I. Introduction

The incorporation of the patient perspective in the evaluation of medical products (i.e., drugs, biologicals, devices) is increasingly important and considered essential in many cases. Medical products aimed at relieving patients’ symptoms and/or improving levels of self-reported functioning will require measures of patient-reported outcomes (PROs) as endpoints in clinical trials. A PRO instrument systematically collects treatment benefit data directly from patients, without interpretation by clinicians or others (FDA, 2009). As stated in the 2009 US Food and Drug Administration (FDA) Guidance for Industry titled Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (“PRO Guidance”), “Use of a PRO instrument is advised when measuring a concept best known by the patient or best measured from the patient perspective” (FDA, 2009).

There is no doubt that the release of the FDA’s PRO Guidance has focused increased attention on the development and use of scientifically sound measurement of PRO endpoints in clinical trials. In addition, clinician-reported outcome (ClinRO) measures, observer-reported outcome (ObsRO) measures, and performance outcome (PerfO) measures are receiving increasing attention as well (FDA 2013). ClinRO measures are completed by clinicians and are often based on clinical interviews (e.g., Hamilton Depression Rating Scale [HAM-D] in depression trials). ObsRO measures are completed by non-clinical informants (e.g., spouse, caregiver, parent, or teacher) and report on observable disease- and/or treatment-related concepts (e.g., activities of daily living inventory completed by caregivers in Alzheimer’s disease [AD] trials).

PerfO measures are assessments based on a task performed by a patient according to instructions administered by a health care professional. They rely on the cooperation, ability, and motivation of the subject. Examples of PerfO measures include tests of sensory function (e.g., visual acuity), cognitive function (e.g., Digit Symbol Substitution Test in AD trials), and physical performance (e.g., timed 25-foot walk test in multiple sclerosis trials). Although the final PRO Guidance only addressed PRO instruments when it was released in 2009, the FDA held a public workshop in October 2011 where they discussed the need for the same level of evidence: well-defined and reliable measurement for all clinical outcome assessment (COA) tools (i.e., PRO, ClinRO, ObsRO, and PerfO measures) intended to support medical product labeling claims (FDA, 2011; FDA, 2013). Likewise, this Task Force Report focuses on PRO instruments but the recommendations apply to ClinRO and ObsRO measures as well.

In addition to the FDA’s increased focus on well-defined and reliable assessment of clinical trial endpoints, one of the most important developments in the field of PRO measurement has been the emergence of technologies that enable the collection of data electronically. Advantages of using electronic data collection include less subject burden, avoidance of secondary data entry errors, easier implementation of skip
patterns, date and time stamping, and more accurate and complete data (Hyland et al. 1993; Tourangeau et al. 1996; Taenzer et al. 1997; Bloom 1998; Velikova et al. 1999; Stone et al. 2002; Bushnell et al. 2006). With the increasing availability of multiple modes\(^1\) of PRO data collection, including both paper and various electronic types, the opportunity exists to mix these modes within and across clinical trials in a medical product development program.

While mixing of modes within and across clinical trials may meet the needs of global product development programs where the patient population and access to technology vary considerably within and across regions, such mixing may in fact be an avoidable source of measurement error. It is, therefore, the general recommendation of this ISPOR PRO task force report that PRO data collection modes not be varied within a single clinical trial or between trials that seek to pool or compare the data without prior evidence of sufficient measurement equivalence between the modes.

This general recommendation is based upon the basic research design tenet that anything with the potential to introduce measurement error into a trial should be avoided (Streiner and Norman, 2008). Measurement error is, in essence, noise (error variance) that reduces statistical power and attenuates the ability of the trial to detect real change (i.e., treatment effect) in the trial endpoint. In the context of collecting PRO data – where patients are providing information directly – there are many unavoidable sources of measurement error, including: differences introduced by the need to translate and culturally adapt multiple versions of a PRO instrument; specific cultural biases introduced by differing experiences of the medical condition being studied (Gnanasakthy et al 2013); and, the variability in patient’s ability to reflect and provide a response.

Potential error variance can also be introduced into the trial design by different data collection modes used within the trial that do not provide comparable data (i.e., the modes lack sufficient measurement equivalence.) As the mode of PRO data collection is a part of the research design, it should be possible (even though challenging at times) to decide on and deploy a single consistent mode of PRO data collection in the trial. The recommendation of this task force report is to avoid, where possible, all potential sources of measurement error, including mixed modes of PRO data collection.

---

\(^1\) Before proceeding further, a clarification regarding terminology is in order. It should be noted that the term mode of data collection as used in this report differs from the FDA’s terminology. The PRO Guidance makes a distinction between PRO instrument administration modes and data collection methods. According to the PRO Guidance, administration mode refers to self- vs. interviewer-administered PRO measurement, while data collection method refers to the tool used for capturing the data such as paper-based questionnaires, web-based data entry, interactive voice response systems (IVRS), or any of the other ePRO devices (FDA, 2009). (Note: An interviewer administered PRO measure is not a ClinRO measure because the patient’s responses are not interpreted, but simply recorded, by the interviewer.)

