude of the difference, on average 0.05 ± 0.19, was quite small and not clinically relevant. For the negative affect sub-scale of the PANAS mean ± SD of the difference between IVR and paper versions was 0.18 ± 1.24 ($p=0.45$) on day 1 and 0.25 ± 0.75 ($p=0.04$) on day 2. Discrepancies between the IVR and paper versions of the PHQ-9 were also larger than observed for other measures, with mean ± SD difference between IVR and paper on day 1 was 0.68 ± 2.52 ($p=0.09$) and 1.09 ± 2.58 ($p<0.01$) on day 2, suggesting significantly higher IVR scores than paper scores on day 2. Spearman correlation coefficients, however, between IVR and paper results were high for each assessment (0.76–0.97, $p<0.0001$ for all measures), including the PANAS negative and PHQ-9. As there was no significant difference

between paper and IVR responses for the PDQ-Q4, FACIT-F or PANAS positive, no trend was observed for the difference in IVR and paper responses relative to the mean response. For the SF-36, the correlation of the difference and mean responses was r=0.35 (p=0.02) on day 2, but differences between IVR and paper responses were not significantly different from 0. The PANAS negative showed evidence of larger differences between paper and IVR with higher response values on day 2 but not day 1 (r=0.32, p=0.04); PHQ-9 responses trended toward larger differences for higher responses although correlations were not significant on either day 1 (r=0.23, p=0.12) or day 2 (r=0.28, p=0.06).

Bland and Altman plots showed a small, tolerable number of points falling outside of the limits of agreement for the PDQ-Q4 (Figure 1), FACIT-F (Figure 2), and PANAS positive (Figure 4a). SF-36 Bland-Altman plots for agreement between IVR and paper versions (Figure 3) showed several outliers with large discrepancies between IVR and paper responses, but no evidence of systematic bias of IVR relative to paper was observed. Several outliers with greater IVR scores than paper scores were also noted for the PANAS negative (Figure 4b), but Bland and Altman plots showed reasonable concordance with most points falling within the 95% limits of agreement. The Bland and Altman plots for the PHQ-9 agreement between IVR and paper versions show a fair amount of deviation from the perfect agreement 1-1 line, with several points having much higher IVR scores than paper scores on day 2 (Figure 5). These points likely contributed heavily to the significant overall difference observed between IVR and paper on day 2. Despite such discrepancies, differences were largely within 95% limits of agreement (Figure 5).

Test-Retest Reliability

Complete test-retest reliability results are shown in Table 2. Spearman correlation coefficients for test-retest ranged from 0.57 for 0.76 for IVR administration and 0.59 to 0.76 for paper administration. Correlations were generally similar between IVR and paper administrations for each assessment. The largest difference in test-retest reliability between IVR and paper versions was observed for the PANAS positive subscale, for which the Spearman correlation coefficient between days 1 and 2 was=0.57 ($p<0.0001$) for IVR and r=0.75 ($p<0.0001$) for paper.

Internal Consistency by Measure

Internal consistency as measured by Cronbach's alpha was similarly high for paper and IVR assessments. For the PDQ-Q4, FACIT F, PANAS positive, and PHQ-9 Cronbach's alpha was 0.80–0.9 for test and retest of both paper and IVR versions, except for the PDQ-Q4, which had an initial test but no retest. On day 1 of the SF-36 Cronbach's alpha was 0.67 and 0.80 on for IVR and paper, respectively, but day 2 values were higher at 0.91 and 0.86 for IVR and paper, respectively. For the PANAS negative internal consistency was a bit lower (Cronbach's alpha day 1 0.73 IVR, 0.71 paper), particularly on day 2 (Cronbach's alpha 0.68 IVR, 0.65 paper), but similar for IVR and paper.

