Published in final edited form as:
Qual Life Res. 2014 February ; 23(1): 217–227. doi:10.1007/s11136-013-0451-4.

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

Jakob B. Bjorner,
QualityMetric, Lincoln, RI, USA; Department of Public Health, University of Copenhagen, Copenhagen, Denmark; National Research Centre for the Working Environment, Copenhagen, Denmark

Matthias Rose,
Department of Psychosomatic Medicine and Psychotherapy, Medical Clinic, Charité, Universitätsmedizin, Berlin, Germany; Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

Barbara Gandek,
Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

Arthur A. Stone,
Department of Psychiatry and Behavioral Science, Stony Brook University, Stony Brook, NY, USA

Doerte U. Junghaenel, and
Department of Psychiatry and Behavioral Science, Stony Brook University, Stony Brook, NY, USA

John E. Ware Jr.
John Ware Research Group, Worcester, MA, USA; Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

Jakob B. Bjorner: jbj@nrcwe.dk

Abstract

Purpose—To test the impact of method of administration (MOA) on the measurement characteristics of items developed in the Patient-Reported Outcomes Measurement Information System (PROMIS).

Methods—Two non-overlapping parallel 8-item forms from each of three PROMIS domains (physical function, fatigue, and depression) were completed by 923 adults (age 18–89) with chronic obstructive pulmonary disease, depression, or rheumatoid arthritis. In a randomized crossover design, subjects answered one form by interactive voice response (IVR) technology, paper questionnaire (PQ), personal digital assistant (PDA), or personal computer (PC) on the Internet, and a second form by PC, in the same administration. Structural invariance, equivalence of item responses, and measurement precision were evaluated using confirmatory factor analysis and item response theory methods.

Correspondence to: Jakob B. Bjorner, jbj@nrcwe.dk.

Results—Multigroup confirmatory factor analysis supported equivalence of factor structure across MOA. Analyses by item response theory found no differences in item location parameters and strongly supported the equivalence of scores across MOA.

Conclusions—We found no statistically or clinically significant differences in score levels in IVR, PQ, or PDA administration as compared to PC. Availability of large item response theory-calibrated PROMIS item banks allowed for innovations in study design and analysis.

Keywords

Patient-reported outcomes; Quality of life; Questionnaire; Mode of administration; Method of administration; Item response theory

Introduction

Advances in survey data collection technologies are enabling substantial improvements in the measurement of patient-reported outcomes (PRO). Technologies such as telephone interactive voice response (IVR), computer-based interfaces, and handheld devices enable electronic data capture, with many advantages including cost savings for data collection and processing and score estimation in real time. Additionally, electronic data capture technologies are flexible, allowing for use of static (pre-selected) or dynamic (matched to the respondent) selection of survey items. However, the migration from paper–pencil to electronic data capture technology may influence item responses in ways that are not related to the health concept being measured. These effects of differences in methods of administration warrant further study.

A meta-analysis of 65 studies of method of administration (MOA) effects in PRO measurement found that the average absolute mean difference was 1.7 % of the score range (approximately 1.7 points on a 0–100 scale) for comparisons of paper questionnaire (PQ) and personal computer (PC) MOA and 2.4 % of the score range for comparison of PQ and personal digital assistant (PDA). While some inconsistencies were found, the mean difference was within +/− 5 % for 93 % of the studies [1]. For studies evaluating PQ and computerized MOA on the same persons, the weighted summary correlation between MOA was 0.90 (95 % CI 0.87–0.92) [1], not significantly smaller than test–retest reliabilities in the studies where this was examined. Thus, the meta-analysis and subsequent studies [2, 3] generally support the equivalence of PQ, PC, and PDA MOA [1].

Several studies have compared PQ and phone interview MOA [4–12], but fewer studies have assessed IVR technology for health outcomes measurement. In this field, the studies have found only non-significant [13] or small score differences [14, 15] between PQ and IVR and high agreement between the two MOA [14].

Thus, most studies support equivalence of self-administered PQ, PC, PDA, and IVR MOA. However, many studies were small and usually did not include explicit statements on the minimal important difference or the power of the study to evaluate equivalence. Further, most studies have compared only two modes and have not performed a comprehensive evaluation of equivalence across the most frequently used modes.

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a National Institutes of Health (NIH) Roadmap initiative that began in 2004 with the goal of generating improved PRO measures for use in clinical research and practice (http://www.nihpromis.org). Using the latest test development procedures and extensive input from patients, large item banks have been created to measure common PRO domains, including physical function, fatigue, pain, social role, and emotional distress [16]. A number of brief

questionnaires (short forms) and more sophisticated assessments using computerized adaptive testing (CAT) software were constructed using item response theory (IRT). In addition to developing item banks and conducting studies to evaluate the validity of the new instruments, PROMIS has also carried out studies on topics pertinent to all self-report assessment, for example, on various retrospective recall periods, accessibility to physically challenged individuals, and MOA, the topic of this paper [17–19].

The present study was designed to examine how differences in MOA affect psychometric properties and score differences and to evaluate the consistency of any differences between MOA across PROMIS health domains, using alternate forms constructed from the PROMIS item banks. Four MOA were compared: PQ, IVR, PC, and PDA. The study has two main purposes: to test equivalence across MOA, and if MOA effects are found, to estimate the magnitude of the MOA effects to allow calibration of scores across MOA.