We find that the distinction made by the FDA is potentially misleading because the term “mixed methods” in the larger PRO measurement field refers to mixing qualitative and quantitative methods in research. This term is not associated with multiple methods of data collection. On the other hand, the PRO measurement field has a long history of using the term “mixed modes” to refer both to administration, as well as data collection (i.e., ePRO vs. paper). Therefore to simplify the discussion in this report, we use the term mode in the context of both modes of administration per the PRO Guidance and modes of data collection per the PRO measurement field.
However, although it may not be optimal, mixing of PRO data collection modes within trials does occur and has to be addressed pragmatically. When modes have been directly compared in cross-sectional studies, there is evidence that PRO data collected electronically can be comparable to that obtained by paper-based data collection, particularly with screen-based devices (Gwaltney et al. 2008). However, the literature is not definitive and can be limited by selective reporting; it has been well-documented that studies with positive findings are more likely to be published than those with inconclusive or negative results (Rosenthal 1979; Song et al. 2009). In addition, comparability of data collected on different modes is likely dependent on the specific PRO measure being used; hence, a general assumption of measurement equivalence between or among modes may not always hold. While some evidence has shown comparability between paper and visual modes (Gwaltney et al. 2008; Ramachandran et al. 2008) or between paper, Web-based, and/or IVR modes (Bennett et al. 2013a; Bennett et al. 2013b; Lundy and Coons 2011; Lundy et al. 2013), more evidence is needed to support mixing modes within a trial setting to ensure it has minimal impact on the trial results. Nevertheless, this task force does not rule out the possibility that, at some point in the future, sufficient evidence will be available to support the assumption of measurement equivalence across modes in most circumstances where an optimal migration has occurred.

Further, it is clear that this issue of mixing modes was contemplated by FDA in development of its PRO Guidance. Specifically, the PRO Guidance states that “We intend to review the comparability of data obtained when using multiple data collection methods or administration modes within a single clinical trial to determine whether the treatment effect varies by methods or modes” (FDA, 2009). The PRO Guidance does not, however, discuss ways for clinical trial designs to ensure the comparability of the data when mixed modes are used. Although this task force report specifically addresses multiple modes of PRO data collection, it is assumed that many of the same issues are involved when multiple modes of ClinRO and ObsRO data collection are considered. Therefore, this report addresses the key issues in mixing modes, specifically focusing on how to reduce the impact on measurement error when these situations arise.

The launching point for this task force report is the prior ISPOR ePRO task force report by Coons et al. (2009) that addressed the evidence needed to support measurement equivalence when migrating from paper to electronic modes of PRO data collection. According to that report, “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.” (p 2) Thus, establishing measurement equivalence is essential in demonstrating that the migration from paper to electronic, or for that matter from any data collection mode to another, did not impact the instrument’s meaning, interpretation, and resulting responses. In the context of the current task force report, we use the term ‘measurement equivalence’ to emphasize the need for the instrument to be measuring the same thing regardless of the mode.
Coons et al. (2009) did not address the issues to take into account when considering mixing two or more modes of PRO data collection in a single trial or across trials intended to be compared or pooled. This current report builds on the recommendations for changing modes of administration in the original ISPOR ePRO task force report by providing additional recommendations regarding good research practices for migration across modes of data collection and an in-depth exploration of the assessment of measurement equivalence between original and migrated versions of PRO instruments, particularly in the context of mixing data collection modes. In addition, we discuss issues that must be considered in order to avoid sources of measurement error that materially impact the meaning and interpretation, and consequently the measurement properties, of the instrument being used to assess PRO endpoints in clinical trials. The report concludes with recommendations for operational and statistical considerations when modes are mixed in a clinical trial setting. The overall objective of the report is to address the use and mixing of data collection modes within and between trials where the PRO endpoints are intended to be used to support medical product labeling.

The report is organized as follows: After describing the task force process, we first describe the most common modes of PRO data collection currently available. We then address the factors that should be considered when selecting a mode or modes of PRO data collection in a clinical trial. Next, a summary of how to ‘faithfully’ migrate instruments is presented, followed by a section on study designs used to evaluate measurement equivalence of the new and original modes of data collection. Finally, we discuss a number of issues that must be taken into account when mixing modes is deemed necessary or unavoidable within or between trials.

II. Task Force Process

<We will describe task force process after the report is finalized.>

III. Modes of PRO Data Collection

The emergence of new technologies allows trial protocols to be written in which data collection schedules and locations can support more timely and convenient assessment of endpoints. Selecting the appropriate mode of data collection is essential to the success of the trial. Prior to examining the key issues in selecting one or more modes of data collection, we first present those that are most widely available.

Paper-and-Pencil

While the trend seems to be changing over time, the initial version of most PRO measures has been developed on paper. Hence, self-administration involves a paper questionnaire and a writing implement
such as a pencil or pen. Unlike most newer technologies, the data are not captured electronically; the source data remains the completed, hard-copy questionnaire. Written responses on the questionnaire are entered into an electronic database through manual (keyed/typed) data entry or optical scanning. This step has the potential to introduce error associated with secondary data entry. In addition, the lack of real-time electronic notation of the date and time the data are collected is a significant weakness of paper-and-pencil data collection. This is particularly the case with data intended to be collected at relatively frequent intervals in unsupervised settings (e.g., daily symptom diary) as prospective or retrospective (parking lot syndrome) data entry is possible (Stone 2002). Nevertheless, paper-and-pencil PRO data collection remains prevalent.

**Digital Pen**

A hybrid of the paper-and-pencil mode is the digital pen. Like traditional paper-and-pencil data collection, study subjects write their responses on a paper questionnaire but, with digital pen technology, the paper has been specially printed to enable the digital pen to identify where it is on the questionnaire and what response is being written. Hence, other than the need for customized, proprietary printing software, the level of change made to the PRO instrument during migration may be minimal.