DISCUSSION

IVR methods of data collection have important advantages over traditional paper versions of administration in the accurate recording of time and date, standardization of administration across conditions, built-in data recording and transfer opportunities, and minimization of data entry errors. However, it is important to show that different methods of data collection for subjective endpoints can yield comparable and reliable results. Overall, we observed few differences in the results of IVR administered questionnaires compared to traiditional paper-and-pencil formats. Two studies, instead of one, were performed to avoid excessive participant burden and to facilitate recruitment and timely completion. Results were consistent across and within studies that IVR administration results in highly reliable and accurate data which may have benefits in longer-term studies compared to traditional paper versions of the test.

Results were generally consistent across the studies and showed that both IVR and paper formats are acceptable to patients, feasible for administration in a clinical trial setting, and provide reliable means of assessing these domains. All outcome measures generally had high rates of completion in both formats. The mean values and range of scores observed were consistent with values reported in previous publications (1, 3). The level of agreement between IVR formats and traditional paper versions of the questionnaires was as high or higher for most measures compared to previous studies (3), and discrepancies that were identified were generally small and not clinically meaningful. The IVR format showed equal or slightly better performance overall on most of our tests. Moreover, we observed between a very high level of test-retest reliability and internal consistency and overall strong performance of the IVR-administered versions of all of the measures.

Some limitations of the studies should be noted. Sample sizes were relatively small, and our period of evaluation was relatively brief. Over a longer duration of follow up, adherence rates would likely have declined, affecting other important reliability and validity comparisons. Longer duration validation studies are needed to assess potential drift in adherence or reliability. The measures were only administered in English, so the results would not necessarily apply to other languages.

We conclude that administering the PDQ-Q4, FACIT-F, SF-36, PANAS, and PHQ-9 by IVR yields results that are similar to those by paper. Despite potential limitations in using IVR for older participants with hearing or comprehension difficulties, the potential benefits of IVR administration, including enhanced privacy, potentially improved adherence and real-time monitoring of response dates and times, are key factors which led to the IVR format being selected for use in The Testosterone Trials.

Acknowledgments

The Testosterone Trials are supported by a grant from the National Institute on Aging, National Institutes of Health [U01 AG030644, R01 AG037679 (Bone Trial)]. AbbVie (formerly Solvay and Abbott Laboratories) generously provided additional funding, AndroGel and placebo gel. UAB Diabetes Research and Training Center (DRCT), Grant Number DK079626 from the National Institute for Diabetes, Digestive and Kidney Diseases, National Institutes of Health. The University of Florida site is supported by the Claude D. Pepper Older Americans Independence Center [NIH/NIA P30AG028740]. Funding for Rancho Bernardo Study has been supported by

National Institutes of Health/National Institute on Aging grants AG07181 and AG028507 and the National Institute of Diabetes and Digestive and Kidney Diseases, grant DK31801.