Methods

Two substudies were conducted. Both used a randomized cross-over design, in which two non-overlapping parallel forms (Form A and Form B) consisting of eight items from each of three PROMIS domains (physical function (PF), fatigue (FAT), and depression (DEP)) were administered. To limit response burden, we restricted the study to three domains. Based on the theoretical framework of the physical to mental health continuum [20], we selected the domain that is the strongest measure of physical health (physical function), the domain that is the strongest measure of mental health (depression), and a domain—fatigue—that reflects both physical and mental health. In study 1, participants were randomized to complete one form by IVR, PQ, or PC, and the second form by PC. The order in which forms were administered and the combination of form and MOA were randomized. In Study 2, all participants were assessed by PDA and PC. This study was performed separately, because the PDA administration required in-person contact. The order of forms and the combination of form and MOA were randomized. The overall design of the study is presented in Fig. 1.

Sample and procedures

Study 1—Data for the IVR–PC, PQ–PC, and PC–PC arms were collected by YouGovPolimetrix, an Internet panel company [21]. YouGovPolimetrix contacted panelists age 18 or older who were fluent in English and had previously indicated that they had rheumatoid arthritis, chronic obstructive pulmonary disease (COPD), or depression. These three disease groups were chosen to represent a broad spectrum of conditions that have well-documented impact on the selected PRO domains [22]. Subjects had to verify that they were diagnosed by a physician and taking diagnosis-specific medication (and for depression were undergoing treatment by a mental health professional). We stratified sampling to achieve equal representation of each group. Subjects reporting more than one condition were randomly assigned to one diagnostic group.

All participants started the assessment on a PC, were screened for eligibility and consented, and answered one item about the impact of their disease on everyday life. To ensure a sufficient distribution of impairment within each diagnostic group, a quota was imposed aiming to achieve equal representation of low, medium, and severe disease impact. This study included 723 persons, well above the target sample of 600 (see Fig. 1).

After qualifying for the study, participants were randomized to study arm through computer-generated random numbers (Fig. 1). If the participant was randomized to an arm where the PC MOA was first and the PQ or IVR MOA was second, a PROMIS static form and user experience questionnaire were administered by PC, followed by sociodemographic and health literacy items. Depending on randomization, the subject then was instructed to

complete a previously mailed paper–pencil questionnaire or call a toll-free number for the IVR assessment to complete the alternate PROMIS form. After completing the second PROMIS form, the subject returned to the PC and completed the study (Fig. 2). If the subject was randomized to an arm where the PQ or IVR MOA was first and the PC MOA was second, the subject completed a PROMIS form by PQ or IVR first, then completed the first user experience questionnaire and all remaining items by PC (Fig. 2). Subjects assigned to the PQ arm (Fig. 1) were requested to mail back the form after completing both assessments. If the participant was randomized to a PC-PC arm, all assessments were completed by PC. Presentation of items followed PROMIS conventions; the PC administration displayed one item per screen, while the PQ layout grouped items with the same response categories together in a grid. IVR recordings were developed for the PROMIS initiative using a female voice.

Study 2—Data for the PDA–PC arms were collected through a multiphysician rheumatology practice on Long Island. Two hundred individuals participated. Eligibility criteria were (1) rheumatology patient at the practice, (2) age 18 or older, (3) fluent in English, (4) able to hold a writing implement, and (5) no visual impairment that would interfere with study participation. Recruitment through flyers and posters was conducted in the waiting room of the practice with the help of a trained research assistant. Participants could complete the study before or after their doctor's visit, or on a day where they did not have a scheduled appointment but were willing to come to the practice. Upon consent, participants were asked to complete the PDA–PC sequence. The order of survey elements was similar to Study 1. Randomization to one of four arms (order of MOA and order of form, Fig. 1) was accomplished by the research assistant opening the next envelope in a numbered sequence that had one of the four administration orders. The two assessments were separated by a short interval (e.g., 5–10 min), to allow participants to switch from one MOA to another and answer health literacy and sociodemographic questions. Both PC and PDA displayed one item per screen. Participants were compensated with up to $50 for participation.

Both studies were approved by the New England Institutional Review Board (#09-107). Further, Study 2 was approved by the Stony Brook University Institutional Review Board (#2008–0280).

Measures

Parallel static forms—Item response theory (IRT) [23, 24] methods were used to develop two parallel static short forms containing eight non-overlapping items from each of three PROMIS item banks (physical function, fatigue, and depression). The item banks had been shown to satisfy typical psychometric assumptions [25] and had been calibrated, which eliminated the need to equate parameters and scores [26]. According to IRT theory, each individual should receive the same estimated score on both short forms from the same domain, and the PC–PC arm allowed for evaluation of this assumption. However, due to the balanced design, the study was robust to departures from perfectly parallel forms.

Our goal was to construct parallel static forms that reflected the content of the larger PROMIS item banks. We selected items for each domain such that the number of items per content category within each form was proportional to the number of items per category in the full item bank. The categories were as follows: upper, central, and lower extremity functions and instrumental activities of daily living (for physical function), experience and impact (for fatigue), and mood and cognition (for depression). In addition to these content validity considerations, we used the PROMIS IRT item parameters to select items within each domain so that the parallel forms had similar test information functions. We used

PROMIS Wave 1 data to check that the parallel forms provided equivalent score estimates and had equivalent known groups validity (results not shown).

Subjects also provided information on demographics, health care utilization and previous computer use, and were screened for impaired health literacy status using three items shown to be effective at detecting inadequate health literacy in relation to the Short Test of Functional Health Literacy in Adults [27].