Digital pens are somewhat thicker than standard pens; they contain a pressure sensor, a camera that captures the written data, a microprocessor, memory, a battery and a ballpoint ink cartridge. Some digital pens also include a Bluetooth transceiver. The responses captured by the pen’s camera can be uploaded in real time via a wireless connection or through a pen docking cradle attached (USB) to a computer that uploads the data via the internet to a web-based server. In addition, the completed paper questionnaire is available for verification of any data in question in the case of queries and is often treated as the source data during source data verification processes.

**Handheld Devices**

The initial handheld electronic devices, introduced in the late 1980s, were referred to as personal digital assistants (PDAs), and were primarily intended to assist people in managing their calendars, and more broadly, their lives. This platform was adapted to capture diary or field-based data electronically. This was among the first types of electronic PRO data collection - ePRO. Since the 1980s, the small handheld device platform has exploded. These are functionally small platform computers that integrate wireless and telephony capabilities with scaled down user interface (UI) elements. These devices are now mobile computers and telephones, referred to as smart phones. This explosion in platform options, and sizes (while still relatively small to fit in the user’s hand), has been leveraged to capture a broad array of PRO data in a broad range of settings.
Another subset of the handheld device category are purpose built devices such as the Asthma Monitor 3 (AM3; manufactured by ERT) and the In2itive eDiary (manufactured by Vitalograph). Both devices are used in respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) and integrate diary data collection and the ability to capture peak expiratory flow (PEF) readings or forced expiratory volume in 1 second (FEV$_1$) in a single device. These proprietary devices were developed to provide subjects with a single device to collect parallel data streams which can reduce burden in clinical trials.

The primary advantage of the small devices is portability. Handheld devices have become the mainstay of electronic PRO data collection, particularly for trials in which frequent field-based data capture is required (e.g., symptom diaries; episodic, event-based diaries). Due to the small size of the screen, usually only a single item/question and its response options are displayed at one time. Responses are typically entered by touching the screen at the appropriate position using a stylus or fingertip; however some older devices require a keypad or keyboard for entry. With wireless and cellular phone technology, data can then be immediately transferred from the handheld device to a central server location.

**Tablet Computers**

Tablet computers (aka tablets) are also relatively mobile data collection devices with integrated touch screens that are operated by touching the screen with a stylus or fingertip. The Apple iPad™ is an example of a tablet computer. Tablets are generally smaller than laptop computers but larger than traditional PDAs or mobile phones. This is both an advantage and a disadvantage. The bigger screen size enables larger presentation of items and response options and allows the possibility of multiple items per screen. However, the larger device size decreases its portability/mobility and usability as an off-site data collection mode, particularly for event-based data capture or field-based data collection.

**Desktop or Laptop Computers**

The final type of computing platform used to collect PRO data is computers that have larger screens. These generally lack touch screen functionality and utilize UI elements such as a keyboard and/or mouse to enter responses. Within the context of clinical trials, they are not routinely intended to be mobile. Desktop, and more recently, laptop computers, are among the original platforms used for collecting PRO data, prior to the introduction of handheld and tablet devices. Both historically, and today, such platforms are used to capture site-based PRO data. These are typically completed when a subject comes to the investigative site, but such platforms can be used from the subject’s home, or anywhere else as well, if the PRO data collection system utilizes Web-based technology. The primary advantage of the desktop or laptop is that the screen size is relatively large, compared to handheldds and tablets, and can accommodate virtually any representation of a PRO instrument.
Interactive Voice Response Systems

Automated telephone-based data collection is commonly referred to as interactive voice response (IVR). IVR systems interact with callers using a pre-recorded voice question and response option script. Contrary to IVR’s name, a touch-tone response entry using the phone’s keypad, not voice, is the primary response mechanism currently used. An assumption may be that future voice recognition technology will allow for true interactive voice capability for this data collection mode. IVR systems can be implemented to allow incoming or outgoing calls. Hence, trials can be designed such that the study subject is called at data collection time points, the subject places the call at data collection time points, or both. Some of the advantages of IVR systems are that no additional hardware is required for the subject other than a landline or cellular/mobile telephone, little if any, subject training is necessary, and data are stored directly to the central database.

Stand-Alone vs. Web-based Systems

The previous descriptions of PRO data collection options have focused on the form factor, either the size of the computer, or whether the system uses auditory or visual representation of the PRO instrument. Another dimension of PRO data collection systems worthy of consideration is whether the system is stand-alone or Web-based. In this context, a stand-alone system is self-contained with all of the software and functionality located in the device (although often with a link to web-based database and management systems) vs. a Web-based system where the device is simply a vehicle to access the system through a Web browser.

This distinction is important and worth mentioning here because there is a substantial difference in the control of the presentation of the PRO instrument between these systems. With a stand-alone system, the subject will only have access to the PRO instrument from the specific device on which it is deployed for that trial, whether a handheld, tablet, desktop, or laptop. There will be consistency of the presentation and interface with the PRO instrument. In contrast, with a Web-based system, there are a number of variations that warrant consideration. One typical implementation is the use of a browser (e.g., Internet Explorer) to access a central system that then displays the PRO instrument on a Web-enabled device screen.

The subject can potentially utilize multiple device types (though this may be restricted to selected browsers/technologies), and therefore, the presentation and interface with the PRO instrument can vary within the trial. Such Web-based systems can, depending on the variability of devices allowed to access the PRO instrument, functionally be different modes of PRO data collection. Thus, the Web-based system, while potentially offering considerable flexibility for completion of the PRO instrument, can also, in some circumstances, be considered a mixed mode of PRO data collection. The amount of error variance caused by this issue is unknown and worthy of further research if we are to move to this more flexible system. It
should also be noted that many Web-based systems cannot be utilized where there is no internet access and require a constant internet connection in contrast to stand-alone systems that can be used offline.

