The PROMIS of Better Outcome Assessment: Responsiveness, Floor and Ceiling Effects, and Internet Administration

JAMES FRIES, MATTHIAS ROSE, and ESWAR KRISHNAN

ABSTRACT. ***Objective.*** Use of item response theory (IRT) and, subsequently, computerized adaptive testing (CAT), under the umbrella of the NIH-PROMIS initiative (National Institutes of Health – Patient-Reported Outcomes Measurement Information System), to bring strong new assets to the development of more sensitive, more widely applicable, and more efficiently administered patient-reported outcome (PRO) instruments. We present data on current progress in 3 crucial areas: floor and ceiling effects, responsiveness to change, and interactive computer-based administration over the Internet.

Methods. We examined nearly 1000 patients with rheumatoid arthritis and related diseases in a series of studies including a one-year longitudinal examination of detection of change; compared responsiveness of the Legacy SF-36 and HAQ-DI instruments with IRT-based instruments; performed a randomized head-to-head trial of 4 modes of item administration; and simulated the effect of lack of floor and ceiling items upon statistical power and sample sizes.

Results. IRT-based PROMIS instruments are more sensitive to change, resulting in the potential to reduce sample size requirements substantially by up to a factor of 4. The modes of administration tested did not differ from each other in any instance by more than one-tenth of a standard deviation. Floor and ceiling effects greatly reduce the number of available subjects, particularly at the ceiling.

Conclusion. Failure to adequately address floor and ceiling effects, which determine the range of an instrument, can result in suboptimal assessment of many patients. Improved items, improved instruments, and computer-based administration improve PRO assessment and represent a fundamental advance in clinical outcomes research. (J Rheumatol 2011;38:1759–64; doi:10.3899/jrheum.110402)

Key Indexing Terms:
ITEM RESPONSE THEORY PROMIS PHYSICAL FUNCTION DISABILITY
COMPUTERIZED ADAPTIVE TESTING SAMPLE SIZES

Successful treatment of the symptoms and functional limitations associated with the several forms of arthritis, especially rheumatoid arthritis (RA), depends upon the availability of sensitive and valid tools that can evaluate meaningful change over time and guide appropriate and timely interventions. Over the past quarter-century, assessment methods have been characterized by self-report instruments, with questionnaire items assessing some of the important aspects of arthritis-associated disability[1,2,3].

The major instruments currently in use are 25 or more years old and were created without a thorough review of alternative configurations, careful study of domain definitions, context, timeframe, response options, translatability, clarity, and importance to the patient. The advent of modern psychometrics employing item response theory (IRT) offers a unique opportunity for precise and efficient assessment of Physical Function (PF) for patients with RA[4].

The Patient-Reported Outcomes Measurement Information System (PROMIS) was inaugurated as a US National Institutes of Health (NIH) Roadmap multicenter project charged with developing improved tools for assessing patient-reported outcome (PRO) endpoints for clinical studies using IRT[5,6]. "Improvement" in these tools can take many forms, perhaps the most important of which is responsiveness to change, which is in turn a result of using items with greater precision, and selection of the best of these items for new short questionnaire forms or computerized adaptive testing (CAT). Better instruments can lead to improvement by providing increased efficiency and increasing the statistical power of studies or by keeping statistical power constant while decreasing questionnaire burden[7].

PROMIS defines PF as "the ability to perform activities of daily living (ADL) and instrumental activities of daily living" (www.nihPROMIS.org)[8,9]. This definition refers to "ability to

From the Department of Medicine, Stanford University School of Medicine, Palo Alto, California, USA; National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), National Institutes of Health, Bethesda, Maryland; and Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts, USA.

J.F. Fries, MD, Department of Medicine, Stanford University School of Medicine; M. Rose, MD, PhD, Department of Quantitative Health Sciences, University of Massachusetts; E. Krishnan, MD, Department of Medicine, Stanford University School of Medicine.

Supported by National Institutes of Health, Improved Assessment of Physical Function and Drug Safety in Health and Disease, 2 U01 AR052158-06.