Analyses

A detailed data analysis protocol was approved by the PROMIS Steering Committee. This paper reports on structural invariance and equivalence of item responses across MOA as evaluated by confirmatory factor analysis and IRT models. The same analytic approach was applied to the comparison of PQ, IVR, and PC MOA and the comparison of the PDA and PC MOA. However, because selection criteria for the PDA arms were different than for the other MOA, the PDA arms were analyzed separately.

Structural Invariance was evaluated by multigroup confirmatory factor analyses, using MOA to define each group. A separate analysis was conducted for each form within each domain. The analyses used the WLSMV estimator as implemented in the program Mplus V5 [28]. The analyses tested the equivalence of factor loadings and item thresholds across MOA using chi-squared difference tests of nested models [28]. We adjusted the significance level using the Hochberg approach [29] to take the number of tests (24) into account.

The potential effect of MOA on *item response* was evaluated using IRT methods. This approach allowed us to separately estimate the impact of MOA on score level and on score precision, while controlling for order effects. These analyses used an extension of the graded response IRT model fitted in SAS using the NLMIXED procedure (see “Appendix”). We used the item parameters estimated in the PROMIS item bank development [21] as fixed constants, but we also evaluated an alternative model, where the item parameters for each item were estimated in the current sample. We tested the potential impact of MOA on score level by estimating a MOA-specific adjustment to the IRT threshold parameters and tested the potential impact of MOA on score precision/variance by estimating a MOA-specific adjustment to the IRT slope parameter. When evaluating significance of these parameters, we adjusted for multiple comparisons [29].

According to standard PROMIS practice, the IRT item parameters were standardized so that the latent score distribution in the general population was standard normal (Mean = 0, SD = 1), while the IRT score is reported in a *T*-score metric, which sets the general population mean to 50 and the standard deviation to 10. The study was powered to evaluate equivalence across MOA within a difference in threshold parameters of 0.2 (corresponding to a 2-point difference in the IRT score using a T-score metric) with a power of 0.85. This difference, corresponding to the 0.2 effect size suggested by Cohen and Cohen [30] as the smallest relevant effect, was chosen since we believed that potential MOA effects should be smaller than score differences that would be considered clinically relevant.

Results

Participants ranged from 18 to 89 years, with a mean of 56 years (SD = 13). A majority of participants in the rheumatoid arthritis and depression groups were women, while the COPD group had equal gender distribution. More than 90 % of participants were white and a little more than half were married. Most participants had at least some college education. About 24 % were full-time employees, 24 % were on disability, and 27 % were retired (Table 1). The range of scale correlations (across Forms A and B and across the two substudies) were

as follows: PF and FAT r = -0.73 to -0.59, PF and DEP r = -0.37 to -0.26, and FAT and DEP r = 0.50 to 0.57.

Multigroup confirmatory factor analyses supported equivalence of the factor structure across MOA. Table 2 shows results for tests of equality of loadings and thresholds across all evaluated domains and forms. Specifying separate thresholds or loadings for each MOA did not lead to significant improvement in fit. The smallest *p*-value (0.04), concerning the equality of thresholds for fatigue Form A in Study 1, was far from significant, considering the number of comparisons.

Table 3 reports the results concerning the estimated impact of MOA on score level and on score precision. A negative effect on the location parameter indicates that the item is "easier" in the particular MOA in the sense that participants, all other things being equal, would tend to pick a response choice with a higher score. For physical function, a higher item score indicates better physical function; for fatigue and depression, a higher item score indicates more severe or frequent symptoms. The table shows no significant effect of MOA on item location, except for the effect of PDA MOA on physical functioning items. PDA MOA makes an item harder so that participants on average will score lower on physical function items. This effect is significant at a 0.05 level, but not after adjustment for multiple comparisons.

Figure 3 illustrates what the implications of the estimated item parameter differences would be, if they were taken as true differences and used for score adjustments. The figure shows the adjusted PROMIS scores for a person providing a response combination that would result in a score of 50.0 if provided by PC. For physical function, for example, the adjusted score for PDA would be 50.4 (95 % CI 50.0–50.8) since physical function items are slightly harder when administered by PDA (see Table 3). This potential adjustment is far below the pre-specified minimal important difference (shown by the vertical broken lines), indicating that the implied mean score levels are equivalent. Since our model assumes a constant methods effect across score levels, the same adjustments would apply to other score levels (e.g., 30 or 60).

Table 3 also reports the effect of MOA on IRT slopes. A positive number indicates that an item is more discriminant for the particular MOA than for the PC MOA, while a negative number indicates less discrimination for the particular MOA. The table shows three significant results concerning slopes: PQ MOA and PDA MOA result in significantly higher discrimination for fatigue items, while IVR MOA results in significantly less discrimination for depression items.

Since no minimal important difference was specified for slope effects prior to analysis, the potential clinical significance of these results was evaluated by post hoc analyses. For example, Fig. 4 shows the item category response functions for the item "*Felt nothing could cheer me up*" for PC and IVR MOA. While the functions are slightly flatter for the IVR MOA, the difference seems negligible. Similar results were observed for other items.

Table 3 also reports the estimated effect on the item parameters if the item is placed in the second part of the form as opposed to the first part of the form. For the purpose of MOA evaluation, these parameters can be regarded as nuisance parameters. Several significant, although small, effects are seen on thresholds and several significant effects are seen on the item slopes. Finally, Table 3 shows the estimated IRT score means and standard deviations for the four clinical groups. In Study 1, the rheumatoid arthritis and COPD groups predictably had the worst physical function scores and the rheumatoid arthritis group had the worst fatigue scores, while the depression group had the worst depression score. Participants

in Study 2 generally had better scores; in particular, the depression score is close to the general population mean.