References

1. Kobak KA, Greist JH, Jefferson JW, Katzelnick DJ, Mundt JC. New technologies to improve clinical trials. J Clin Psychopharmacol. 2001; 21(3):255–256. [PubMed: 11386486]
2. Piette JD. Interactive voice response systems in the diagnosis and management of chronic disease. Am J Manag Care. 2000; 6(7):817–827. [PubMed: 11067378]
3. Corkrey R, Parkinson L. Interactive voice response: review of studies 1989–2000. Behav Res Methods Instrum Comput. 2002; 34(3):342–353. [PubMed: 12395550]
4. Midanik LT, Greenfield TK. Interactive voice response versus computer-assisted telephone interviewing (CATI) surveys and sensitive questions: the 2005 National Alcohol Survey. J Stud Alcohol Drugs. 2008; 69(4):580–588. [PubMed: 18612574]
5. Mundt JC, Kobak KA, Taylor LV, Mantle JM, Jefferson JW, Katzelnick DJ, et al. Administration of the Hamilton Depression Rating Scale using interactive voice response technology. MD Comput. 1998; 15(1):31–39. [PubMed: 9458661]
6. Wang-Weigand S, Watissee M, Roth T. Use of a post-sleep questionnaire-interactive voice response system (PSQ-IVRS) to evaluate the subjective sleep effects of ramelteon in adults with chronic insomnia. Sleep Med. 2011; 12(9):920–923.10.1016/j.sleep.2011.06.008 [PubMed: 21925941]
7. Calloway M, Bharmal M, Hill-Zabala C, Allen R. Development and validation of a subjective post sleep diary (SPSD) to assess sleep status in subjects with restless legs syndrome. Sleep Med. 2011; 12(7):704–710.10.1016/j.sleep.2010.09.020 [PubMed: 21733752]
8. DeRogatis LR, Allgood A, Auerbach P, Eubank D, Greist J, Bharmal M, et al. Validation of a Women's Sexual Interest Diagnostic Interview--Short Form (WSID-SF) and a Daily Log of Sexual Activities (DLSA) in postmenopausal women with hypoactive sexual desire disorder. J Sex Med. 2010; 7(2 Pt 2):917–927.10.1111/j.1743-6109.2009.01528.x [PubMed: 19832932]
9. Osborn CY, Mulvaney SA. Development and feasibility of a text messaging and interactive voice response intervention for low-income, diverse adults with type 2 diabetes mellitus. J Diabetes Sci Technol. 2013; 7(3):612–622. [PubMed: 23759393]
10. Tucker JA, Simpson CA, Huang J, Roth DL, Stewart KE. Utility of an interactive voice response system to assess antiretroviral pharmacotherapy adherence among substance users living with HIV/AIDS in the rural South. AIDS Patient Care STDS. 2013; 27(5):280–286.10.1089/apc.2012.0322 [PubMed: 23651105]
11. Simpson CA, Xie L, Blum ER, Tucker JA. Agreement between prospective interactive voice response telephone reporting and structured recall reports of risk behaviors in rural substance users living with HIV/AIDS. Psychol Addict Behav. 2011; 25(1):185–190.10.1037/a0022725 [PubMed: 21443312]
12. Hettema JE, Hosseinbor S, Ingersoll KS. Feasibility and reliability of interactive voice response assessment of HIV medication adherence: research and clinical implications. HIV Clin Trials. 2012; 13(5):271–277.10.1310/hct1305-271 [PubMed: 23134627]
13. Dalal AA, Nelson L, Gilligan T, McLeod L, Lewis S, DeMuro-Mercon C. Evaluating patient-reported outcome measurement comparability between paper and alternate versions, using the lung function questionnaire as an example. Value Health. 2011; 14(5):712–720.10.1016/j.jval.2010.12.007 [PubMed: 21839410]
14. Skolarus TA, Holmes-Rovner M, Hawley ST, Dunn RL, Barr KL, Willard NR, et al. Monitoring quality of life among prostate cancer survivors: the feasibility of automated telephone assessment. Urology. 2012; 80(5):1021–1026.10.1016/j.urology.2012.07.038 [PubMed: 22990056]
15. Snyder PJ, Ellenberg SS, Cunningham GR, Matsumoto AM, Bhasin S, Barrett-Connor E, et al. The Testosterone Trials: Seven coordinated trials of testosterone treatment in elderly men. Clin Trials. 2014; 11(3):362–375.10.1177/1740774514524032 [PubMed: 24686158]
16. Lee KK, Berman N, Alexander GM, Hull L, Swerdloff RS, Wang C. A simple self-report diary for assessing psychosexual function in hypogonadal men. J Androl. 2003; 24(5):688–698. [PubMed: 12954659]

17. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. Journal of Rheumatology. 2005; 32:811–819. [PubMed: 15868614]
18. Ware, J.; Kosinski, M. SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1. 2. Lincoln, RI: Quality Metric Inc; 2005.
19. Ware, J.; Snow, K.; Kosinski, M. SF-36 Health Survey: Manual and Interpretation Guide. Lincoln, RI: Quality Metric Inc; 1993, 2000.
20. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol. 1988; 54(6):1063–1070. [PubMed: 3397865]
21. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001; 16(9):606–613. [PubMed: 11556941]
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476):307–310. [PubMed: 2868172]
23. Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951; 16(3): 297–334.