**Mobile Applications (“apps”)**

Another, more recent, option is the downloading of a software application, also known as an “app,” to a smartphone device (e.g., Apple iPhone™ or Android phone). The user runs the app locally on the device and completes the PRO instrument. The data are then transmitted via the internet or cellular signal. When the app is downloaded onto the smartphone, the questionnaire resides on the device and the device becomes, in essence, a stand-alone system. There is increasing interest in developing such instrument apps for PRO data collection, which would allow subjects to use their own technology with which they are comfortable and familiar, rather than having to carry around yet another device. This approach is gaining momentum in what are called “bring your own device” (BYOD) trial contexts (Taylor 2013). The use of such apps in BYOD trial settings has not yet been fully evaluated by the scientific community nor the FDA, but it has significant potential.

### IV. Process for Selecting the Appropriate Mode of Data Collection

The selection of a PRO-based clinical trial endpoint measure and the mode of PRO data collection should not be an afterthought in a drug development program. All too often, the PRO data collection mode appears to be given insufficient attention with providers of ePRO technologies and services asked to accomplish the near impossible prior to the launch of a trial. Hence, as early as possible, a substantial amount of thought and deliberation should be invested in the selection and evaluation of the mode of PRO data collection for a clinical trial. A lead time of 6 months is ideal as ePRO development activities are front-loaded and must occur prior to the launch of the trial. There are a number of factors that must be considered, including patient population, location of data collection, characteristics of the instrument (e.g., length, format of responses/answers), data collection schedule (which is driven by the type of outcome being assessed), feasibility, and cost.

**Patient Population**

The primary consideration for selection of a PRO data collection mode is the patient population that will be asked to provide the self-reported data. The characteristics of clinical trial subjects, particularly sensory and physical abilities, will be important drivers of the choice of modes. It should be noted that this is not a new consideration. Historically, when data collection mode options were limited to paper and pencil, or an interviewer reading to the patient, we had a limited ability to respond to the variability in patients’ capabilities. Given the diversity of options now available, we can (and should) be more responsive to the patient population’s needs. For example, subjects that have non-correctable visual or hearing impairments will require an auditory or visual based data collection system, respectively. Furthermore, in conditions
where there are decrements in physical function (e.g., joint stiffness, tremors) or patients’ physical abilities are compromised, such as rheumatoid arthritis or Parkinson’s disease, both the selection and the specifics of the data collection mode will be important. Auditory systems may be good for such patients, or visual systems that have larger font sizes for reading or larger stylus sizes for arthritis sufferers would be helpful.

**Location of Data Collection**

In a clinical trial, PRO data may be collected from subjects at the investigative site (e.g., clinic), in the field or away from the study site (e.g., subject’s home or workplace), or both. At the study site, the portability of the data collection mode is not as critical; hence, all modes are potentially viable. However, if the data collection takes place in the field, then subject convenience and portability are important considerations.

**The Characteristics of the PRO Instrument**

The characteristics of the PRO questionnaire can be a critical driver in selecting the mode of data collection. With regard to the length of the PRO instrument, both the number of items and the amount of time necessary to complete the items should be considered. It should be noted that subject burden is an issue to bear in mind, regardless of the data collection mode. Raymond (2010) makes the distinction between “questionnaires” and “diary-type reports,” with the latter comprising fewer concepts with questions that are completed at least daily. Handheld devices have become the mainstay of field-based data collection (eDiaries) in clinical trials, but they are less than optimal for longer PRO instruments. Long and time-consuming questionnaires can be physically and/or cognitively fatiguing and should be avoided. Likewise, there may be aspects of the data collection mode to consider that may mitigate or aggravate the fatigue factor.

Depending on the length or complexity of the response options, screen size can be a limitation. If information on a screen is needed to inform or interpret the task or content on a subsequent screen, the implementation of the PRO instrument is far from optimal (Tiplady 2010). Hence, the response options should appear on the same screen as the question. Scrolling to access response options should not be required. In the context of IVR systems, memory may be required to enable the subject to select among response options; numerous or lengthy response options complicate task completion. In addition, there are some types of response formats that are not easily operationalized on all data collection platforms. For instance, a traditional visual analog scale (VAS), which is a line with descriptive anchors at each end (e.g., “No pain” to “Pain as bad as it could be”) with no intermediate positions along the continuum, does not lend itself to administration on an auditory system (e.g., IVR). Open-ended or free-text responses tend to be more burdensome on tablets and handheld devices because an on-screen keyboard is required for text entry.

**Data Collection Schedule**

The frequency of protocol-driven data collection points in a clinical trial should also be considered. Technology has enabled greater flexibility and functionality for designing trials with more frequent data
capture. Trial protocols may require that PRO data is collected monthly, weekly, daily, or multiple times per day. Within a day, data capture can be scheduled for specific times (e.g., 7 AM and 7 PM) or based on an event (e.g., bowel movement) or symptom (e.g., pain). The choice of mode must consider the multiple places where the subject may be when the data collection is to occur. Hence, for multiple data collection points during the day, the portability of the device is a major consideration assuming the patient is mobile.

**Feasibility of Implementation**

Another consideration for mode selection is the infrastructure available in the selected locale of an investigative site or in the trial more broadly. Some regions within a country and some countries more broadly may not be able to support certain technologies. For example, if high-speed internet access or a cellular phone network is critical for a data collection technology, this will be a criterion for selecting this mode in trial or at a specific site. Variability in feasibility of implementation across investigative sites within a trial can potentially lead to mixing of modes.