We tested the robustness of the results in Table 3 in several ways. (1) *Dependence on standard PROMIS item parameters* was evaluated by reestimating the item parameters in the current sample and then rerunning the MOA analyses. No additional significant results were found, the estimates of MOA effects on location changed very little, and the estimates of effects on slope generally changed slightly toward zero (results not shown). (2) *Possible MOA effects on single items* were evaluated by estimating a separate effect of MOA for each item. A borderline significant effect was found for one item, but this effect was non-significant after adjusting for multiple comparisons. (3) We performed *subgroup* analyses focusing on three groups for which electronic data capture might be problematic: persons 60 years or older (n = 275), persons with at most a high school or GED education (n = 142), and persons with low health literacy (n = 235). After adjustment for multiple comparisons, subgroup results were similar to the results in the total sample, with the exception that PQ items on depression had lower discrimination among those with lower education.

Discussion

This study of the effects of different MOA within the domains of physical function, fatigue, and depression was conducted as part of the PROMIS initiative [16]. We found neither statistically nor clinically significant effects on mean score levels of PQ, IVR, or PDA administration as compared to PC administration. Thus, in line with the rather substantial number of studies comparing PQ with PC or PDA [1] and with the more limited number of studies comparing IVR and PQ [31], our results provide strong support for the equivalence of scores from PC, PQ, IVR, and PDA MOA.

We found a few significant effects of MOA on score precision as assessed by IRT discrimination parameters. In particular, item discrimination was significantly lower for IVR administration in the depression domain. These results suggest a slightly lower score precision for IVR administration. The impact of this lower IRT discrimination for overall scale performance should be investigated in future studies. Results of satisfaction surveys in a non-health context found that respondents to aural MOA (telephone interviews and IVR) were significantly more likely to provide extreme responses [32]. In an IRT context, more extreme responses would be seen if the item threshold parameters were clustered more closely together for each item and the item discrimination was higher. This was not the pattern found in our study; the differences may be due to the different concepts studied or the way IVR was implemented.

Our study differs from other studies of mode effects in a number of aspects that represent both strengths and weaknesses.

Sample

While most other studies have sampled from a limited number of clinical centers, Study 1 used an internet sample with explicit criteria for the selected clinical diagnoses. Judging from the score levels on the outcome scales, the three groups were indeed severely impacted by their disease. Thus, in line with standard recommendations [31], we believe that our sample represents the most important future users of PROMIS tools: patients in clinical trials. It cannot be ruled out that our Study 1 sample has more familiarity with and skills in computers, compared to standard patient groups. However, subgroup analyses of elderly patients, patients with little education, and patients with low health literacy did not provide results that differed notably from the results in the total sample.

Parallel forms study design

While most controlled intervention studies on MOA effects either use a cross-sectional design or a test–retest design, the existence of large IRT-calibrated item banks within PROMIS allowed us to use a parallel forms design. In a cross-sectional design, MOA differences can be due to either MOA-specific response style or MOA-specific non-completion. Our design eliminated the possibility of MOA-specific non-completion, which is the most appropriate approach if you want to combine results from different MOA within a single study. The parallel forms design avoids potential problems of the test–retest design such as participants' recollection of previous responses or a change in health between assessments. Still the parallel forms design is as powerful as the test–retest design, since analyses can be done through within-person comparisons. We also used balanced randomization to avoid possible confounds by order of administration or if the forms were not completely parallel. Finally, we performed power analyses to make sure that the study had sufficient power to evaluate equivalence between MOA within the specified minimal important difference. In fact, since we used conservative estimates for the reliability of the forms, our study had even more power, as witnessed by the narrow confidence intervals for MOA effects (Fig. 3).

Data analysis

The characterization of MOA effects as potential adjustments to the IRT parameters allows for independent evaluation of MOA effects on score level and on score precision. To be able to estimate the MOA effects and simultaneously control for order of administration, we had to expand standard IRT models and develop an estimation program. While experience with this type of analysis is limited, we believe it is a powerful approach with clear advantages over analysis of mean score levels or ICC reliabilities, since the evaluation of score levels and score precision is performed simultaneously, but with two distinct sets of parameters.

In our opinion, there is strong support for the generalization of these results to CAT for the three PROMIS domains: physical function, fatigue, and depression. While the use of a PQ comparison necessitated the use of fixed short forms in our study, the test experience for the participants was exactly the same for the electronic MOA as if a CAT had been used. The only issue that could cause MOA effects in a CAT that would be missed by our study would be if a group of items that would be prone to MOA effects was not included in the static forms used in this study. However, the static forms were developed to represent all the subdomains found in the PROMIS item banks. Further, we found no indications of MOA effects pertaining to particular items or subgroups of items.

Our findings also suggest that the MOA results may generalize to the other PROMIS domains. In studies where MOA effects have been found—in particular studies comparing PQ and phone interviews [10]—the MOA effect seems to particularly concern domains related to mental health. For this reason, we selected the three domains of our study to represent both physical and mental health. The fact that no major MOA effects were found over a very diverse set of health outcomes supports the position that no MOA effects may exist for PROMIS domains such as role participation, pain, anxiety, and anger. A further theoretical cause of MOA effects would be response choices that were well suited for some MOA, but not for others. However, since PROMIS researchers have decided on a limited number of standardized response choices that are used across most domains, the response choices tested in this study are also the ones used in most other PROMIS domains.