Address correspondence to Dr. Fries; E-mail: jff@stanford.edu

perform" rather than "actual performance," as have the greater majority of previous instruments[9]. The term "Physical Function" is preferred to the term "disability," since it was felt desirable to develop instruments that could measure both ability and disability. One of the ways in which the term "disability" can be interpreted is as the magnitude of decrements in PF/disability compared to the ability expected of a "normal," "typical," or "average" person. Disability has been commonly measured by PRO, including instruments such as the traditional (Legacy) Health Assessment Questionnaire Disability Index (HAQ or HAQ-DI)[10,11] and the 10-item PF scale of the Medical Outcome Study Short-Form 36 (SF-36)[3].

An instrument is a collection of items, such as, "Are you able to walk a block?". PROMIS instruments are developed from large and exhaustive item banks with items that have been refined by qualitative methods for attributes such as clarity, importance, and ease of translation. Quantitative methods also are used including IRT-based calibration, which assumes unidimensionality. The most informative items in an item bank may be aggregated to develop improved instruments[12,13].

OBJECTIVE

We seek to document PROMIS advances in assessment of PF including systematic improvements in: (1) responsiveness; (2) evaluation of equivalence between paper and pencil questionnaire (PP) administration and Internet (Web browser-based) administration of the same items; and (3) floor and ceiling effects. Three articles with full descriptions of these projects and their results are in preparation. For this reason and because of space limitations, we cannot provide as detailed a discussion as we would like.

All subjects provided appropriate consent as specified by the governing institutional review board.

Responsiveness. The HAQ and PF-10, among other Legacy instruments, yield familiar, sensitive, and valid clinical PF endpoints. IRT-based assessments, however, permit aggregation of items with the greatest information content into more powerful instruments. We compared Legacy instruments with the PROMIS instruments. We performed extensive qualitative analyses of Legacy scale items that had been revised for clarity and consistency, and had common response scales and 5-option response sets[10,14]. We then compared the performance of Legacy instruments to instruments that were improved using these qualitative approaches.

We also compared the responsiveness of Legacy scales to subsets of the PROMIS PF item bank. We developed tests by selecting items with the highest information using IRT. A full introduction to the assessment of item information is beyond the scope of this report; a useful introduction is provided elsewhere[15].

Our objective was to compare responsiveness between change scores on subsets of PROMIS items and change scores on Legacy instruments to these alternative PRO measures and to test whether more informative items would reduce sample size requirements. A change score includes the true change (unobservable) and the error terms of the baseline and final scores. Item improvement is intended to decrease the standard deviation (SD) of baseline and final scores, thus permitting a closer estimate of the true change score.

Our hypotheses: (1) PROMIS instruments will efficiently measure changes in PF over time; and (2) PROMIS instruments in comparison to Legacy instruments will detect changes in PF better and will require smaller sample sizes.

Mode of administration. We systematically tested the impact of mode of administration on PROMIS items. The hypothesis is that mode of administration does not have a substantial effect on measurement characteristics of PROMIS PRO instruments.

Floor and ceiling. Most, if not all, existing PF instruments were designed to measure health status in the context of clinical settings. Such instruments do not discriminate between PF of individuals who are at the extremes of PF and are insensitive to changes at both ends of the spectrum. We hypothesized that lack of discriminative ability and precision leads to decreased study power and increased sample size requirements to detect a given effect size.

METHODS

Responsiveness. We compared 5 PF scales including 2 Legacy instruments, their item-improved derivatives, and an IRT-based Short-Form selected to maximize information. We assessed sensitivity to detect 12-month disease progression in 451 patients with RA. Metrics for change/responsiveness between baseline and 12-month measures included effect sizes, standardized response mean (SRM), and sample size requirements to detect a specified change score.

Mode of administration. Our study is designed as a randomized crossover study (Figure 1). Two non-overlapping forms (Forms A and B) with 8 unique items each from 3 of the PROMIS domains (emotional distress-depression, fatigue, PF) were developed. Respondents answered one of the forms by automated telephone interview using interactive voice response (IVR) technology, PP, or personal digital assistant (PDA) technology. The other mode was Internet-based administration. Forms were administered in random order. The 2 assessments were separated by a short interval (e.g., 5 to 10 minutes), but took place on the same day. The study was powered to detect a mean mode score difference of 1.5 on a T-score metric (SD of 10) with 85% power. Data collection through IVR and PP were performed by YouGov Polimetrix® and data for the PDA mode were collected by the Stony Brook Clinics. Respondents had one or more of the following chronic conditions: chronic obstructive pulmonary disease (COPD), depression, or RA.

