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

Stephen Joel Coons PhD,[1] Chad J. Gwaltney PhD,[2] Ron D. Hays PhD,[3] J. Jason Lundy MS,[4] Jeff A. Sloan PhD,[5] Dennis A. Revicki PhD,[6] William R. Lenderking PhD,[7] David Cella PhD,[8] and Ethan Basch MD, MSc,[9] on behalf of the ISPOR ePRO Task Force

1. Center for Health Outcomes and PharmacoEconomic Research, College of Pharmacy, University of Arizona, Tucson, AZ, USA
2. Brown University, Providence, RI, USA and PRO Consulting, Pittsburgh, PA, USA
3. Division of General Internal Medicine and Health Services Research, Department of Medicine, UCLA School of Medicine, Los Angeles, CA, USA and RAND, Santa Monica, CA, USA
4. Department of Pharmaceutical Sciences, College of Pharmacy, University of Arizona, Tucson, AZ, USA
5. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
6. Center for Health Outcomes Research, United BioSource Corporation, Bethesda, MD, USA
7. Center for Health Outcomes Research, United BioSource Corporation, Lexington, MA, USA
8. Center on Outcomes, Research and Education, Evanston Northwestern Healthcare and Northwestern University Feinberg School of Medicine, Evanston, IL, USA
9. Health Outcomes Research Group, Departments of Biostatistics and Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Corresponding Author: Stephen Joel Coons, College of Pharmacy, University of Arizona, PO Box 210202, Tucson, AZ, 85721-0202 USA

Tel: 520-626-3566

Fax: 520-626-4063

Email: coons@pharmacy.arizona.edu

Source of financial support: Although there were sponsoring firms that underwrote the initial meeting that led to this report, the publication of these recommendations was not contingent on the sponsors' approval.

Keywords: patient-reported outcomes, effectiveness, evaluation studies, health-related-quality of life

Running Head: Recommendations of the ISPOR ePRO Task Force

ABSTRACT

Background: Patient-reported outcomes (PROs) are the consequences of disease and/or its treatment as reported by the patient. The importance of PRO measures in clinical trials for new drugs, biologic agents, and devices was underscored by the release of the US Food and Drug Administration's draft guidance for industry titled "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims." The intent of the guidance was to describe how the FDA will evaluate the appropriateness and adequacy of PRO measures used as effectiveness endpoints in clinical trials. In response to the expressed need of ISPOR members for further clarification of several aspects of the draft guidance, ISPOR's Health Science Policy Council created three task forces, one of which was charged with addressing the implications of the draft guidance for the collection of PRO data using electronic data capture modes of administration (ePRO). The objective of this report is to present recommendations from ISPOR's ePRO Good Research Practices Task Force regarding the evidence necessary to support the comparability, or measurement equivalence, of ePROs to the paper-based PRO measures from which they were adapted.

Methods: The Task Force was composed of the leadership team of ISPOR's ePRO Working Group and members of another group (i.e., *ePRO Consensus Development Working Group)* that had already begun to develop recommendations regarding ePRO good research practices. The resulting Task Force membership reflected a broad array of backgrounds, perspectives, and expertise that enriched the development of this report. The prior work became the starting point for the Task Force report. A subset of the Task Force members became the writing team that prepared subsequent iterations of the report that were distributed to the full Task Force for review and feedback. In addition, review beyond the Task Force was sought and obtained. Along with a presentation and discussion period at an ISPOR meeting, a draft version of the full report was distributed to roughly 220 members of a reviewer group. The reviewer group comprised individuals who had responded to an e-mailed invitation to the full membership of ISPOR. This Task Force report reflects the extensive internal and external input received during the 16-month good research practices development process.

Results/Recommendations: An ePRO questionnaire that has been adapted from a paper-based questionnaire ought to produce data that are equivalent or superior (e.g., higher reliability) to the data produced from the original paper version. Measurement equivalence is a function of the comparability of the psychometric properties of the data obtained via the original and adapted administration mode. This comparability is driven by the amount of modification to the content and format of the original paper PRO questionnaire required during the migration process. The magnitude of a particular modification is defined with reference to its potential effect on the content, meaning, or interpretation of the measure's items and/or scales. Based on the magnitude of the modification, evidence for measurement equivalence can be generated through combinations of the following: cognitive debriefing/testing, usability testing, equivalence testing, or, if substantial modifications have been made, full psychometric testing. As long as only minor modifications were made to the measure during the migration process, a substantial body of existing evidence suggests that the psychometric properties of the original measure will still hold for the ePRO version. Hence, an evaluation limited to cognitive debriefing and usability testing only may be sufficient. However, where more substantive changes in the migration process has occurred, confirming that the adaptation to the ePRO format did not introduce significant response bias and that the two modes of administration produce essentially equivalent results is necessary. Recommendations regarding the study designs and statistical approaches for assessing measurement equivalence are provided.

Conclusions: The electronic administration of PRO measures offers many advantages over paper administration. We provide a general framework for decisions regarding the level of evidence needed to support modifications that are made to PRO measures when they are migrated from paper to ePRO devices. The key issues include (1) the determination of the extent of modification required to administer the PRO on the ePRO device and (2) the selection and implementation of an effective strategy for testing the measurement equivalence of the two modes of administration. We hope that these good research practice recommendations provide a path forward for researchers interested in migrating PRO measures to electronic data collection platforms.

INTRODUCTION

Overview

Patient-reported outcomes (PROs) are the consequences of disease and/or its treatment as reported by the patient, including perceptions of health, well-being, symptom experience, functioning, and treatment satisfaction. PROs are increasingly being used to complement safety data, survival rates, and other traditional indicators of clinical efficacy in therapeutic intervention trials [1]. They enrich the evaluation of treatment effectiveness by providing the patient perspective. In some cases, such as pain assessment or fatigue, a PRO may be the only viable endpoint since there are no observable or measurable physical or physiological markers of disease or treatment activity [2-4]. In other cases, where PROs are not the only available endpoint, they may still be among the most important.

A number of reports and consensus papers addressing the use of PROs in clinical research and labeling claims have been published during the past several years [5-11]. Regulatory agencies are being asked increasingly to review and approve protocols that include PRO measures [12, 13]. As of 1994, the majority of Phase II-IV clinical trials collected some type of PRO data [14]. Willke et al. [12] reviewed the effectiveness endpoints reported in FDA-approved product labeling for new molecular entities approved from 1997 through 2002 and found that PRO endpoints were included in 30% (64) of the 215 product labels examined. For 23 products, PROs were the only endpoints reported.

Concurrent with the increased use and significance of PROs in clinical trials has been the steady growth in electronic data capture (EDC) in clinical trials. There have been missteps along the way, most notably the perceived lack of adequate technical support for clinical investigators [15-17]. Adaptation of case report forms to electronic formats, including electronic modes of PRO administration (ePROs), must ensure the data collected via the different methods are equivalent or account for any identified differences.