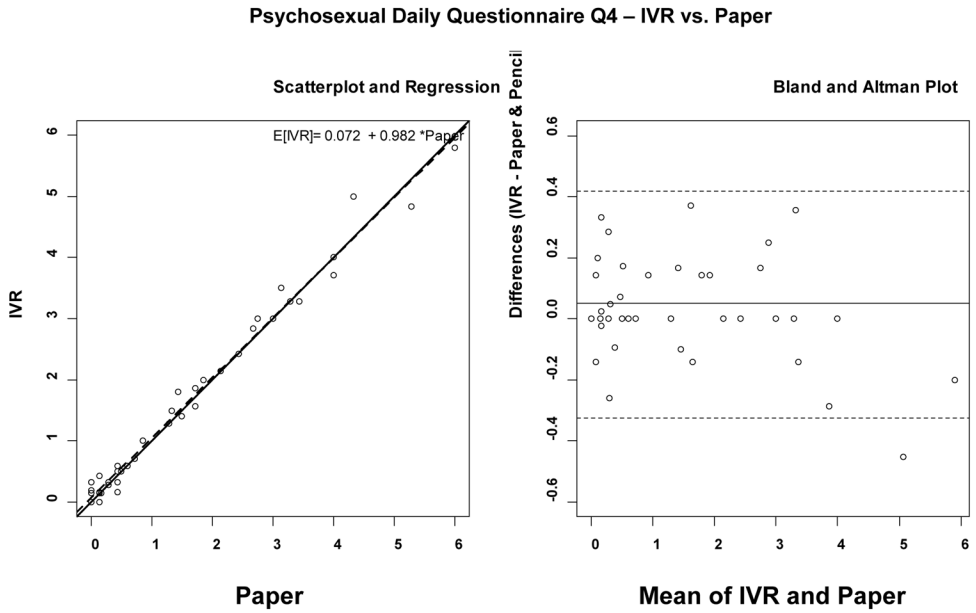

Fig. 1. Concordance of assessing the Psychosexual Daily Questionnaire, Question 4 (PDQ-Q4) by paper and by Interactive Voice Response (IVR)

In the left panel, the dotted line shows the best-fit regression line and the solid line the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panel, the dotted lines show the Bland-Altman interval bounds, and the solid line shows the mean difference of IVR vs. paper.

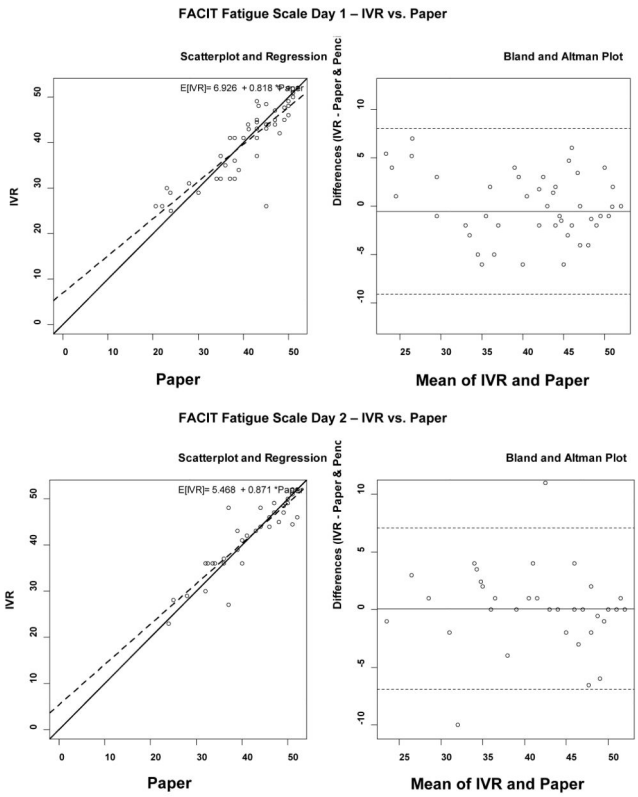

Fig. 2. Concordance of assessing the FACIT-Fatigue scale (FACIT-F) by paper and by Interactive Voice Response (IVR)

The top two panels show the results on day 1 and the bottom two panels show the results on day 2. In the left panels, the dotted line shows the best-fit regression line and the solid lines the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panels, the dotted lines show the Bland-Altman interval bounds, and the solid lines shows the mean difference of IVR vs. paper.