Finally, there is the issue of whether the results can be generalized to other groups, in particular to participants with other levels of health. The general recommendation for MOA studies is to use a study population similar to the intended users of the instrument in

subsequent research or in clinical work [31]. For this reason, we selected a mixed population for Study 1, including patients with somatic and mental disorders. The clinical differences between these groups were clearly seen from their mean scores on the three outcome measures (Table 3). Thus, we believe that the results can be generalized across a broad range of clinical conditions.

The issue of equivalence for PROMIS tools for personal or phone interviews using a 'live' interviewer is still not settled. It is for these MOA that lack of equivalence has been documented most consistently, and the frameworks of social desirability and interviewer style provide good theoretical explanations as to why MOA effects could happen. We caution that we did not evaluate MOA using a live interviewer and we cannot provide any insights about this MOA.

In conclusion, our results provide strong support for the equivalence of score levels from the evaluated MOA: PC, PQ, IVR, and PDA. This conclusion is in line with the rather substantial number of studies comparing PQ with PC or PDA [1] and with the more limited number of studies comparing IVR and PQ [31].

Acknowledgments

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Northwestern University PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztein, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. The authors would like to thank two anonymous PROMIS reviewers and two journal reviewers for comments on a previous version of this manuscript. See the web site at www.nihpromis.org for additional information on the PROMIS cooperative group.

Appendix

The standard graded response IRT model can be formulated:

$$\log\left(\frac{P(x_{ji}\geq c)}{P(x_{ji}<c)}\right)=\alpha_i\,(\theta_j-(\lambda_i-\tau_{ic}))$$

where θ_j, is the latent health of person *j*: (here: physical functioning, fatigue, or depression), a_i is the discrimination parameter for item *i*, λ_i is the location parameter for item *i*, and τ_{ic} is the item category parameter. An extended graded response model can be formulated in the following way:

$$\log\left(\frac{P(x_{jiopqa}\geq c)}{P(x_{jiopqa}<c)}\right)=(\alpha_i+\alpha_o+\alpha_p+\alpha_q)\,(\theta_j-(\lambda_i+\lambda_o+\lambda_p+\lambda_q-\tau_{ic}))$$

where a_o, λ_o represents the potential effect of *item order* (being administered in the second part of the form as opposed to the first) on item discrimination and location parameters. α_p, λ_p represents the potential effect of IVR *phone administration* (as opposed to Internet

administration). α_p, λ_q represents the potential effect of *paper & pencil questionnaire administration* (as opposed to Internet administration).

The model was estimated using SAS proc MLMIXED. The item parameters α_i, λ_i, and τ_{ic} were initially treated as known constants and fixed to the values estimated in the PROMIS item bank development calibrations. In additional analyses, α_i, λ_i, and τ_{ic} were estimated for each item using the current sample. The mean and standard deviation of θ was estimated separately for each diagnostic group.

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 Health. 2008; 11(2):322–333. [PubMed: 18380645]
2. Raat H, Mangunkusumo RT, Landgraf JM, et al. Feasibility, reliability, and validity of adolescent health status measurement by the Child Health Questionnaire Child Form (CHQ-CF): Internet administration compared with the standard paper version. Quality of Life Research. 2007; 16(4): 675–685. [PubMed: 17286197]
3. Yu SC. Comparison of Internet-based and paper-based questionnaires in Taiwan using multisample invariance approach. CyberPsychology & Behavior. 2007; 10(4):501–507. [PubMed: 17711357]
4. Duncan P, Reker D, Kwon S, et al. Measuring stroke impact with the Stroke Impact Scale: Telephone versus mail administration in veterans with stroke. Medical Care. 2005; 43(5):507–515. [PubMed: 15838417]
5. Hepner KA, Brown JA, Hays RD. Comparison of mail and telephone in assessing patient experiences in receiving care from medical group practices. Evaluation and the Health Professions. 2005; 28(4):377–389. [PubMed: 16272420]
6. de Vries H, Elliott MN, Hepner KA, et al. Equivalence of mail and telephone responses to the CAHPS Hospital Survey. Health Services Research. 2005; 40(6 Pt 2):2120–2139. [PubMed: 16316441]
7. Powers JR, Mishra G, Young AF. Differences in mail and telephone responses to self-rated health: Use of multiple imputation in correcting for response bias. Australian and New Zealand Journal of Public Health. 2005; 29(2):149–154. [PubMed: 15915619]
8. Beebe TJ, McRae JA, Harrison PA, et al. Mail surveys resulted in more reports of substance use than telephone surveys. Journal of Clinical Epidemiology. 2005; 58(4):421–424. [PubMed: 15868697]
9. Kraus L, Augustin R. Measuring alcohol consumption and alcohol-related problems: Comparison of responses from self-administered questionnaires and telephone interviews. Addiction. 2001; 96(3): 459–471. [PubMed: 11255585]
10. McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. Medical Care. 1994; 32(6):551–567. [PubMed: 8189774]
11. Hanmer J, Hays RD, Fryback DG. Mode of administration is important in US national estimates of health-related quality of life. Medical Care. 2007; 45(12):1171–1179. [PubMed: 18007167]
12. Hays RD, Kim S, Spritzer KL, et al. Effects of mode and order of administration on generic health-related quality of life scores. Value Health. 2009; 12(6):1035–1039. [PubMed: 19473334]
13. Agel J, Rockwood T, Mundt JC, et al. Comparison of interactive voice response and written self-administered patient surveys for clinical research. Orthopedics. 2001; 24(12):1155–1157. [PubMed: 11770093]
14. 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]
15. 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]

16. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. [PubMed: 17443116]
17. Broderick JE, Schwartz JE, Vikingstad G, et al. The accuracy of pain and fatigue items across different reporting periods. Pain. 2008; 139(1):146–157. [PubMed: 18455312]
18. Broderick JE, Schneider S, Schwartz JE, Stone AA. Interference with activities due to pain and fatigue: Accuracy of ratings across different reporting periods. Quality of Life Research. 2010; 19(8):1163–1170. [PubMed: 20535565]
19. Schneider S, Stone AA, Schwartz JE, Broderick JE. Peak and end effects in patients' daily recall of pain and fatigue: A within-subjects analysis. J Pain. 2011; 12(2):228–235. [PubMed: 20817615]
20. Ware JE Jr, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. Medical Care. 1995; 33(4 Suppl):AS264–AS279. [PubMed: 7723455]
21. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology. 2010; 63(11):1179–1194. [PubMed: 20685078]
22. Ware, JE., Jr; Snow, KK.; Kosinski, M.; Gandek, B. SF-36 health survey. Manual and interpretation guide. Boston: The Health institute, New England Medical Center; 1993.
23. Hambleton RK, Jones RW. An NCME Instructional Module on the comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice. 1993; 12(3):38–47.
24. van der Linden, WJ.; Hambleton, RK. Handbook of modern item response theory. New York: Springer; 1997.
25. Reeve BB, Hays RD, Bjorner JB, et al. 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 Suppl 1):S22–S31. [PubMed: 17443115]
26. Kolen, ML.; Brennan, RL. Test equating, scaling, and linking: Methods and practices. New York: Springer; 2004.
27. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Family Medicine. 2004; 36:588–594. [PubMed: 15343421]
28. Muthen, BO.; Muthen, L. Mplus user's guide. 5. Los Angeles: Muthén & Muthén; 2007.
29. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800–803.
30. Cohen, J. Statistical power for the behavioral sciences. Hillsdale NJ: Erlbaum; 1988.
31. 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 Health. 2009; 12(4): 419–429. [PubMed: 19900250]
32. Dillman DA, Phelps G, Tortora R, et al. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Social Science Research. 2009; 38:1–18.

Abbreviations

CAT	Computerized adaptive testing
COPD	Chronic obstructive pulmonary disease
DEP	Depression
FAT	Fatigue

IRT	Item response theory
IVR	Interactive voice response
MOA	Method of administration
PC	Personal computer
PDA	Personal digital assistant
PF	Physical functioning
PQ	Paper questionnaire
NLMIXED	SAS procedure for estimating mixed models
PRO	Patient-reported outcomes
PROMIS	Patient-Reported Outcomes Measurement Information System
WLSMV	Weighted least squares with mean and variance adjustment

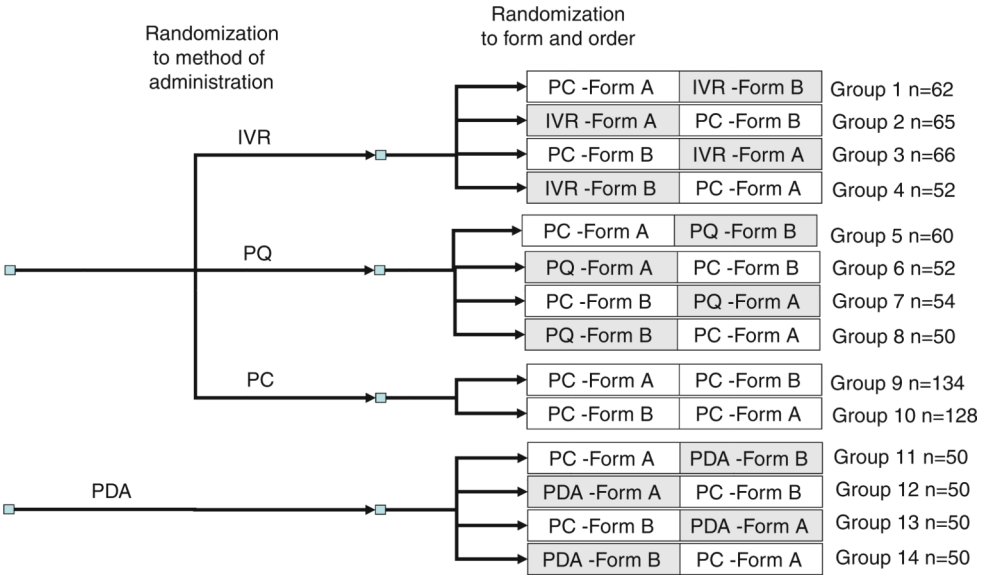

Fig. 1.