The importance of PRO measures in clinical trials for new drugs, biologic agents, and devices was underscored by the release of the US Food and Drug Administration's draft guidance for industry, "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims" [18]. The intent of the guidance was to describe how the FDA will evaluate the appropriateness and adequacy of PRO measures used as effectiveness endpoints in clinical trials. The FDA guidance was created to make the process of developing and reviewing PRO measures more efficient and transparent for both the FDA and clinical trial sponsors by outlining basic evaluation standards. A series of articles commenting on various aspects of PRO development, selection, testing, analysis, and interpretation contained in the FDA guidance document were recently published [19-25]. Nevertheless, this process continues to evolve and remains challenging due, in part, to the myriad of possible PRO measures, the need for various language and cultural adaptations, and the multiple existing and emerging modes of administration. Furthermore, the draft guidance raised specific issues associated with ensuring the comparability of electronic and paper-based PRO measures [18].

Many PRO measures were originally developed for paper-and-pencil administration, but may be able to be adapted to ePRO formats. EDC adaptation of existing PRO measures may lead to less administrative burden, high patient acceptance, avoidance of secondary data entry errors, easier implementation of skip patterns, and more accurate and complete data [26-31]. The FDA has indicated openness to considering the advances promised by the use of ePRO measures in clinical trials [25]. However, the ePRO measure will be subject to the same scrutiny as would a paper-based measure. Empirical evidence will be required to demonstrate that the measurement properties of the ePRO application are comparable if not superior to the original PRO format. Needless to say, it would be unwise to consider moving a paper-based PRO measure to an electronic format for use in a clinical trial if the original measure does not meet the standards of the FDA guidance. In addition, migration of existing PRO measures to ePRO devices should be planned, conducted, and evaluated with permission and in cooperation with the measure's developer whenever possible.

The purpose of this manuscript is to present recommendations for the evidence necessary to support the comparability or measurement equivalence of ePROs to the paper-based PRO measure from which they were adapted. Although a brief review is provided, this manuscript is not intended to comprehensively compare and contrast the various modes of ePRO administration. Furthermore, these recommendations are predicated on the assumption that for use in clinical trials, the ePRO data collection and storage infrastructure complies with regulatory requirements for sponsor and investigator record keeping, maintenance, and access. We will not discuss this issue in detail. Record keeping requirements (addressed in 21 CFR 312.50, 312.58, 312.52, 312.68, 812.140, and 812.145) include the preparation and maintenance of case histories, record retention, and provision for the FDA to access, copy, and verify records [32]. In addition, collection of ePRO data must be compliant with the Guidance for Industry: E6 Good Clinical Practice (Section 5.5.3) [33], Guidance for Industry: Computerized Systems Used in Clinical Investigations [34], and 21 CFR Part 11 [35-37]. Hence, records must be maintained or submitted in accordance with the underlying requirements set forth in the Federal Food, Drug, and Cosmetic Act, the Public Health Service Act, and applicable FDA regulations (other than part 11).

Task Force Process

After release of the draft PRO guidance in February 2006, the FDA solicited comments and suggestions to inform the finalization of the guidance. ISPOR membership provided comments to the FDA identifying the need for clarity on several specific issues, including the FDA's expectations regarding the use of existing PRO instruments and their modifications, translating and/or adapting PRO measures from one language/culture to another, and changing the mode of administration of PRO measures, specifically to electronic data capture (ePRO).

Based on January 2007 recommendations from ISPOR's Health Science Policy Council, the ISPOR Board of Directors in March 2007 approved the formation of three PRO Task Forces to address the above issues. The task force that came to be called the *ePRO Task Force* initially was composed of the leadership team of ISPOR's ePRO Working Group, which was chaired by Stephen Joel Coons. Another

group already in existence and also chaired by Prof. Coons, the *ePRO Consensus Development Working Group,* was merged into the ePRO Task Force soon afterward. The resulting Task Force membership reflected a broad array of backgrounds, perspectives, and expertise that enriched this good research practices development process.

The work that had been begun by the *ePRO Consensus Development Working Group* became the starting point for the ePRO Task Force report. A subset of the Task Force members became the writing team that prepared subsequent iterations of the report. Monthly Task Force teleconferences were held to review the progress and provide feedback to the writing team. In addition, review beyond the Task Force members was sought and obtained. An outline of the initial recommendations and future direction of the ePRO Task Force report was presented as part of a PRO Forum at the May 2007 ISPOR 12th Annual International Meeting. Questions and feedback from the PRO Forum participants informed and further defined the content of the Task Force report. Once a draft version of the full report was completed, it was distributed in November 2007 to a roughly 220 member reviewer group. The reviewer group comprised individuals who had responded affirmatively to an e-mailed invitation to the full membership of ISPOR to join the ePRO Working Group. A considerable amount of substantive feedback was received from the reviewer group. Based on both the internal and external input, innumerable iterations of the report were distributed to the Task Force members over a 16-month period. This final report reflects the culmination of that extensive process.

Types of ePRO Data Collection Devices/Systems

There are two main categories of ePRO administration platforms: voice/auditory devices and screen text devices. Voice/auditory devices are primarily telephone based and are commonly referred to as interactive voice response (IVR). Screen text devices provide the respondent with a computerized version of the measure's items and responses in a visual text format. Screen text devices include desktop and laptop computers, which may include a touch screen; tablet or touch-screen notebook computers; handheld/palm computers; and web-based systems. Computer touch screen systems differ from

traditional computer keyboard and mouse systems by having a touch-sensitive monitor screen that allows the patient to respond to questions by the touch of a finger or stylus. Touch-screen applications may be used with or without a keyboard or mouse; however, the stand-alone desktop systems are limited in mobility.

Touch-screen tablet or laptop systems are usually full function computers that have few practical limits on the number of ePRO questions, graphical displays (e.g., body diagrams, visual analog scales), computational complexity, data storage, or data transfer options. Because tablet or laptop computers offer more screen space than other screen-based options, the question and response text can be presented in larger font and displayed on the same screen in practically all languages.

With handheld computer systems/devices, data are entered via the touch sensitive screen using a special pen/stylus. Handheld computers offer the advantage of being lightweight and the most portable of the screen text devices, but the drawback can be limited screen space. This may require the respondent to scroll to view the entire question and response set. It also limits the use of larger, easier to read fonts. However, portability of the handheld computer gives it the advantage of being potentially more useful for real-time assessment of patient experience such as eDiaries [3, 38].

Web-based systems offer the advantage of capturing the PRO data in the data file as the patient is responding to the questionnaire. The data does not need to be transferred to a central server, which is the process required by the other screen-based systems and has been known to present challenges to study subjects and study site staff. In addition, web-based systems can accommodate protocol and other changes during a study much more easily and at much less cost than the other screen-based systems since the changes only need to be made to the software residing on the central server. Other screen-based systems require software changes to be uploaded to each device, which can create significant logistical and technical challenges. Web-based ePRO systems require access to a computer with internet service or a device enabled with access to a wireless network. Depending on the study protocol, web-based systems potentially offer the respondent the convenience of completing the questionnaire in their

home. The touch-screen and mobility advantages may be lost unless the computer has touch-screen and internet capabilities; however, the latter is becoming increasingly available in most countries.

Audiovisual computer-assisted self interviewing (A-CASI) is an EDC hybrid device that combines screen text and voice/auditory functionality into one platform. Respondents are presented with a questionnaire on a computer monitor, with or without a touch screen, accompanied by an audible reading of the questions and responses. Hybrid devices can offer the respondent the choice of disabling the audio reading of the questionnaire and responding to the visual presentation only, or vice versa, which can be useful for assessing special populations (low literacy or visually impaired) [39].