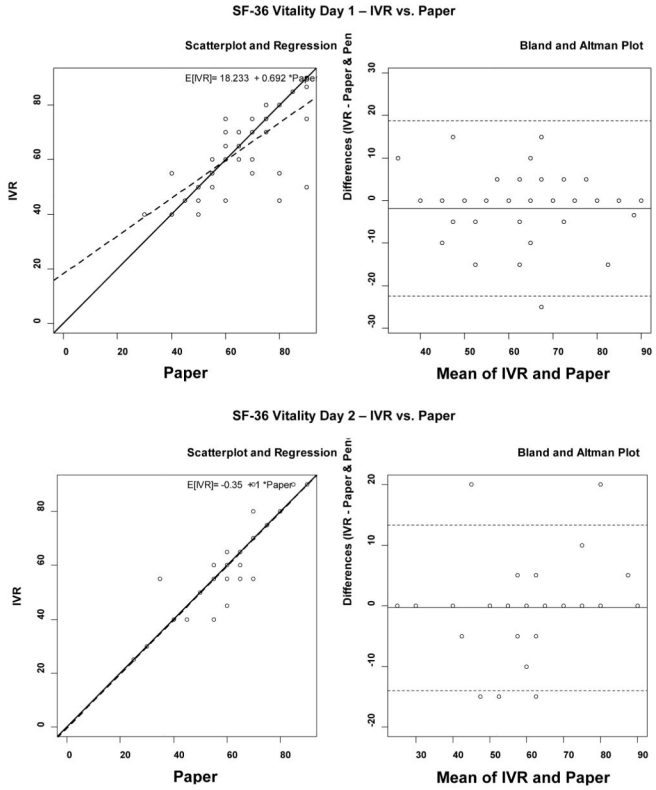

Fig. 3. Concordance of assessing the SF36 Vitality scale (SF-36)by paper and by Interactive Voice Response (IVR)

The top two panels show the results on day 1 and the bottom two panels show the results on day 2. In the left panels, the dotted line shows the best-fit regression line and the solid lines the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panels, the dotted lines show the Bland-Altman interval bounds, and the solid lines shows the mean difference of IVR vs. paper.