Study design and sample size. Target sample size = 50 for each group except groups 9 and 10 where target sample size = 100. *PC* personal computer, *IVR* interactive voice response, *PQ* paper questionnaire, *PDA* personal digital assistant

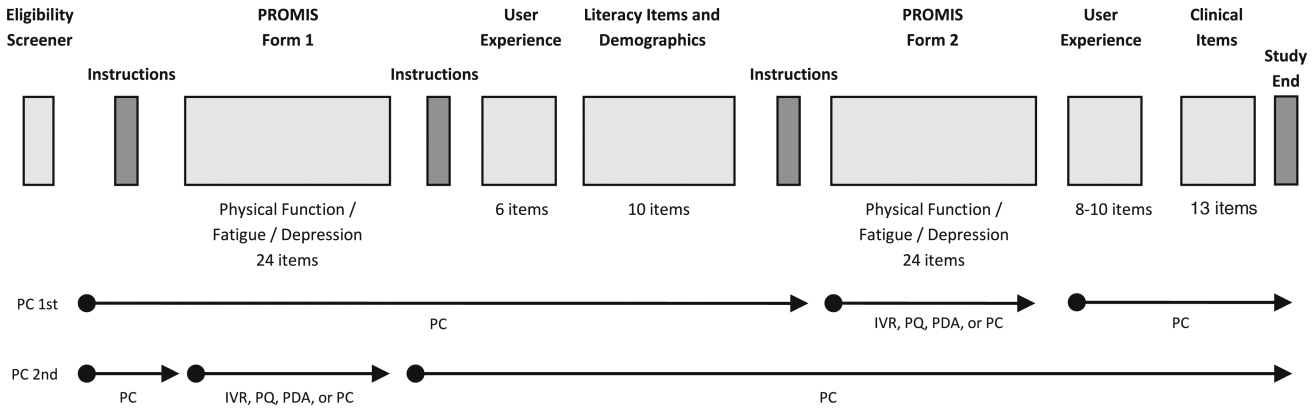

Fig. 2.
Order of Assessments. *PC* personal computer, *IVR* interactive voice response, *PQ* paper questionnaire, *PDA* personal digital assistant

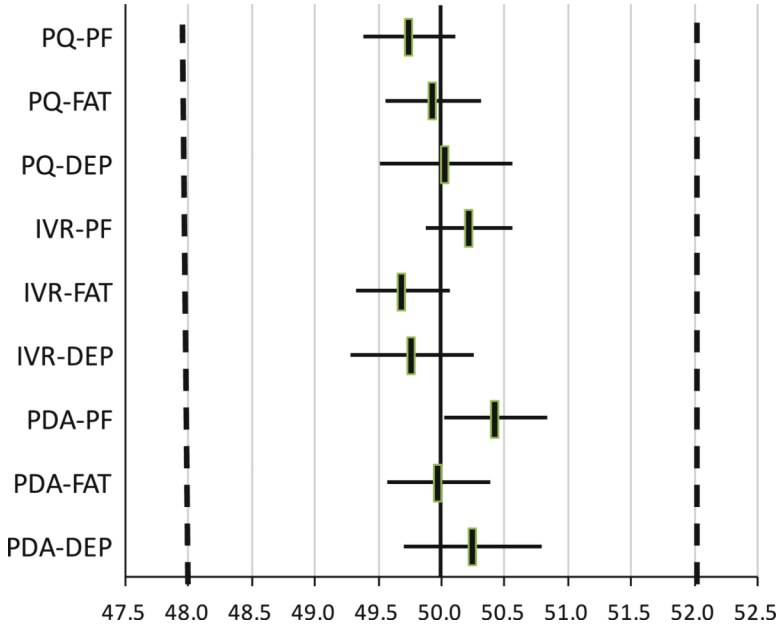

Fig. 3.

Adjusted PROMIS score estimates for different methods of administration. The figure illustrates the adjusted PROMIS scores for a person providing a response combination that would result in a score of 50.0 if provided by PC. The *horizontal axis* is rescaled to a 50–10 metric to confirm with standard PROMIS reporting. *PF* physical functioning, *FAT* fatigue, *DEP* depression, *PC* personal computer, *IVR* interactive voice response, *PQ* paper questionnaire, *PDA* personal digital assistant

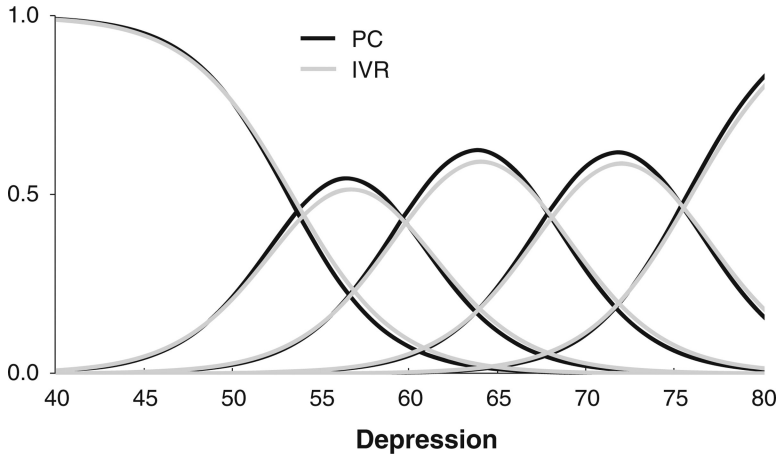

Fig. 4.
Item category response functions for the item *Felt nothing could cheer me up* using PC or IVR administration. *PC* personal computer, *IVR* interactive voice response