Voice/auditory devices provide the respondent with an audio version of the questions and response choices. Specifically, IVR systems are automated telephone-based systems that interact with callers using a pre-recorded voice question and response system. Some of the advantages of IVR are that no additional hardware is required for the respondent other than a telephone, little if any respondent training is necessary, data are stored directly to the central database, and IVR systems can record voice responses. The use of the recorded voice prompts has been shown to reduce the literacy skill requirements of study participants [40, 41]. IVR systems accept a combination of voice input and touch-tone keypad selection to facilitate the completion of questionnaires. IVR systems allow for respondents to call in or for the system to call respondents; however, it is recommended that researchers provide written complementary materials for questions and response options at the start of the study, particularly for lengthy questionnaires. The auditory presentation of IVR systems departs from the visual medium in which most PRO measures were developed, but it is very similar to telephone interview-administered modes of data collection. Few studies have directly compared IVR and paper-based versions of PRO measures. Further research is needed to assess whether and under what conditions (e.g., length of assessment or item, number of options, respondent cognitive capacity) transfer from PRO written modalities to IVR yields equivalent data.

The choice among the different ePRO platforms should consider the type of PRO measure being adapted, the target population, the complexity of data capture requirements or scoring calculations, and the timeframe required for patient reporting (e.g., immediate vs. recall). For all the above ePRO applications where the data are not stored immediately in a central database, once the data are collected, they should be transferred as soon as possible via internet, intranet, or server-based system to a centralized storage and processing facility.

Comparisons of Electronic and Paper Modes of PRO Administration in the Literature

A number of studies have directly compared data obtained with electronic and paper modes of PRO administration. Gwaltney et al. [42] performed a meta-analysis that included 46 studies and over 275 PRO measures to examine the relationship between paper PROs and computer screen-based ePROs. The average mean difference between the modes was very small (0.2% of the scale range or 0.02 points on a 10-point scale) and the average correlation between the paper and ePRO measures indicated redundancy (0.90). The cross-mode correlation was often similar to the test-retest reliability of the paper measure, which indicates equivalence of the measures. In such circumstances, administering the paper measure and then the ePRO is essentially the same as administering the paper PRO measure twice.

Several different computer screen-based devices were used for administering PROs in the reviewed literature; including computer touch screen, handheld computers, web-based platforms, as well as traditional computer monitor, keyboard, and mouse. There was little evidence that the size of the computer screen, respondent age, or amount of computer experience meaningfully influenced the equivalence of the ePRO [42].

Studies in which IVR systems have been used to collect patient-reported data have provided support for the reliability and feasibility of the data collection mode [43,44]. Other studies have compared traditionally clinician-administered/completed clinical rating forms with IVR-administered patient-completed versions [45-47]. Mundt et al. [46] compared an IVR-administered version of the Montgomery-Asberg Depression

Rating Scale to clinician administration in a small sample (n=60) of patients. The findings provided initial evidence of the equivalence of the administration modes based on the lack of statistically significant or clinically meaningful total score mean difference. Rush et al. [47] compared three modes of administration (clinician rating, paper-based self report, and IVR) of the Quick Inventory of Depressive Symptomatology (QIDS). They found that in non-psychotic patients with major depressive disorder, the IVR and self-report versions of the QIDS performed as well as the clinician-rated version in terms of internal consistency reliability and all three versions provided comparable mean total scores. Agreement between the three self-report versions of the QIDS regarding pre-defined response to treatment (Yes/No) was acceptable based on kappa coefficients (0.72 to 0.74).

There are few publications comparing PRO measures originally developed for paper-and-pencil self-administration with an IVR-adapted administration. Alemi et al. [48] compared IVR administration of a follow-up questionnaire for recovering drug addicts with a mailed, self-administered version. They found no significant differences between the responses collected via the two modes but that the IVR mode had a higher response rate. Agel et al. [49] compared the responses obtained on an IVR-administered version of the Short Musculoskeletal Function Assessment (SMFA) questionnaire to those obtained with a paper self-administered version. Based on the crossover design, there were no significant differences between the means of the responses on the versions of the questionnaire. Dunn et al. [50] tested correspondence between the original paper version and an IVR version of the Changes in Sexual Functioning Questionnaire (CSFQ). The authors reported high Pearson product-moment correlations between the versions for both the CSFQ total score and the individual subscales scores.

The published literature addresses other types of comparisons between ePROs and paper PROs, including time to completion, satisfaction/ease of use, and missing data [51]. Although time to completion was often used as a comparison measure between the paper-based and the electronic adaptation of the PRO questionnaires, the findings are equivocal and the implications are unclear. In some studies, respondents were faster on the electronic version than the paper version [29, 52, 53] and in other studies respondents were faster on the paper version [54-56]. Results have indicated that less computer

experience, greater age, poorer physical condition, and lower education were associated with greater time needed to complete the ePRO [29, 56, 57]. Other than level of computer experience, these influences are not unique to ePROs. Some studies found that although patients took longer to complete the ePRO form, they reported that they thought completion took less time for ePROs compared with the paper version [58].

Other outcomes used to evaluate ePROs, such as satisfaction and ease of use were usually measured through the administration of follow-up questions after PRO completion. Typically, respondents were asked about the ease of using the electronic format, the adequacy of the instructions, ability to read the screen, and the acceptability of the time taken to complete the questionnaires. Respondents generally reported that they preferred the ePRO over the paper PRO [29, 52-56, 59].

Quantity of missing data was another important comparison between paper PRO and ePRO modes of administration [29, 53, 60, 61]. ePROs typically produce less missing data than paper-based measures, but the amount of usable data from each format should be compared. One potential problem regarding missing data with handhelds is that the devices themselves can be lost. In order to allow respondents the ability to opt out of answering individual items, ePRO instruments should have “choose not to respond” or “skip question” response options or some other means of moving forward without answering. In addition, the ability to review and change prior responses are a characteristic of paper-based forms that can be implemented with all ePRO devices.

EVIDENCE NEEDED TO SUPPORT MEASUREMENT EQUIVALENCE

Definition of Measurement Equivalence

An ePRO measure that has been adapted from a paper-based measure ought to produce data that are equivalent or superior (e.g., higher reliability) to the data produced from the original paper version. Measurement equivalence is a function of the comparability of the psychometric properties of the data

obtained via the original and adapted administration mode. This comparability is driven by the amount of modification to the content and format of the original paper PRO measure required during the adaptation process. Hence, the amount of change that occurs during migration to the electronic platform/device will dictate the amount of evidence necessary to demonstrate that the change did not introduce response bias and/or negatively affect the measure's psychometric properties. As noted in the FDA draft guidance [18, lines 582-583], "The extent of additional validation recommended depends on the type of modification made."