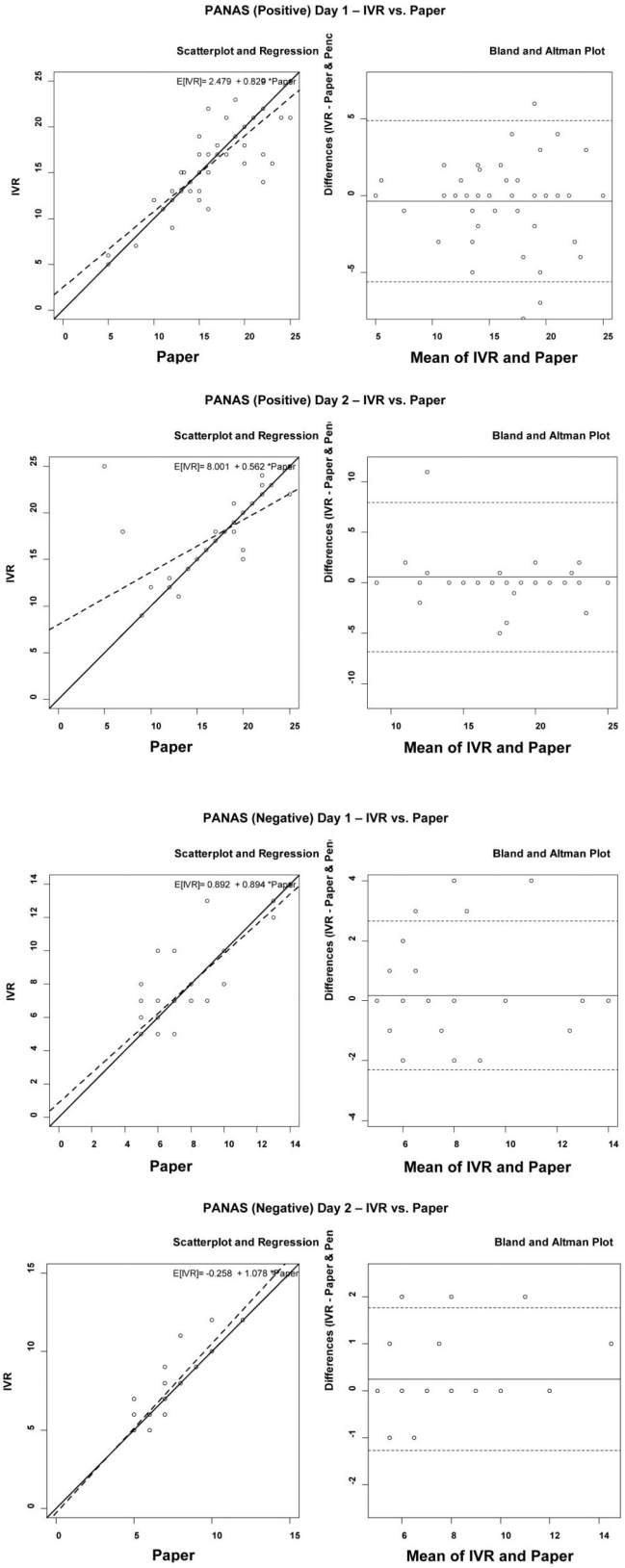

Fig 4.

Fig. 4a. Concordance of assessing the PANAS (positive) by paper and by Interactive Voice Response (IVR). The top two panels s how the results on day 1 and the bottom two panels show the results on day 2. In the left panels, the dotted line shows the best-fit regression line and the solid lines the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panels, the dotted lines show the Bland-Altman interval bounds, and the solid lines shows the mean difference of IVR vs. paper.

Fig. 4b. Concordance of assessing the PANAS (negative) by paper and by Interactive Voice Response (IVR). The top two panels show the results on day 1 and the bottom two panels show the results on day 2. In the left panels, the dotted line shows the best-fit regression line and the solid lines the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panels, the dotted lines show the Bland-Altman interval bounds, and the solid lines shows the mean difference of IVR vs. paper.

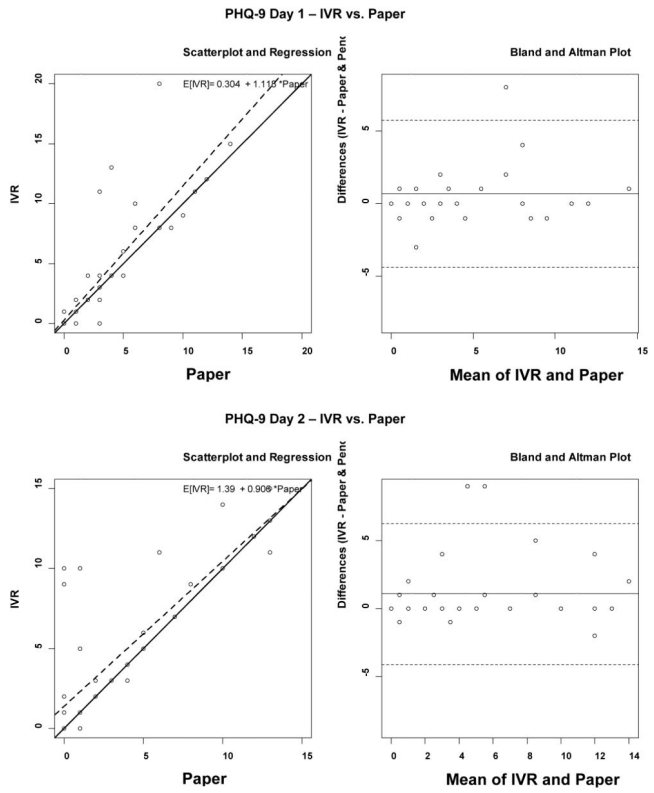

Fig. 5. Concordance of assessing the PHQ-9 by paper and by Interactive Voice Response (IVR)