Floor and ceiling. We performed a simulation study using items from the PROMIS databank where we modeled the power sample size estimates as a function of the number of items and the distribution of PF impairment in various settings. We simulated the sample size-power relationships of 4, 6, and 8 item scales in the general population and in populations where the mean PF was one SD above and below that of the mean PF in the general population. We also calculated the extent of the "floor effect" by assessing the distribution of HAQ scores in diseased and general populations.

RESULTS

Responsiveness. Four hundred fifty-one patients met American College of Rheumatology criteria for RA. The

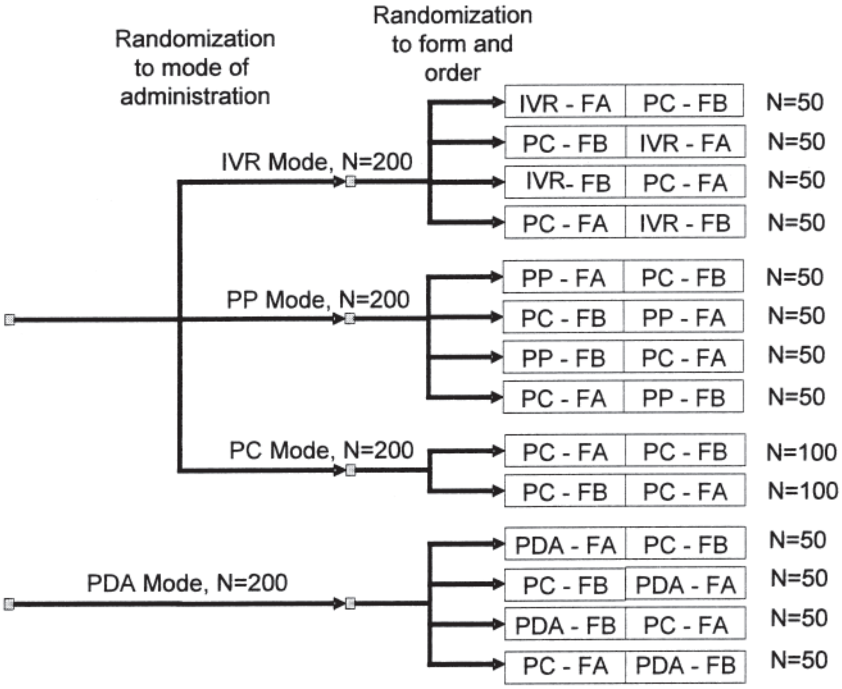

Figure 1. Four modes of administration are compared using a randomized design, which accounts for order effects and pairwise comparisons among the 4 modes. IVR: interactive voice response; PP: paper and pencil; PC: Internet connected computer; PDA; personal digital assistant. Form A (FA) and Form B (FB) are mutually exclusive 8-item questionnaires so that carry-forward effects from the previous administration are eliminated.

patients were 65 years of age with 14 years of education, 81% female and 87% Caucasian, with moderate baseline disability. 41% (N = 185) had been exposed to anti-tumor necrosis factor (TNF) treatment. All instruments were sensitive to change in PF status, with p values for changes in PF scores ranging from 0.001 to 0.05 and SRM and effect size computations mirroring these results. The most responsive were the PROMIS 20-item Short-Forms. Under study conditions, IRT-improved instruments could detect 1.2% difference with 80% power, while reference instruments could detect only a 2.4% difference ($p < 0.01$). Sample sizes required for the best IRT-improved instruments were only 24% of the worst Legacy comparator (100 vs 427).

Mode of administration. To date, we have been able to analyze the data for the PP, IVR, and Internet modes. The results presented at the OMERACT conference and in this report are preliminary first reports. We recruited 721 participants with RA, depression, and/or COPD. Two parallel forms were developed; both included 3 items measuring daily life functions, one item measuring back-neck function, 2 items lower, and 2 items upper extremity functions. First results show that they are highly consistent (Cronbach $\alpha = 0.93$) and highly correlated ($r = 0.92$).

The analysis of a generalized linear model (Table 1) demonstrated that there is no relevant mean effect for the different modes of administration. Compared to the Internet mode, the PP assessment would provide a mean score of 0.3 units higher, i.e., less than 1 point on a scale with SD of 10.