In Table 1 we provide a framework for assessing the magnitude of a particular change and match the degree of change with a recommended strategy for assessing measurement equivalence. The magnitude of a particular change is defined with reference to its potential effect on the content, meaning, or interpretation of the measure's items and/or scales. Note that the FDA draft PRO guidance does not make the distinction between minor, moderate, or substantial modifications. The draft guidance indicates that additional validation is required when "an instrument is altered in item content or format" [18, line 619]. Our goal is to be more explicit about how much additional validation is needed given the modifications to the paper version to convert it to an ePRO mode of administration. Full psychometric validation for every modification is impractical and, furthermore, not necessary based on current evidence.

- A minor modification is not expected to change the content or meaning of the items and response scales. Simply placing a scale from a paper-and-pencil format into a screen text format without significantly reducing font size, altering item content, recall period, or response options qualifies as a minor modification. This includes an appreciation of the fact that a one item per screen electronic format differs from the many items per page paper format. The large literature on migrating from paper to screen-based platforms suggests that these common modifications will not have a substantive effect on the performance of the measure [42]. However, it is still important to provide some evidence for the equality of the ePRO measure to other modes of data collection. In these cases, small-scale (5-10 patients) cognitive interviewing [63] and usability

testing (see below) can establish that participants are responding to the assessment items in the intended manner and that the ePRO software works properly when used by the target population.

- A moderate level of modification may change the meaning of the assessment items, but this change might be subtle. Examples of changes to items that could fall in this category include splitting a single item into multiple screens, significantly reducing the font size, and requiring the patient to use a scroll bar to view all item text or responses. Another example might include changing the order of item presentation. When these types of modifications are made to a PRO, it is advisable to formally establish the equivalence of the electronic measure. Designs that can be used to establish equivalence are discussed below. We include migrating from paper PROs to IVRS in this category, as (a) it remains unclear whether there are reasons to be concerned about the changes involved in moving from paper to IVRS (e.g., visual to aural presentation); and (b) the available literature supporting the equivalence between IVRS and paper is emerging and still not conclusive. In addition to assessing measurement equivalence, usability testing should be conducted in the target population.

- Substantial modifications almost certainly will change the content or meaning of the assessment. Examples of changes that could fall in this category include removing items to decrease the amount of time it takes to complete an assessment or making dramatic changes to item text, such as removing references to a recall period or scale anchors, in order to fit an item on a screen. In this case, equivalence of the assessments may be irrelevant and the modified measure should be treated as a new measure. Estimating the comparability of the old and new versions of the measure may be still be valuable for some purposes such as bridging scores [64]. Little or none of the data on the reliability and validity of the original measure will be informative in judging the quality of the modified measure. Therefore, studies designed to assess the psychometric characteristics of the new measure are required along with large scale usability testing in the target population.

Levels of Evidence

Cognitive Debriefing

Cognitive debriefing (a.k.a., cognitive interviewing or cognitive testing) is becoming increasingly important in the development and testing of many types of questionnaires [63]. Cognitive interviewing techniques are used to explore the ways in which members of the target population understand, mentally process, and respond to the items on a questionnaire [65]. Although most often associated with questionnaire development, cognitive debriefing is directly applicable to the pre-testing of alternative modes of administration for existing measures. Cognitive debriefing consists of the use of both *verbal probing* by the interviewer (e.g., "What does the response 'some of the time' mean to you?") and *think aloud* in which the interviewer asks the respondent to verbalize whatever comes to mind as he or she answers the question [66].

In this context, cognitive debriefing would be used to assess whether the ePRO application changes the way respondents interpret the questions, decide on an answer, and respond. In addition, it can help to determine whether the instructions were clear or if anything was confusing. The cognitive debriefing should be conducted with 5 to 10 patients [67], but more may be necessary to adequately reflect the target study population. It is important to fully document the process along with the qualitative findings and any resulting changes.

Usability Testing

Usability testing examines whether respondents from the target population are able to use the software and the device appropriately. This process includes formal documentation of respondents' ability to navigate the electronic platform, follow instructions, and answer questions. The overall goal is to demonstrate that respondents can complete the computerized assessment as intended. The scale of the usability testing process should be based on the complexity of the physical and cognitive tasks required

for the specific ePRO application. The characteristics of the PRO measure (e.g., number and format of items, types of response scales, number of response options) in combination with the characteristics of the ePRO device/platform (e.g., visual vs. aural, touch-tone vs. touch-screen, stylus vs. finger) drives the number of subjects needed. Usability testing may require a small number of subjects (5 to 10) for an ePRO device that is simple to use or a larger sample (20 or more) for one that is more physically and/or cognitively complex.

Usability testing as described above is not the same as another process called *user acceptance testing* (UAT). The purpose of UAT is to determine whether the software complies with the written system specification or user requirements document. It is not intended solely to determine if respondents like or can use the system. UAT is one aspect of an extensive system/software validation process that is far beyond the scope of this manuscript.

Equivalence Testing

Equivalence testing is designed to evaluate the comparability between PRO scores from an electronic mode of administration and paper-and-pencil administration. The intent is to ensure that PRO scores from the ePRO do not vary significantly from those scores from a paper questionnaire (except for measurement error). There are several study designs and statistical methods that can be used to assess the comparability of measurement obtained on two (or more) different occasions. First, we discuss study designs followed by statistical methods for equivalence testing.

Study Designs for Testing Measurement Equivalence

When it is necessary to test the measurement equivalence of an ePRO adaptation, as in the second level of modification listed in Table 1, there are two recommended study designs: 1) the randomized parallel groups design; and 2) the randomized crossover design. The study sample should be representative of

the intended patient group in which the ePRO will be used, particularly in regard to age, gender, race/ethnicity, education, and disease severity.

Randomized parallel groups design

In the randomized parallel groups design, patients are randomly assigned to one of two study arms. In this design, patients in one study arm would complete the original paper version of the PRO measure and patients in the other arm would complete the ePRO measure. Comparisons of mean score differences can then be made between groups. The random assignment of an adequate number of patients to each of the two study arms is designed to yield equivalence of the characteristics of the two groups. More elaborate studies based on a parallel groups design could involve more than two comparison groups (e.g., paper PRO vs. tablet ePRO vs. IVRS ePRO) or could incorporate a repeat administration (within mode) after a two-day to two-week interval. The latter would provide directly comparable test-retest reliability for the paper PRO and ePRO measures.

There are two possible approaches for testing of equivalence in a parallel groups design: 1) set a mean difference “d” that would be the minimum effect size that is indicative of a lack of equivalence and calculate a sample size to detect the difference “d” with sufficient power; or 2) set a level of difference “d” that is the maximum that would be tolerated for equivalence to be accepted, express the hypothesis testing in terms of ruling out differences smaller than “d,” and calculate a sample size that would be required to rule out such a difference being present. The first approach would be erroneous [68, 69]; it is inherent in the logic of statistical inference that one draws a definitive conclusion when a hypothesis is rejected, not when it fails to be rejected. Blackwelder [70] provides an accessible summary of carrying out equivalence testing procedures and Atherton and Sloan [71] provide convenient design algorithm macros. Compared to classical hypothesis testing, the equivalence approach will inflate the sample size required to demonstrate equivalence by as much as one third greater [69]. To rule out differences between a paper-based PRO and ePRO assessment of 0.3 standard deviations (a small effect size), a two-sample t-

test based on 234 patients per group would provide 80% power with a two-tailed alternative and a 5% Type I error rate.