**Cost**

The reality of conducting clinical trials is that cost will be a factor in determining the mode of data collection. While this is less than ideal with regard to choosing the optimal mode to collect high quality data, cost is, nonetheless, a significant driver for selecting a mode. The team making the mode of data collection decision will need to balance the above selection criteria against available funds to make the best decision for a specific circumstance.

It should be noted that if a less expensive option is chosen early in a medical product’s development, the team will need to consider using a different mode (i.e., mixing modes) later in the program. *It is likely that choosing one mode of data collection will be more cost effective than multiple modes in the same trial.*

Furthermore, it cannot be assumed that paper mode of data collection will always be less expensive than an electronic mode; there are hidden costs with paper including the time required for secondary data entry into the system and time spent on “data cleaning” and queries prior to database lock. Therefore, when cost comparisons are made, the full cost of a data collection mode(s) accrued across the lifetime of the study -- not just the upfront costs associated with ePRO implementations -- should be taken into consideration. Finally, it should be noted that the costs associated with electronic modes of data collection change as technology evolves.

After the appropriate mode of data collection has been selected, the next step in the process is to migrate the PRO instrument to the newly selected mode before implementing PRO measurement in the planned clinical trial program, assuming the chosen mode is a new one for the PRO instrument. In some instances, multiple modes may be selected for data collection and migration may occur concurrently. Following migration, it is important to assess if the new mode has measurement equivalence with the original mode, that is, *subjects interpret and respond to the instrument the same way regardless of the mode.*
V. Migration

The discussion of the optimal migration process is relevant to the discussion of potentially using mixed modes because it is necessary to migrate and evaluate measurement equivalence before mixing so that the desired modes are available and appropriate for use. Since there is little, if any literature, available on the migration process, these recommended good practices are based on previous successful migrations conducted by members of the task force that led to demonstrations of measurement equivalence.

The goal of any migration is to have minimal or no impact on the measurement characteristics of the instrument. A “faithful migration” refers to the development of alternative modes of data collection that do not introduce response bias that results from changes in the way the instrument is presented/formatted or how the subject interacts with it. The most common path is migrating from paper to electronic modes, but migration from one electronic mode to another or from electronic to paper will occur. The primary goal of the migration process is to ensure that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of data collection mode. It is possible to evaluate this by conducting cognitive interviews with subjects from the target population and/or assessing response equivalence between modes.

Furthermore, it is possible to achieve measurement equivalence even if the instructions or item presentation may not be the same as the original mode. In fact, there may be an opportunity to present items or instructions within the instrument more clearly in a specific data collection mode. For example, if the instrument has a skip pattern, the electronic version could show only the items that subjects need to complete; that is, if there are any skips or jumps over items based upon a previous answer, the subject will never see the item. This can avoid completion of a non-relevant item and resulting data entry queries due to conflicting responses. Such enhancements may bring clarity to instrument completion, but leaves the instrument ‘faithful’ to its original intent and meaning.

This section on migration issues builds on the recommendations of the previous ISPOR ePRO Task Force Report (Coons et al., 2009) to provide more detailed guidelines for the “faithful migration” process and to discuss the mixing of data collection modes, particularly the special considerations when mixing paper and electronic modes.

Perform a “Faithful Migration”

Most migrations involve making changes to the PRO instrument that are required due to characteristics of the new mode. A “faithful migration” is conducted carefully to ensure that only necessary changes to the format and instructions are made—item and response content has not changed. A “faithful migration” of an instrument does not need to look exactly like it did originally, but it does need to collect the same data.
The degree of modification is a key consideration in determining the level of evidence needed to evaluate equivalence as presented by Coons et al. (2009), and is a direct result of the migration process. For example, migrating from paper to an IVR presentation has been categorized as a moderate modification (Coons et al, 2009) because the necessary changes are more extensive than with most other data collection modes. Modification includes revisions to the instructions and may include non-substantive changes to wording of questions and responses for effective implementation on the IVR system. These changes, along with the change from visual to auditory cognitive processing, may result in systematic differences in responses between the modes (Coons et al, 2009). However, it should be noted that traditional telephone-based data collection using a live interviewer rather than the recorded scripts of an IVR system would require the same type of changes.

The following are recommended steps to conduct a “faithful migration”:

1. Contact the instrument’s developer/copyright holder to obtain licensing and/or permission to conduct the migration and to determine if there are requirements for the migration process.

Some instrument developers are now restricting which modes are suitable for migration, and in some cases, which vendors are permitted to conduct such migrations. It is essential to obtain all information related to conducting a migration well before beginning the process to ensure the appropriate procedures are followed. In cases where an instrument is in the public domain or has no identifiable copyright holder or licensor, the sponsor and ePRO vendor must proceed using their best judgment to conduct a “faithful” migration without instrument developer involvement or approval.

2. Review the original version of the instrument to identify necessary changes that need to be made to suit the new mode.

The most common changes are to revise the instructions to be suitable to the new mode and to present one item at a time on a handheld device rather than multiple items per page or in a grid format as is often done on paper. For example, paper questionnaires often instruct the subject to circle a response or to check/tick a box to choose his or her response. Either type of wording will be confusing on an electronic device, so the instructions are reworded to read “select a response”, which can be understood across most modes of data collection. Moving from a grid format on paper to a single item per screen presentation may require rewording of items if a format is used in which the item stem begins at the top of the page and the remainder of each item continues as a fragment in the grid below. In such cases, a complete question will need to be formed from each part of the grid, and the responses are typically presented in a vertical scale rather than horizontally as on the grid.
3. Contact instrument developer/copyright holder for approval of wording and formatting changes needed to migrate the instrument to the chosen platform.