Table 1. Generalized linear model analyses examining the effect of mode of item administration. The analysis is treating the mode effect as a main effect, after the potential effect of administration order (Time 1 vs Time 2) and form (Form A vs Form B) has been taken into account. The estimates show the mean differences that can be expected on a scale with standard deviation of 10 units. All differences are less than 10% of a standard deviation.

	Estimate (Units)	Standard Error	95% CI
Internet	0		
Paper and pencil	0.30	0.33	–0.34 to 0.94
Interactive voice recognition	0.01	0.32	–0.56 to 0.71
Time 1	0		
Time 2	0.06	0.02	0.40 to 0.52
Form A	0		
Form B	–0.68	0.30	–0.26 to –0.10

Floor and ceiling. Figure 2 shows sample size power estimates for different population characteristics. The longer the instrument, the better the power for a given sample size, and the smaller the sample size for a given power requirement. However, in the population with better PF than the general population, the sample size requirements were much larger. For ceiling effects, HAQ scores of zero (HAQ ceiling) were

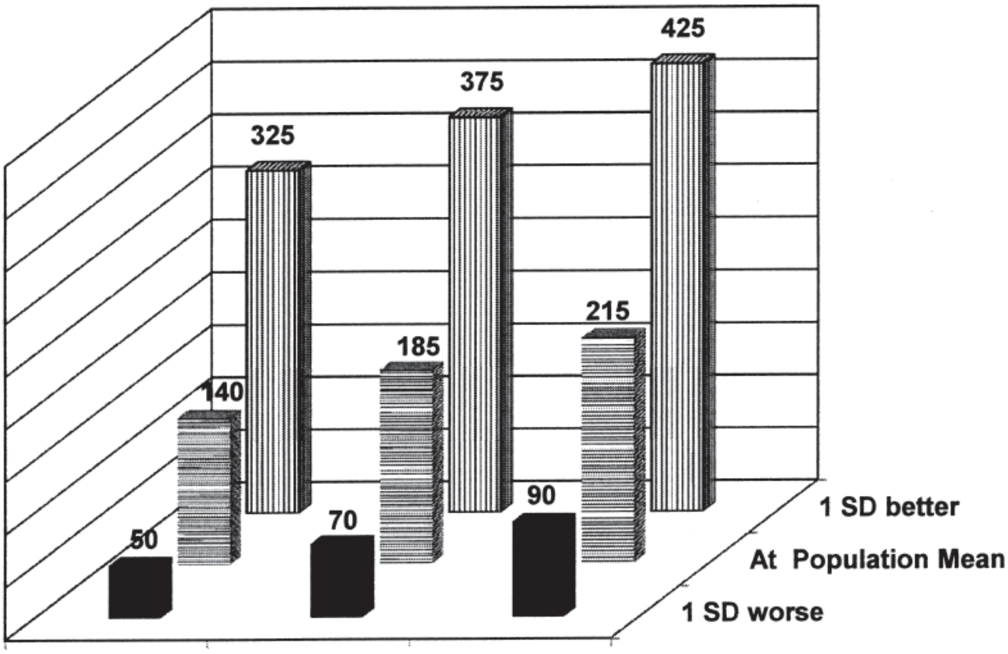

Figure 2. Sample sizes required at 80% power. Power-sample size estimates for 3 questionnaires with 4, 6, and 8 items; the fewer-item sets contain subsets of the larger item sets. Data are from simulation studies setting the effect size to 0.2 and using a population with a mean physical function score 1 SD below the population mean (as with a population of moderately affected patients with RA), at the population mean, and 1 SD above the population mean. Excellent power is achievable with 8 (or more) items with the population below the mean. The improvement effect of increasing item numbers is illustrated. However, when the population is 1 SD above the population mean, power is poor and sample size requirements large. This is the effect of studying a population where many subjects are near the ceiling where few items conveying little information are available. The increased statistical power available by adding items at the floor and ceiling to existing PROMIS item banks is inferred.

observed in about 10%–15% of RA patients and one-half or more of "normal" subjects[16].