Randomized crossover design

The use of the crossover design in ePRO equivalence studies would involve the random assignment of respondents to complete either a paper PRO or ePRO measure for the first administration and then the other mode for the second administration. Adequate time should be allowed between administrations to minimize memory or testing effects from the first administration (referred to as a carryover effect), but not so long that the underlying concept (e.g., pain, fatigue) might actually change. Testing and order effects can weaken the internal validity of this study design, but the within-patient design provides greater statistical power and decreases sample size requirements. Both testing and order effects should be accounted for as described in most statistical textbooks on the analysis of clinical trials. Detailed statistical methods and example studies are described along with a set of computational algorithms in Sloan, Novotny et al. [72] and Sloan and Dueck [73].

By incorporating the reduced variance estimates that arise from using patients as their own controls, the methods for determining sample size for crossover studies are a slight modification of those described for parallel groups designs above. A simple method of estimating the sample size required for crossover design comparisons of means from two different PRO administration modes is to multiply the total sample size required for a parallel groups design by a factor of $(1-\rho)/2$ where ρ is an estimate of the expected correlation between the two modes of administration (or to be conservative, an estimate of the lower bound). For example, as indicated above, a parallel groups design using equivalence methodology with 234 patients per group can exclude a difference between means of 0.3 standard deviations (equivalent to a small effect size [74]). If we assume an expected value of $\rho=0.9$ then the required sample size is 468 * 0.05= 23.4 (i.e., 24); if we assume an expected value of $\rho=0.7$, then the required sample size is 468 * 0.15=70.2 (i.e., 71). The efficiency of the crossover design explains why it is the most popular design as evidenced by the meta analysis performed by Gwaltney et al [43]. Note that the calculated sample sizes

denote the number of completed pairs of assessments necessary for the analysis and appropriate adjustments should be made for non-completions.

The above sample size calculations are all based upon designs involving comparisons of mean scores. If the endpoint of interest is the intraclass correlation coefficient (ICC), the sample size calculations differ somewhat. First, the sample size for this situation only applies to crossover designs as the ICC is not relevant for parallel group designs. Second, the hypothesis to be tested in this situation is whether the population ICC is sufficiently large to indicate that the scores for the paper PROs and the ePROS are psychometrically equivalent. The test is based on a standard normal test statistic (Z-score) and whether the one-sided confidence interval (lower bound) is above the specified equivalence threshold (e.g., 0.70). For example, 43 patients with complete paired observations would be required for a study to have 80% power to declare that true population reliability is above 0.70 with 95% confidence if the underlying population ICC is 0.85 using Walters methodology [75]. Alternative calculations are possible based on the consistency form of ICC [76] or the 2 sided width of the confidence interval around the ICC [77].

Statistical Methods for Evaluating Measurement Equivalence

The ICC and weighted kappa are useful statistics to measure agreement and, in this case, to test measurement equivalence. Use of Pearson's or Spearman's correlation coefficients alone is not recommended because they are not sensitive to systematic mean differences between groups and as a result tends to overestimate agreement. Methods developed by Bland and Altman [78] combine simple graphical techniques with hypothesis testing for measurement equivalence. Several examples of applications of these measurement equivalence procedures have been published [79-82]. In addition, comparison of mean scores and the evaluation of differential item functioning (discussed briefly below) may be appropriate to assess measurement equivalence.

ICC

The ICC, which can assess both the covariance and degree of agreement between score distributions, has been used most frequently in previous studies that examined the equivalence of paper PROs and ePROs [42]. The ICC provides a means to assess the reliability of scores from an instrument given on multiple occasions or across multiple raters [83]. The ICC takes into account both relative position in the group of scores and the amount of deviation above or below the group mean [84].

Kappa coefficient

Rather than computing simple agreement, which may be high due to chance alone, the kappa coefficient corrects for this by examining the proportion of responses in agreement in relation to the proportion of responses that would be expected by chance alone [85]. The traditional kappa computation only considers absolute agreement and does not credit ratings that are close to one another but not in exact agreement. However, an extension of this approach, called the weighted kappa, considers such "partial" agreement [86]. Weighted kappa and the ICC are similar and in some cases equivalent [87]. Hence, we recommend using ICC in most cases. Fleiss [88] suggests that kappa coefficients of less than 0.40 are poor, 0.40 to 0.59 are fair, 0.60 to 0.74 are good, and greater than 0.74 are excellent. For ICC results, we recommend conforming to the standards for acceptable levels of reliability, specifically at least 0.70 for group comparisons and 0.85 to 0.95 for applications at the individual levels [89, 90].

Comparison of mean scores

Comparing the mean scores obtained on the two modes of administration from the same person [52, 91] or from two equivalent groups can be used to assess measurement equivalence. This approach is most appropriate when the calculation of an ICC is not possible (i.e., in a randomized parallel group design). The difference between modes should not exceed what would be considered the measure's minimally important difference (MID). For those measures for which there is an established MID, the mean difference is evaluated relative to that value. If an MID has not been documented in the literature, then an estimate of the MID is required.

A commonly-used framework for expressing such estimates, endorsed in the FDA draft guidance, is based on Cohen's rules of thumb for effect sizes [74]. A "small" effect size (difference of between 0.20 SD and 0.49 SD) may be meaningful and represent an MID [92-97]. Hence, mean differences between modes of administration in this range warrant further consideration before concluding equivalence. When assessing measurement equivalence, the mean difference between modes should be interpreted relative to an estimate of the mean difference within mode in repeated administrations. In addition, the ICC for ePRO vs. paper administration should be compared to the test-retest ICCs within mode. As noted earlier, the ePRO application should not be held to a higher standard than the original paper-based PRO measure. Further, mode differences may be the result of the better measurement properties of the ePRO device.

Differential item functioning

Another approach to assessing mode equivalence is by using item response theory (IRT) or other approaches to evaluate differential item functioning (DIF) [98, 99]. The probability of responding to each response category for an item should be invariant to mode of administration, conditional on the estimate of underlying score on the domain being measured. For example, people who are estimated to have the same level of physical functioning should have the same probability of answering "not limited at all" to a question about running a mile whether they respond on a self-administered paper questionnaire or over the internet. If the probabilities differ, that is an indication of DIF and lack of mode equivalence. A simple analog to the IRT approach to DIF is to condition on the total domain score rather than the IRT estimated score [100]. Note that for DIF analyses, larger sample sizes (200 minimum; 500 preferred) are needed than the sample size needed for ICCs or weighted kappas.

Other considerations

In addition, the variance and distribution of scores and, when appropriate, the internal consistency reliability, should also be compared. Cronbach's alpha coefficient can be used to estimate the internal consistency reliabilities for the different modes and the significance of the difference in reliability between the modes can be computed [101]. As with the ICC, internal consistency reliability coefficients should be at least 0.70 for group comparisons and 0.85 to 0.95 for applications at the individual level [89, 90]. While DIF can provide important information about lack of equivalence at the item level, it is important to evaluate measurement equivalence corresponding to how the measure will be scored. A PRO measure may have a total score and multiple subscale (domain) scores, therefore the total and subscale scores should be evaluated for measurement equivalence. If item-level DIF is present but operates in different directions, it is possible for there to be measurement equivalence at the scale score level.