Table 1

Descriptive information

	Study 1			Study 2
	RA (*n* = 223)	**COPD (*n* = 248)**	**Depression (*n* = 252)**	**Rheumatology (*n* = 200)**
Male	35 %	50 %	31 %	20 %
Mean age (SD)	56 (10)	62 (10)	50 (13)	58 (15)
Age range	26–82	25–89	18–80	18–87
Hispanic ethnicity	2%	2%	5%	5%
Race				
White	89 %	93 %	90 %	95 %
African–American	6%	4%	5%	2%
American Indian/Alaskan Native	1%	1%	2%	0%
Asian	1%	0%	1%	1%
Native Hawaiian/Pacific Islander	0%	0%	0%	1%
Multiracial	4%	2%	3%	2%
Marital status				
Never married	11 %	7%	15 %	10 %
Married	61 %	51 %	48 %	64 %
Living with partner	7%	7%	15 %	2%
Separated	1%	2%	2%	4%
Divorced	16 %	20 %	17 %	9%
Widowed	3%	12 %	2%	12 %
Education				
6th through 11th grade	2%	3%	1%	1%
High school graduate/GED	19 %	21 %	12 %	24 %
Some college/Technical degree/AA	44 %	47 %	36 %	32 %
College degree (BA/BS)	18 %	18 %	30 %	20 %
Advanced degree (MA, PHD, MD)	15 %	10 %	20 %	20 %
Employment				
Full-time employed	28 %	15 %	27 %	28 %
Part-time employed	13 %	7%	8%	9%
Full-time student	2%	0%	3%	3%
Leave of absence	0%	0%	2%	1%
On disability	24 %	31 %	29 %	12 %
Retired	19 %	40 %	15 %	34 %
Unemployed	7%	3%	11 %	3%
Home maker	6%	5%	6%	10 %

Table 2

Tests for equality of loadings and thresholds using multigroup confirmatory factor analysis (across methods of administration)

		Study 1						Study 2					
		Loadings[a]			Thresholds[b]			Loadings[c]			Thresholds[d]		
Domain	Form	Chisq	DF	P	Chisq	DF	P	Chisq	DF	P	Chisq	DF	P
Physical functioning	A	10.1	10	0.43	27.8	26	0.37	5.3	4	0.26	19.2	12	0.08
	B	12.2	10	0.27	40.8	28	0.06	1.6	5	0.90	7.0	13	0.90
Fatigue	A	17.2	12	0.14	42.5	28	0.04	3.1	6	0.79	17.1	12	0.15
	B	20.2	12	0.06	37.0	26	0.07	7.1	5	0.22	20.0	13	0.10
Depression	A	5.3	11	0.92	22.6	28	0.83	7.4	5	0.19	8.9	10	0.54
	B	16.1	11	0.14	31.5	28	0.30	2.6	5	0.76	9.1	12	0.70

[a] Test that the item loadings are the same across MOAs against the alternative that each MOA (PC, IVR, and P&P) has unique loadings. Loading for first item fixed in order to identify the model

[b] Test that the item thresholds are the same across MOAs against the alternative that each MOA (PC, IVR, and P&P) has unique thresholds. First threshold for each item fixed in order to identify the model

[c] Test that the item loadings are the same across MOAs against the alternative that each MOA (PC and PDA) has unique loadings. Loading for first item fixed in order to identify the model

[d] Test that the item thresholds are the same across MOAs against the alternative that each MOA (PC and PDA) has unique thresholds. First threshold for each item fixed in order to identify the model

Table 3

IRT analysis of method of administration effect for IVR, PQ, and PDA administration compared to PC

	Physical function			Fatigue			Depression		
	Est	**SE**	**95 % CI**	**Est**	**SE**	**95 % CI**	**Est**	**SE**	**95 % CI**
Study 1									
Effect on location (mean item threshold)									
PQ	−0.026	0.018	−0.060 to 0.008	−0.006	0.019	−0.043 to 0.030	−0.014	0.022	−0.058 to 0.030
IVR	0.022	0.017	−0.011 to 0.055	−0.031	0.018	−0.067 to 0.005	0.023	0.021	−0.017 to 0.063
Effect on discrimination (precision)									
PQ	0.113	0.124	−0.130 to 0.356	0.381	0.131	0.123 to 0.638	−0.061	0.089	−0.237 to 0.114
IVR	−0.107	0.110	−0.323 to 0.109	−0.067	0.107	−0.277 to 0.142	−0.240	0.077	−0.391 to −0.089
Effect of position as second part of form									
Location	0.024	0.010		0.058	0.011		0.084	0.012	
Discrimination	0.122	0.069		0.539	0.074		0.138	0.053	
Sample means									
RA sample	−1.207	0.048		1.093	0.067		0.751	0.068	
COPD sample	−1.212	0.040		0.930	0.061		0.468	0.068	
Depression sample	−0.635	0.050		0.983	0.057		1.221	0.054	
Sample standard deviations									
RA sample	0.693	0.036		0.966	0.048		0.984	0.051	
COPD sample	0.601	0.029		0.938	0.044		1.015	0.052	
Depression sample	0.769	0.039		0.892	0.042		0.826	0.040	
Study 2									
Effect on location									
PDA	0.043	0.020	0.003 to 0.083	−0.002	0.020	−0.042 to 0.037	0.025	0.027	−0.028 to 0.078
Effect on discrimination									
PDA	0.113	0.142	−0.167 to 0.394	0.665	0.162	0.345 to 0.985	0.221	0.135	−0.045 to 0.487
Effect of position as second part of form									
Location	0.024	0.020		0.080	0.020		0.229	0.027	
Discrimination	−0.321	0.136		0.432	0.154		−0.133	0.128	
Sample mean	−0.803	0.061		0.580	0.072		0.069	0.074	
Sample standard deviation	0.787	0.046		0.939	0.052		0.932	0.058	

The IRT item parameters were standardized so that the latent score distribution in the general population was standard normal (Mean = 0, SD = 1)

PC personal computer, *IVR* interactive voice response, *PQ* paper questionnaire, *PDA* personal digital assistant