DISCUSSION

Responsiveness. The cost of clinical research is in large part a consequence of the number of human subjects required. A large number renders recruitment a larger and longer task, requires additional centers and coordinating personnel, and puts more subjects at risk for unforeseen adverse events. Under typical conditions for studies of interventions for RA, sample sizes required may be reduced by a factor of 2 to 4 by using instruments with a lower SD of the change score relative to the change score itself. In healthier populations, we expect similar improvements in needed sample sizes by including items targeted at healthier persons who previously contributed little to power in trials because their baseline PF had previously been estimated as optimal. An initial HAQ score of zero and a final score of zero does not mean that the patient may not have improved or regressed, but only that changes occurred in the unobservable region of better than average PF.

Floor and ceiling. The sample size requirement for a given effect size and power will depend on the precision of the instrument in terms of detecting small changes across (cross-sectional studies) and within (longitudinal studies and clinical trials) groups. When the maximum sample size is predetermined owing to cost/feasibility/time considerations as in many clinical trials, the power of the study will be inversely proportional to the SD of the change score. The performance of an ideal instrument will not be influenced by the distribution of the underlying trait; it should be able to discriminate a small change regardless of the distribution of the trait in the sample.

Our simulation studies suggest that the existing instruments perform well in subpopulations with significant disability, such as those with RA, but have less discriminatory power among healthier (more able) populations. We have observed before that 68% of the general population has a HAQ score of zero, signifying no detectable disability[13]. With the use of better treatments including TNF inhibitors earlier in the disease course, functional disability in RA has been declining over time[17,18], and the available instruments are insufficient to detect treatment effects in many subjects. Items in the instrument collectively must span the full range of PF in the population under study. As in the case of RA, this range may be wide, from totally impaired to extremely robust[19].

Modes of administration. A number of studies have compared PP and computerized administration modes: PDA, Internet connected computer (PC), and interactive voice recognition

(IVR)[20]. Generally, most studies suggest psychometric equivalence between modes of administration[21,22,23]. Literature on the SF-36® Health Survey has been summarized[24,25,26]. Few studies report differences in scores[27,28]. Recently, mode effects have been discussed, in particular, for mental health assessments using the Center for Epidemiologic Studies Depression Scale[29].

The literature on mode effects between PP versus telephone administration is more limited and provides heterogeneous results. Some studies of healthcare and health status measures suggest no mode effects[30], while others report and account for them[31]. Literature on mode effects using IVR technology is sparse, too, probably due to the novelty of IVR. One large-scale study reports an IVR mode effect[32] and suggests making adjustments.

Because evaluation methods vary, studies of mode of administration are hard to compare. The studies cited above (1) used different questionnaires and/or different concepts; (2) generally did not take into account differences in the presentation of paper and electronic surveys (the paper forms can be reliably reproduced, while there may be various screen formats employed in the display of the same survey across electronic modes); (3) studied different patient populations; (4) employed different study designs (cross-sectional vs longitudinal); (5) focused on comparing only 2 administration modes (e.g., PP vs tlelphone, telephone vs computer, computer vs PP); and (6) often were underpowered to detect small but clinically meaningful differences. Thus, the current project was designed to examine 4 modes of administration within one study and to minimize these problems. The results are reassuring.

CONCLUSIONS

Our report discusses 3 important advances in assessment of PF achieved by the PROMIS network. Outcome scales developed from IRT-improved items result in greater responsiveness and study efficiency, improving the precision of clinical studies and reducing sample size requirements. Potentially, study enrollment periods will shorten, number of centers and investigators will be reduced, and costs of clinical research may be substantially decreased.

Reduction in floor and ceiling effects improves power and allows the use of the same metric to follow severely impaired individuals and those in robust health.

The current mode of administration study is one of the largest of its kind, and results are reassuring as we move into an era where some but not all data for a study will be acquired electronically. Our preliminary results found minimal mode of administration effect on the mean score estimation for PF. This represents a major advance, as it is likely to enable investigators to proceed without requiring major adjustments for mode of administration.

ACKNOWLEDGMENT

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 [Evanston Northwestern Healthcare, Principal Investigator (PI): David Cella, PhD, U01AR52177] and 6 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 have included Deborah Ader, PhD, Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Laura Lee Johnson, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD, PhD. This report was reviewed by the PROMIS Publications Subcommittee before external peer review.