Full Psychometric Evaluation

When substantial change has occurred in the PRO measure migration process that has the potential to impact fundamental psychometric properties of the measure, then the measure should be evaluated as if it were a new measure. The topic of PRO questionnaire development and testing is covered sufficiently elsewhere [20, 21, 77, 102] and is likely to require both qualitative and quantitative components. At minimum, the researchers will need to document the content validity (i.e., conceptual framework in the terminology of the FDA draft guidance) of the new PRO measure, and provide evidence supporting internal consistency and test-retest reliability, and construct validity of the measure [22,103]. The sponsor is advised to also consult the draft FDA PRO guidance document for evidentiary requirements for PRO measures that are intended to be used to support labeling claims [18].

Various study designs can be used to evaluate the measurement properties of these new ePRO measures, although most often PRO instruments are evaluated using stand-alone observational studies or within randomized clinical trials. Detailed explication of the psychometric research methods and study designs for psychometric evaluation studies is beyond the scope of this report. However, the main difference between designs for equivalence testing and psychometric validation is the need to assess

validity in the latter, which necessitates the inclusion of a variety of measures extrinsic to the scale of interest. The interested reader is directed to several publications on psychometric evaluation of PRO measures [22, 77, 102, 104].

DISCUSSION AND CONCLUSIONS

It is unreasonable to expect that each specific ePRO application developed from a paper-based PRO measure should undergo full psychometric testing, as if it were a new measure. The expense associated with that process is high, with the potential for little (if any) scientific gain. As long as only minor modifications were made to the measure during the migration process, a substantial body of existing evidence suggests that the psychometric properties of the original measure will still hold for the ePRO version; hence, an evaluation limited to cognitive debriefing and usability testing only may be reasonable. Nevertheless, as with any instrument, ongoing assessment of reliability and validity should continue regardless of the mode of administration. However, where more substantive changes in the migration process has occurred, confirming that the adaptation to the ePRO format did not introduce significant response bias and that the two modes of administration produce essentially equivalent results is necessary. In those cases, there is a need for a practical approach to assessing the measurement equivalence of the ePRO application to the original paper-based measure. Although it is not typically optimal for two administration modes to be used in the same study, there are situations where it happens and may even be advisable (e.g., in a study of a hard-to-reach population where multiple modes improve the overall response rate [105]). In addition, comparability with data from other trials in which the original PRO measure was used is beneficial.

This paper does not address the cross-cultural adaptation of paper PRO measures from one language to ePRO applications for use in other languages or cultures. When standard cross-cultural translation and adaptation procedures [106-108] have been used with the original PRO questionnaire and an acceptable version has been produced, the adaptation of that translated version to an ePRO platform should only require the level of testing necessary based on the changes made in the migration process. However, it

must be recognized that a translation could result in longer items or response labels. Hence, for small screen-based ePRO devices, the fit or placement of items or responses on the screen may be more problematic than that of the original language version. As recommended in all cases, the new ePRO version of a cross-culturally adapted measure should at least undergo usability testing and cognitive debriefing in the target population prior to its use in a clinical trial.

Although not within the scope of this paper, the migration of a measure developed specifically for an electronic data capture device to a paper-based mode of administration may prove to be more problematic than the other way around. The ease of incorporation of skip patterns that are seamless to respondents in electronic data capture is harder to implement on paper questionnaires. Some respondents on the paper form may respond to questions that should be skipped resulting in uncertainty about which responses are reflecting the respondent's true response.

ePRO use in special populations (e.g., visually or cognitively impaired, depressed, limited fine motor skills) was not substantively discussed here since most of the potential problems also exist for paper-based questionnaires. There are issues that may have particular salience with particular ePRO devices such as font size on handheld computers and auditory volume for the hearing impaired on IVR systems. Practical considerations derived from usability testing and cognitive debriefing can inform the decision to use a particular ePRO platform based on the target patient population [109].

We have provided a general framework for decisions regarding the level of evidence needed to support modifications that are made to PRO measures when they are migrated from paper to ePRO devices. The key issues include (1) the determination of the extent of modification required to administer the PRO on the ePRO device and (2) the selection and implementation of an effective strategy for testing the measurement equivalence of the two modes of administration. Not all contingencies could be covered in the context of this paper, but we have attempted to address the most common circumstances. The electronic administration of PRO measures offers many advantages over paper administration. We hope

that our recommendations provide a path forward for researchers interested in migrating PRO measures to electronic platforms.

ACKNOWLEDGEMENTS

The members of ISPOR's ePRO Task Force were: Stephen Joel Coons (Chair), Ethan Basch, Laurie B. Burke, Donald M. Bushnell, David Cella, Chad J. Gwaltney, Ron D. Hays, Joy Hebert, William R. Lenderking, Paula A. Funk Orsini, Dennis A. Revicki, James W. Shaw, Saul Shiffman, Jeff A. Sloan, Brian Tiplady, Keith Wenzel, and Arthur Zbrozek. The steady and capable support of ISPOR and its staff, particularly Elizabeth Molsen, is genuinely appreciated. In addition, the contributions of James Pierce, Damian McEntegart, and Theron Tabor are gratefully acknowledged. The work reflected here was begun by the *ePRO Consensus Development Working Group*, support for which was provided by the following firms: Allergan, assisTek (formerly Assist Technologies), Centocor, ClinPhone, GlaxoSmithKline, invivodata, Merck, Novartis, Pfizer, Sanofi-Aventis, and Takeda. Subsequently, the ePRO Consensus Development Working Group was merged into ISPOR's ePRO Task Force. Additional support was provided by P01 grant (AG020679-01) from the National Institute on Aging, and the UCLA Center for Health Improvement in Minority Elderly/Resource Centers for Minority Aging Research, NIH/NIA/NCMHD, under Grant P30-AG-021684-08.

REFERENCES

1. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743-50.

2. Shiffman S, Hufford MR. Subject experience diaries in clinical research, Part 2: ecological momentary assessment. Applied Clinical Trials 2001;10(3):42-8.

3. Shiffman S, Hufford MR, Paty J. Subject experience diaries in clinical research, Part 1: The patient experience movement. Applied Clinical Trials 2001;10(2):46-56.

4. Wiklund I. Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. Fundam Clin Pharmacol 2004;18:351-63.

5. Leidy NK, Revicki DA, Geneste B. Recommendations for evaluating the validity of quality of life claims for labeling and promotion. Value Health 1999;2:113-27.

6. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9:887-900.

7. Santanello NC, Baker D, Cappelleri JC. Regulatory issues for health-related quality of life – PhRMA Health Outcomes Committee Workshop, 1999. Value Health 2002;5:14-25.

8. Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the patient-reported outcomes (PRO) harmonization group meeting at the Food and Drug Administration, February 16, 2001.
Value Health 2003;5:521-33.

9. Revicki DA. FDA draft guidance and health outcomes research. Lancet 2007;369:540-2.

10. Revicki DA, Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: The PRO evidence dossier. Qual Life Res 2007;16:717-23.

11. Sloan JA, Halyard MY, Frost MH, et al. The Mayo Clinic manuscript series relative to the discussion, dissemination, and operationalization of the Food and Drug Administration guidance on patient-reported outcomes. Value Health 2007;10(suppl 2):S59–S63.