The copyright holder’s approval is necessary to ensure that the proposed changes are considered appropriate and do not threaten the validity of the instrument. The copyright holder may have preferred solutions to the changes proposed so that consistency can be maintained across migrations conducted by separate ePRO vendors.

4. Conduct the migration process and generate draft screens or script.

Once approval for wording changes is obtained, draft screens or a draft script for IVR should be generated to show how the migrated instrument will look or sound in the new mode. The following recommendations help to ensure a “faithful migration” takes place:

- Retain exact wording of item (questions/stem and responses) where possible
- Retain the order of response options
- Keep question and response options together on the same screen
- Evaluate the need for instructions on the same screen or different screens due to space constraints
- Consider aesthetic elements such as spacing between question and responses, space between response options, and equal spacing of response options to reduce bias and improve usability

Along with ensuring the format is correct in the migrated version, the ease of use on the part of the subject should also be considered during migration. For screen-based devices, this includes ensuring that the font size and screen resolution allow the text to be easily readable to study subjects of all ages, while balancing the likely space constraints of longer translations in future international use. Navigation should be easy and intuitive, with clear buttons to indicate how to move forward and back if applicable. Error messages should be considered, if the item is subject to validation rules as in the case of numeric entries that may be above or below a certain range. Usability testing in patient and non-patient populations can help refine the usability of electronic systems.

5. Send draft screens or IVR script to instrument developer/copyright holder for review and approval.

At this stage, collaboration between the instrument developer/copyright holder and the ePRO vendor is critical to ensure that the migration has been done faithfully and that the new format of the instrument appears to retain the same intent and meaning of the original. There may be several iterations of draft screens/scripts and rounds of review to meet the needs of the instrument developer for the migrated format.

The final stages of ePRO implementation include developing the requirements documentation, programming in the mode of data capture and conducting user acceptance testing in the final platform prior to implementation in usability studies or in the eventual clinical trial. User acceptance testing ensures that the system is operating according to the specifications. However, further testing, either usability testing or feasibility testing (see below) will still need to be conducted to ensure that the specific intended target population can use the system to complete the migrated instrument.

See Appendix A for a checklist of steps in the migration process.

Further cognitive interviewing, usability testing, and/or equivalence testing, may be required to confirm that the migration has been faithful and the new implementation is capturing the same kind of data as the original. These are addressed in later sections of this report.

Mode Specific Considerations for Migration

In addition to the process outlined above, each mode has specific considerations that must be addressed during the migration process. The four most common modes will be addressed here.

Migration to a Smartphone/Handheld Device

The main factor for consideration in migrating to a smartphone/handheld device is the space constraints of the smaller screen. Regardless of how the instrument was originally formatted on paper, the default on a handheld device is one item per screen with all responses visible on the same screen. In some cases, a long item with long response options will pose a great challenge as it is not possible to fit all of the text of the question and responses on the screen at the same time. Some solutions are to display the item on the first screen and responses on the second screen with a reminder of what the question was asking, or to present partial responses on a line with a popup that displays the entire response for clarification.

Another possibility is to allow scrolling on the screen to accommodate longer questions or to view response options that do not fit on the small screen. Scrolling is not ideal as it greatly increases the risk that subjects will not be aware that it is necessary to scroll to view the missing text or response options. It also increases the risk of subjects interpreting the question differently or answering differently.

As previously mentioned, migration from some paper-based questionnaires will require the rejoining of split item stems, as illustrated in Figure 1. In addition, due to space constraints of the screen, response options must be displayed vertically instead of horizontally as in the original presentation. Nevertheless, the possibility of utilizing a landscape format in order to take advantage of wider screen real estate should not be ruled out. Figure 2 also illustrates a case where the recall period “past 4 weeks” is not bolded on the
handheld device format although it was bolded on paper, as some platforms are not able to render this type of formatting in the electronic version.

Figure 1. Example of a Paper Format

<table>
<thead>
<tr>
<th>How much of the time during the past 4 weeks . . .</th>
<th>None of the time</th>
<th>A little of the time</th>
<th>Some of the time</th>
<th>Much of the time</th>
<th>All of the time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Did you feel worn out?</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 2. Example of Handheld Device Format

1. How much of the time during the past 4 weeks did you feel worn out?
- None of the time
- A little of the time
- Some of the time
- Much of the time
- All of the time

Migration to a Tablet Device

Due to the larger screen area available, a decision needs to be made in consultation with the instrument developer regarding whether to present one item per screen or to present multiple items per screen on the tablet. With larger size tablets it may be possible to display the entire page as it appeared on the paper version, with radio buttons or checkboxes for responding. A grid format can be retained on the tablet, although it may not be possible to retain all items on the screen as on paper.

It is important to remember that a “faithful migration” does not necessarily mean that the electronic version must look the exact same as the paper version. It means that all of the language from the paper version has been retained, and the migration does not impact the way in which the subject enters the data.

Presenting multiple items per screen may save time with a very long questionnaire, but it also runs the risk of missing data or confusion. It may not be clear to the subject that multiple responses need to be selected on the same screen. Moreover, if functionality is not programmed to prevent skipping questions or advancing to the next screen without completing all questions, the risk of missing data is increased.

On the other hand, a tablet presentation allows for larger fonts and more space to display text, so there are fewer concerns over fitting the instructions, question and responses on the same screen. However, it
should be noted that although space may be available for multiple items per screen on a tablet, a single item per screen can provide consistency across multiple screen-based migrations of the instrument. Consistent presentation, i.e., less variation in the presentation of an instrument’s items across different size screen-based devices, is optimal and a recommendation of this task force, especially if more than one data collection mode is being used in a clinical trial.