The top two panels show the results on day 1 and the bottom two panels show the results on day 2. In the left panels, the dotted line shows the best-fit regression line and the solid lines the 1-1 line, where the 1-1 line indicates where points would fall if IVR and paper responses were identical for each participant. In the right panels, the dotted lines show the Bland-Altman interval bounds, and the solid lines shows the mean difference of IVR vs. paper.

Table 1

Participant Characteristics in IVR Validation Studies

Participant Characteristic		Study 1	Study 2
N		**50**	**51**
Age	Mean (SD)	75.9 (5.4)	75.5 (4.9)
Ethnicity	Not Hispanic/Latino	100%	100%
Race	White/Caucasian	96%	96%
	Black/African American	4%	4%
Marital Status	Currently Married	90%	92%
	Single/Never Married	–	2%
	Divorced	6%	2%
	Widowed	4%	4%
Education	Professional/Graduate degree	30%	37%
	College graduate	34%	23%
	Some college education	20%	12%
	High School graduate or equivalent	12%	10%
	Technical or Vocation School degree	4%	14%
	7^{th} – 12^{th} Grade	4%	4%

Table 2

Summary of Bland-Altman Agreement Analyses

		IVR		Paper		IVR vs. Paper	Difference (IVR –Paper)					Limits of Agreement		Difference vs. Mean	
Instrument	**N**	**Mean**	**SD**	**Mean**	**SD**	**Correlation**[a]	**Mean**	**SD**	**Min**	**Max**	**p-value**[b]	**Lower Bound**	**Upper Bound**	**Correlation**[a]	**p-value**
Psychosexual Daily Questionnaire – Q4	50	1.49	1.55	1.44	1.57	0.97	0.05	0.19	−0.45	0.67	0.06	−0.33	0.42	−0.05	0.74
FACIT-Fatigue Scale Day 1	50	40.33	7.89	40.85	8.35	0.87	−0.52	4.24	−19.00	7.00	0.58	−9.05	8.01	−0.07	0.61
FACIT-Fatigue Scale Day 7	37	42.00	7.73	41.92	8.03	0.88	0.08	3.45	−10.00	11.00	0.70	−6.91	7.06	−0.25	0.14
SF-36 Vitality Day 1	50	63.33	13.44	65.20	14.07	0.72	−1.87	10.24	−40.00	15.00	0.47	−22.44	18.71	0.00	0.97
SF-36 Vitality Day 2	46	63.26	16.57	63.59	15.12	0.93	−0.33	6.78	−15.00	20.00	0.72	−13.99	13.33	0.35	0.02
PANAS (positive) Day 1	50	16.30	4.56	16.67	4.61	0.80	−0.37	2.61	−8.00	6.00	0.43	−5.61	4.88	0.00	0.98
PANAS (positive) Day 2	44	17.55	4.00	16.98	4.58	0.77	0.57	3.66	−5.00	20.00	0.67	−6.82	7.95	−0.08	0.59
PANAS (negative) Day 1	49	6.86	2.52	6.67	2.47	0.81	0.18	1.24	−2.00	4.00	0.45	−2.30	2.67	−0.04	0.77
PANAS (negative) Day 2	44	6.77	2.34	6.52	2.06	0.91	0.25	0.75	−1.00	3.00	0.04	−1.26	1.76	0.32	0.04
PHQ-9 Day 1	50	3.86	4.66	3.18	3.52	0.92	0.68	2.52	−3.00	12.00	0.09	−4.38	5.74	0.23	0.12
PHQ-9 Day 2	46	4.33	4.38	3.24	3.93	0.76	1.09	2.58	−2.00	10.00	<.01	−4.11	6.28	0.28	0.06

[a]**Correlation**: Spearman correlation

[b]**p-value**: Wilcoxon Signed Rank test.