12. Willke RJ, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. Control Clin Trials 2004;25: 535-52.

13. Szende A, Leidy NK, Revicki D. Health-related quality of life and other patient-reported outcomes in the European centralized drug regulatory process: a review of guidance documents and performed authorizations of medicinal products 1995 to 2003. Value Health 2005;8:534-48.

14. Shiffman S. Delivering on the eDiary promise. Applied Clinical Trials 2005;14(9):64.

15. Wiechers OA. The move to EDC. Applied Clinical Trials 2002;11(11):38-40.

16. Saponjic RM, Freedman S, Sadighian A. What monitors think of EDC: results of a survey of U.S. monitors. Applied Clinical Trials 2003;12(5):50-2.

17. Getz KA. The imperative to support site adoption of EDC. Applied Clinical Trials 2006;15(1):38-40.

18. US Food and Drug Administration. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (DRAFT), February 2006. Available at: http://www.fda.gov/cder/guidance/5460dft.pdf [Accessed June 1, 2008].

19. Rothman ML, Beltran P, Cappelleri JC, et al. Patient-reported outcomes: conceptual issues. Value Health 2007;10(suppl 2):S66–S75.

20. Snyder CF, Watson ME, Jackson JD, et al. Patient-reported outcome instrument selection: designing a measurement strategy. Value Health 2007;10(suppl 2):S76–S85.

21. Turner RR, Quittner AL, Parasuraman BM, et al. Patient-reported outcomes: instrument development and selection issues. Value Health 2007;10(suppl 2):S86–S93.

22. Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007;10(suppl 2):S94-S105.

23. Sloan JA, Dueck AC, Erickson PA, et al. Analysis and interpretation of results based on patient-reported outcomes. Value Health 2007;10(suppl 2):S106-S115.

24. Revicki DA, Erickson PA, Sloan JA, et al. Interpreting and reporting results based on patient-reported outcomes. Value Health 2007;10(suppl 2):S116-S124.

25. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007;10(suppl 2):S125-S137.

26. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. Public Opinion Quarterly 1996;60:275-304.

27. Taenzer PA, Speca M, Atkinson MJ, et al. Computerized quality-of-life screening in an oncology clinic. Cancer Practice 1997;5:168-75.

28. Bloom DE. Technology, experimentation, and the quality of survey data. Science 1998;280:847-8.

29. Velikova G, Wright EP, Smith AB, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol 1999;17:998-1007.

30. Stone AA, Shiffman S, Schwartz JE, et al. Patient noncompliance with paper diaries. BMJ 2002;324:1193-4.

31. Bushnell DM, Reilly MC, et al. Validation of electronic data capture of the Irritable Bowel Syndrome—Quality of Life Measure, the Work Productivity and Activity Impairment Questionnaire for Irritable Bowel Syndrome and the EuroQol. Value Health 2006;9:98-105.

32. US Food and Drug Administration. Code of Federal Regulations – Title 21 – Food and Drugs. Available at: http://www.fda.gov/cdrh/aboutcfr.html [Accessed June 1, 2008].

33. US Food and Drug Administration. Guidance for Industry: E6 Good Clinical Practice: Consolidated Guidance, April 1996. Available at: http://www.fda.gov/cder/guidance/959fnl.pdf [Accessed June 1, 2008].

34. US Food and Drug Administration. Guidance for Industry: Computerized Systems Used in Clinical Investigations, May 2007. Available at: http://www.fda.gov/cber/gdlns/compclintrial.pdf [Accessed June 1, 2008].

35. US Food and Drug Administration. Guidance for Industry—Part 11, Electronic Records; Electronic Signatures—Scope and Application, August 2003. Available at: http://www.fda.gov/cder/guidance/5667fnl.pdf [Accessed June 1, 2008].

36. Raymond SA, Meyer GF. Interpretation of regulatory requirements by technology providers. Applied Clinical Trials 2002;11(7):50-8.

37. Farrell J, Cooper M. Navigating the new 21 CFR 11 guidelines. Applied Clinical Trials 2004; 13(3):67-70.

38. Dale O, Hagen KB. Despite technical problems personal digital assistants outperform pen and paper when collecting patient diary data. J Clin Epidemiol 2007;60:8-17.

39. Hahn EA, Cella D, Dobrez D, et al. The talking touchscreen: a new approach to outcomes assessment in low literacy. Psychooncology 2004;13:86-95.

40. Crow JT. Receptive vocabulary acquisition for reading comprehension. Modern Languages Journal 1986;70:242-50.

41. Henriksen B. Three dimensions of vocabulary development. Studies in Second Language Acquisition 1999;21:303-17.

42. 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;11:322-33.

43. Krystal AD, Walsh JK, Laska E, et al. Sustained efficacy of eszopiclone over 6 months of nightly treatment: results of a randomized, double-blind, placebo-controlled study in adults with chronic insomnia. Sleep 2003;26:793-9.

44. Mundt JC, Marks IM, Shear MK, Greist JH. The Work and Social Adjustment Scale: a simple measure of impairment in functioning. Br J Psychiatry 2002;180:461-4.

45. Mundt JC, Kobak KA, Taylor LV et al. Administration of the Hamilton Depression Rating Scale using interactive voice response technology. MD Computing 1998;15(1):31-9.

46. Mundt JC, Katzelnick DJ, Kennedy SH, et al. Validation of an IVRS version of the MADRS. J Psychiatr Res 2006;40:243-6.

47. Rush A, Bernstein I, Trivedi M et al. An evaluation of the Quick Inventory of Depressive Symptomalogy and the Hamilton Rating Scale for Depression: A sequenced treatment alternatives to relieve depression trial report. Biol Psychiatry 2006;59(6):493-501.

48. Alemi F, Stephens R, Parran T, et al. Automated monitoring of outcomes: application to the treatment of drug abuse. Med Decis Making 1994;14:180-7.

49. Agel J, Greist JH, Rockwood T, et al. Comparison of interactive voice response and written self-administered patient surveys for clinical research. Orthopedics 2001;24:1155-57.

50. Dunn JA, Arakawa R, Greist JH, Clayton AH. Assessing the onset of antidepressant-induced sexual dysfunction using interactive voice response technology. J Clin Psychiatry 2007;68:525-32.

51. Fricker RD, Schonlau M. Advantages and disadvantages of internet research surveys: Evidence from the literature. Field Methods 2002;14:347-67.

52. Bushnell DM, Martin ML, Parasuraman B. Electronic versus paper questionnaires: a further comparison in persons with asthma. J Asthma 2003;40:751-62.

53. 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.

54. 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;10:15-21.

55. Caro Sr JJ, Caro I, Caro J, et al. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? Qual Life Res 2001;10:683-91.

56. Crawley JA, Kleinman L, Dominitz J. User preferences for computer administration of quality of life instruments. Drug Information Journal 2000;34:137-44.

57. Allenby A, Matthews J, Beresford J, McLachlan SA. The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. Eur J Cancer Care 2002;11:245-53.

58. Kleinman L, Leidy NK, Crawley J, et al. A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire. Med Care 2001;39:181-9.

59. Cook AJ, Roberts DA, Henderson MD, et al. Electronic pain questionnaires: a randomized, crossover comparison with paper questionnaires for chronic pain assessment. Pain 2004;110: 310-17.