**Migration to a Web-based Format**

While Web-based instruments appear to be device independent, there are still constraints contingent on the type of device used to access the Web interface. Due to the wide range of browsers, devices, and screen sizes, a decision needs to be made whether to allow certain types of browsers, such as those for mobile devices, to access the instrument. The screen design for a Web-based instrument intended to be viewed on a desktop or laptop computer would be more similar to a tablet design due to the assumption of more screen space available. However, the screen design for a mobile device like an iPhone would be closer to a smartphone implementation and would require one item per screen formatting.

Due to the wide range of screen sizes and formats with Web-based instruments, there may be greater risks of differences in interpretation of scales. For example, the VAS is scored by measuring the point selected on a line. Line length will vary greatly by screen size and browser. This could lead to different responses in relation to the overall length of the line. As stated previously, it is best to use the single item per screen format because it provides the potential for less variation in presentation across different screen sizes.

**Migration to an IVR System**

For IVR systems, subjects respond to recorded scripts by using, primarily, zero through nine on the telephone keypad. For the most part, migration considerations for IVR have to do with the manner in which the item text (stem) or response options are formatted. For example, if the item is in the form of a statement, it may make more sense to rephrase it in the form of a question: “I feel tired” becomes “Did you feel tired?”

With regard to the response options, if a verbal rating scale is used, the responses must be associated with numeric entries and these must be incorporated into the IVR script for the instrument. In the case of a numeric rating scale that has each end of the scale anchored by a descriptor (e.g., 0= “None” and 10= “Worst imaginable”), but no descriptors in between, the script needs to describe that response context clearly. For example, such items are often worded as follows: Use a scale from zero to ten where zero means none and ten means worst imaginable.

In addition, a traditional visual analog scale, with verbal anchors on each end and no demarcations or descriptors at interim points, cannot be effectively operationalized on an IVR system without changing it to a numerical rating scale. On a technical level, if responses require pressing two or more numbers on the keypad (e.g., 10), the script should confirm with the subject the intended response since one of the numbers
may not have been recorded by the system (e.g., only “1” was recorded rather than the intended “10” due to insufficient pressure on the keypad).

Usability vs. Feasibility

When performing a migration from one data collection mode to another, establishing the subject’s ability to use the new mode, or usability testing, is an important component of the migration. Coons et al. (2009, p 423) stated: “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.” Usability testing is an indication of the subject’s ability to navigate or use a particular data collection system. Since it is focused on the respondent’s ability to use the system, it may be conducted at an investigative site, in a controlled environment with observation of the subject.

While usability testing is always recommended to establish subjects’ (or end users’) ability to use the system, feasibility testing, or the evaluation of the system within a specific study design, may only be necessary in certain circumstances. The distinction between these two types of testing is best characterized as follows: usability testing assesses whether the data collection mode can work under general conditions, whereas feasibility testing assesses whether it will work in the context of a specific study design or a specific instrument.

The need for feasibility testing will be driven by the novelty of the study design in which the PRO data collection system is to be implemented. For example, if the system is to be implemented for site-based PRO data collection in a standard study design that has been previously implemented in numerous trials, there may be no need for additional feasibility testing. As a counter example, if the system is to be implemented in a novel study design, where field-based data is being collected in a unique way (e.g., multiple times per day) for a given patient population, then feasibility testing will ensure that that PRO data collection system actually works with the patients in the study design and using the new instrument. Thus, the evaluation of whether or not to conduct feasibility testing, in addition to usability testing, will be case-by-case and driven by the novelty of the study design and the instrument.

If feasibility testing is deemed as necessary, the testing plan should include: recruiting subjects similar to those who will participate in the clinical trials; subjects following the study procedures as required by the study design for a reasonable period of time (e.g., using the diary for 7 days); and then performing debriefing interviews with the subjects to assess their compliance with the study procedures (e.g., if they completed the diary every day, as requested) as well as to assess usability. The debriefing of the subjects is best facilitated by review of actual compliance data captured during the study and likely reported on a portal of some kind.
It is important to note that neither usability testing nor feasibility testing as described above is the same as another process called user acceptance testing (UAT). According to Coons et al. (2009, p. 424), “the purpose of UAT is to determine whether the software complies with the written system specification or user requirements document.” It is not intended to determine if respondents like or can use the system. UAT does not include clinical study subjects. We recommend that usability testing and, if necessary, feasibility testing, occur in addition to user acceptance testing, following a migration to a data collection mode.

VI. Equivalence

Any migration process involves some type of modification(s) to implement the instrument in the new mode. The goal of the “faithful migration” is that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of data collection mode. Once a migration has occurred, it is necessary to determine if this goal was achieved through an evaluation of the measurement equivalence between the original and migrated modes. The previous ISPOR Task Force Report focused on the degree of modification as the key factor in determining the level of evidence needed to establish equivalence. In this section, we build upon the work of that previous task force and recommend additional considerations for determining the level of evidence needed to establish equivalence. We also delineate the types of equivalence testing and the typical procedures to execute such work.

Need to Establish Equivalence

One consideration is whether measurement equivalence needs to be established between the original and new data collection mode. In the context of instruments that will be used in registration trials for submission to the FDA, measurement equivalence must be established and documented by the study sponsor if the data are to be used to support labeling for a medical product. If there are sufficiently rigorous published data to support that equivalence, then further equivalence studies are not needed. A decision tree is shown below in Figure 3. It is worth noting that from a scientific perspective, we believe that it is always necessary to have confidence, through evidence, that measurement equivalence exists because this has a direct impact on interpreting any results from a migrated instrument (represented by always following the right hand side of Figure 3).