REFERENCES

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
2. Ware JE Jr, Kosinski M, Keller SD. A 12-item Short Form Health Survey. Med Care 1996;34:220-33.
3. Ware JE, Jr, Kosinski M. SF-36 physical and mental health summary scales: a manual for users of version I. 2nd ed. Lincoln, RI: QualityMetric, Inc.; 2001.
4. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61:17-33.
5. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45 Suppl 1:S22-31.
6. Hambleton R, Swaminathan H, Rogers J. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
7. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care 2007;45 Suppl 1:S32-38.
8. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45:S3-S11.
9. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23 Suppl 39:S53-57.
10. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
11. Bruce B, Fries J. The Stanford Health Assessment Questionnaire (HAQ): a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167-78.
12. Patient Reported Outcomes Measurement Information System. [Internet home page. Accessed April 5, 2011.] Available from: http://www.nihpromis.org.
13. Krishnan E, Sokka T, Hakkinen A, Hubert H, Hannonen P. Normative values for the Health Assessment Questionnaire disability index. Arthritis Rheum 2004;50:953-60.
14. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45 Suppl 1:S12-21.
15. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38 Suppl:1128-42.

16. Krishnan E, Tugwell P, Fries JF. Percentile benchmarks in patients with rheumatoid arthritis: Health Assessment Questionnaire as a quality indicator (QI). Arthritis Res Ther 2004;6:R505-513.
17. Krishnan E, Fries JF. Reduction in long-term functional disability in rheumatoid arthritis from 1977 to 1998: a longitudinal study of 3035 patients. Am J Med 2003;115:371-6.
18. Krishnan E, Hakkinen A, Sokka T, Hannonen P. Impact of age and comorbidities on the criteria for remission and response in rheumatoid arthritis. Ann Rheum Dis 2005;64:1350-2.
19. Bruce B, Fries JF. The Arthritis, Rheumatism and Aging Medical Information System (ARAMIS): still young at 30 years. Clin Exp Rheumatol 2005;23 Suppl 39:S163-167.
20. Dillman D. Mail and Internet surveys: The tailored design method. New York: John Wiley & Sons; 2000.
21. Bettinville A, Rabenberg H, Hansgen KD. An enquiry into the WILDE-Intelligence-Test (WIT): comparability of application of the paper-and-pencil version vs. the computer based application — an analysis based on data of the Leipzig Vocational Retraining Center [German]. Rehabilitation (Stuttg) 2005;44:237-43.
22. Folk LC, March JZ, Hurst RD. A comparison of linear, fixed-form computer-based testing versus traditional paper-and-pencil-format testing in veterinary medical education. J Vet Med Educ 2006;33:455-64.
23. Norman GJ, Sallis JF, Gaskins R. Comparability and reliability of paper- and computer-based measures of psychosocial constructs for adolescent physical activity and sedentary behaviors. Res Q Exerc Sport 2005;76:315-23.
24. Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. Qual Life Res 2001;10:15-22.
25. Burke J, Burke KC, Baker JH, Hillis A. Test-retest reliability in psychiatric patients of the SF-36 Health Survey. Int J Methods Psychiatr Res 1995;5:189-94.
26. Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. Qual Life Res 2002;11:19-26.
27. Beebe T, Harrison P, McRae J, Evans J. The effects of data collection mode and disclosure on adolescent reporting on health behavior. Soc Sci Comput Rev 2006;24:476-88.
28. DeAngelis S. Equivalency of computer-based and paper-and-pencil testing. J Allied Health 2000;29:161-4.
29. Swartz RJ, de Moor C, Cook KF, Fouladi RT, Basen-Engquist K, Eng C, et al. Mode effects in the Center for Epidemiologic Studies Depression (CES-D) Scale: personal digital assistant vs. paper and pencil administration. Qual Life Res 2007;16:803-13.
30. Duncan P, Reker D, Kwon S, Lai SM, Studenski S, Perera S, et al. Measuring stroke impact with the Stroke Impact Scale: telephone versus mail administration in veterans with stroke. Med Care 2005;43:507-15.
31. 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. Aust NZ J Pub Health 2005;29:149-54.
32. Rodriguez HP, von Glahn T, Rogers WH, Chang H, Fanjiang G, Safran DG. Evaluating patients' experiences with individual physicians: a randomized trial of mail, internet, and interactive voice response telephone administration of surveys. Med Care 2006;44:167-74.