60. Drummond HE, Ghosh S, Ferguson A, et al. Electronic quality of life questionnaires: a comparison of pen-based electronic questionnaires with conventional paper in a gastrointestinal study. Qual Life Res 1995;4:21-6.

61. Palermo TM, Valenzuela D, Stork PP. A randomized trial of electronic versus paper pain diaries in children: impact on compliance, accuracy, and acceptability. Pain 2004;107:213–9.

62. Shields A, Gwaltney C, Tiplady B, et al. Grasping the FDA's PRO Guidance. Applied Clinical Trials 2006;15(8):69-72.

63. Willis GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, California: Sage Publications, 2005.

64. Quigley D., Elliott MN, Hays RD, et al. Building a bridge: continuity in measuring patient experiences of care when converting to the CAHPS® hospital survey. Med Care. In press..

65. Willis G, Reeve BB, Barofsky I. The use of cognitive interviewing techniques in quality of life and patient-reported outcomes assessment. In: Lipscomb J, Gotay CC, Snyder C (Eds.). Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005:610-22.

66. Willis GB, DeMaio TJ, Harris-Kojetin B. Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. In: Sirken MG et al. (Eds.). Cognition and Survey Research. New York: John Wiley & Sons, 1999:133-53.

67. Ojanen V, Gogates G. A briefing on cognitive debriefing. Good Clinical Practice Journal 2006;12:25-9.

68. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigourous methods. BMJ 1996;313:36-9.

69. Fleiss JL, Kingman A. Statistical management of data in clinical research. Crit Rev Oral Biol Med 1990;1:54-66. Available at: http://crobm.iadrjournals.org/cgi/reprint/1/1/55 [Accessed June 13, 2008].

70. Blackwelder WC. Current issues in clinical equivalence trials. J Dent Res 2004;83(Spec Iss C):C113-C115.

71. Atherton SP, Sloan JA. Design and analysis of equivalence trials via the SAS system. SUGI Proceedings 1998;23:1166-71.

72. Sloan JA, Novotny P, Loprinzi CL, Ghosh M. Graphical and analytical tools for two-period crossover clinical trials. SUGI Proceedings 1997;22:1312-7. Available at: http://www2.sas.com/proceedings/sugi22/STATS/PAPER280.PDF [Accessed June 13, 2008].

73. Sloan JA, Dueck A. Issues for statisticians in conducting analyses and translating results for quality of life end points in clinical trials. J Biopharm Stat 2004 Feb;14(1):73-96.

74. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1988.

75. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101-10.

76. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med 2002;21:1331-5.

77. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use (3rd ed.). New York: Oxford University Press, 2003.

78. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995;24(3):7-14.

79. Gonin R, Lloyd S, Cella DF. Establishing equivalence between scaled measures of quality of life. Qual Life Res, 1996;5:20-26.

80. Marshall, GN, Hays RD, Nicholas R. Evaluating agreement between clinical assessment methods. Int J Methods Psychiatr Res 1994;4:249-57.

81. Sloan JA. Statistical issues in the application of cancer outcome measures. In: Lipscomb J, Gotay CC, Snyder C, editors. Outcomes Assessment in Cancer: Measures, Methods, and Applications. New York: Cambridge University Press, 2005:362-85.

82. Smith DJ, Huntington J, Sloan JA. Choosing the "correct" assessment tool. Curr Probl Cancer 2005;29:272-82.

83. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.

84. Anastasi A, Urbina S. Psychological Testing (7th ed.). Upper Saddle River, New Jersey: Prentice Hall, 1996.

85. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

86. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.

87. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 1973;33:613-9.

88. Fleiss JL. Statistical Methods for Rates and Proportions (2nd ed.). New York: Wiley, 1981.

89. Nunnally JC, Bernstein IH. Psychometric Theory (3rd ed.). New York: McGraw-Hill, 1994.

90. Weiner EA, Stewart BJ. Assessing Individuals. Boston: Little Brown, 1984.

91. Ramachandran S, Lundy JJ, Coons SJ. Testing the measurement equivalence of paper and touch-screen versions of the EQ-5D visual analog scale (EQ-VAS). Submitted for publication.

92. Guyatt GH, Osoba D, Wu AW, et al. Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002; 77(4):371-83.

93. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582-92.

94. Sloan J, Symonds T, Vargas-Chanes D, Fridley B. Practical guidelines for assessing the clinical significance of health-related quality of life changes within clinical trials. Drug Information Journal 2003;37:23-31.

95. Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported questionnaire data: another step toward consensus. J Clin Epidemiol 2005;58:1217-9.

96. Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? Expert Rev Pharmacoeconomics Outcomes Res 2004;4:515-23.

97. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2005;2:63-7.

98. Teresi JA. Overview of quantitative measurement methods: equivalence, invariance, and differential item functioning in health applications. Med Care 2006;44 (suppl 3):S39-S49.

99. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res 2007;16:33-42.

100. Crane PK, Gibbons, LE, Ocepek-Welkson K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res 2007;16:69-84.

101. Feldt LS, Woodruff KJ, Saith FA. Statistical inference for coefficient alpha. Applied Psychological Measurement 1987;11:93-103.

102. Hays RD, Revicki D. Reliability and validity (including responsiveness). In: Fayers P, Hays RD (Eds). Quality of Life Assessment in Clinical Trials (2nd ed.). Oxford, UK; Oxford University Press, 2005:25-39.

103. Lohr K. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11:193-205.

104. 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 Outcome Measurement Information System (PROMIS). Med Care 2007;45:S22-S31.

105. Hepner KA, Brown J A, Hays RD. Comparison of mail and telephone in assessing patient experiences in receiving care from medical group practices. Eval Health Prof 2005;28:377-89.

106. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417-32.

107. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcome (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8:94-104.

108. Marquis P, Keininger D, Acquadro C, de la Loge, C. Translating and evaluating questionnaires: cultural issues for international research. In: Fayers P, Hays RD (Eds). Quality of Life Assessment in Clinical Trials (2nd ed.). Oxford, UK; Oxford University Press, 2005:77-93.

109. Hahn EA, Cella D. Health outcomes assessment in vulnerable populations: Measurement challenges and recommendations. Arch Phys Med Rehabil 2003;84(Suppl.):S35-42.

Table 1: PRO to ePRO Measurement Equivalence: Instrument Modification and Supporting Evidence

Level of Modification	Rationale	Examples	Level of Evidence
Minor	The modification can be justified on the basis of logic and/or existing literature. No change in content or meaning.	1) Non-substantive changes in instructions (e.g., from circling the response to touching the response on a screen). 2) Minor changes in format (e.g., one item per screen rather than multiple items on a page).	Cognitive debriefing Usability testing
Moderate	Based on the current empirical literature, the modification cannot be justified as minor. May change content or meaning.	1) Changes in item wording or more significant changes in presentation that might alter interpretability. 2) Change in mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]).	Equivalence testing Usability testing
Substantial	There is no existing empirical support for the equivalence of the modification and the modification clearly changes content or meaning	1) Substantial changes in item response options 2) Substantial changes in item wording	Full psychometric testing Usability testing

Adapted from Shields et al. [62].