---

2 UAT is one aspect of an extensive system/software validation process that is far beyond the scope of this manuscript. Another ISPOR PRO Task Force Report, Validation of Electronic Systems to Capture Patient Reported Outcomes (PRO) Data - Recommendations for Clinical Trial Teams: A report of the ISPOR ePRO Systems Validation Task Force addresses this topic.
Figure 3. Decision-Tree regarding Need to Establish Measurement Equivalence

Levels of Equivalence Evaluation

Following the “faithful migration” and the determination that equivalence needs to be established, the appropriate level of equivalence evidence needs to be identified. The level of equivalence evidence is dependent on the extent that the changes or modifications are likely to have had an effect on the subjects’ interpretation and responses to the items in the instrument.

Table 1 below summarizes the levels of modification that might occur during “faithful migration,” and is an adapted version of the one presented by Coons et al. (2009). In the course of performing the migration, two types of changes may occur and need to be considered when equivalence is evaluated: format and procedural. Format changes refer to differences between the modes in terms of format, including how the items and responses are presented to the subject. For example, formatting modifications include adapting instructions from a paper to an electronic mode, such as changing ‘circle’ to ‘select.’ Procedural changes refer to the different ways modes are actually implemented in studies and include aspects such as edit or validation checks, introducing a jump or skip sequence so subjects do not see questions that are not relevant to them, completion windows, and compliance with protocol requirements such as when to complete data collection. In general, the procedural changes between modes may have a greater impact on how the subject responds to questions because the electronic modes can limit possible responses to those
that are within an appropriate range, prevent unintentionally skipped questions, and enforce completion windows; none of these are/were possible with paper.

The breadth and extent of the modifications made during the migration process, some of which are necessary and others that facilitate easier administration of the instrument, will vary in terms of their impact on influencing the subject’s interpretation and responses. Table 1 incorporates examples of minor and moderate levels of procedural as well as format modifications to illustrate how both types of changes can be evaluated in terms of levels of equivalence. Minor modifications have a low likelihood of affecting interpretation and response, and therefore cognitive interviews and usability testing are sufficient to confirm the equivalence between the modes. Moderate modifications introduce the possibility of affecting interpretation and response, and therefore it is recommended that a quantitative equivalence study along with usability testing be conducted to evaluate equivalence between the modes.

It is critical to determine the level of equivalence needed as part of the equivalence study planning process. The table below is based upon Coons et al. (2009), and additional detail has been added to illustrate what constitutes minor and moderate modifications. In cases where modifications fall into more than one level, the recommendations associated with the higher level of modification and, therefore, evidence should be followed. It must be noted that if substantial changes to the item content and/or response options are needed to enable migration of the instrument to a new mode of data collection, the instrument is considered a new instrument and full psychometric evaluation would be necessary.

**Table 1. Levels of Modification and Equivalence**

<table>
<thead>
<tr>
<th>Level of Modification</th>
<th>Rationale</th>
<th>Examples</th>
<th>Level of Evidence</th>
</tr>
</thead>
</table>
| Minor                 | The changes to instrument are *not likely* to have changed interpretation or responses. | Format:  
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 Interviewing Usability testing |
<table>
<thead>
<tr>
<th>Procedural:</th>
<th>Format:</th>
<th>Equivalence Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Implementation of tablet at the site with differences in edit checks, validation rules, branching logic</td>
<td>1) Changes in item wording or more significant changes in presentation that might alter interpretability. (e.g., splitting an item over two screens, changing the structure of the response options.)</td>
<td>Usability testing</td>
</tr>
<tr>
<td>Moderate</td>
<td>2) Change in mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]).</td>
<td></td>
</tr>
<tr>
<td>3) Change in mode of administration to Web-based administration (e.g., variance between screen sizes too great to be considered minor modification.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedural:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) Migration of paper diary to electronic platform with differences in completion windows, compliance with administration recall period primarily and edit checks, validation rules, branching logic secondarily.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) Differences in the ways that subjects are alerted to complete instruments (e.g., alerts on a</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Types of Measurement Equivalence Studies

Two major types of studies may be conducted to evaluate measurement equivalence between modes:

1) Qualitative studies involve cognitive interviews that provide qualitative data to evaluate equivalence between modes. These studies have previously been associated with minor degree of modification between modes.

2) Quantitative studies are intended to evaluate statistical equivalence of responses, and involve much larger sample sizes and focus only on statistical comparison of responses to both modes of the PRO measure.

Equivalence Study Descriptions

Common Qualitative Study Designs

Qualitative studies involve small samples of 10 to 15 participants who are from the target population of the confirmatory clinical trial, usually Phase 3. Qualitative study designs are used to evaluate the impact of format changes between the original mode and the migrated mode to ensure that the subject’s interpretation of the items on the migrated mode is comparable to the original.

Cognitive Interviews during the Instrument Migration Process

It bears pointing out that cognitive interviews conducted during the instrument migration process serve a different purpose than cognitive interviews conducted during the instrument development process. In general, cognitive interviewing techniques are used to study the way in which subjects “understand, mentally process, and respond” to materials presented to them (Willis 2005, p 3). Of interest here are the cognitive interviews conducted subsequent to instrument migration, which are aimed at determining if subjects are interpreting and responding to the items the same way on the new mode as they would on the mode from which the instrument was migrated. See below for different cognitive interviewing approaches. In contrast, cognitive interviewing during instrument development is primarily aimed at supporting the instrument’s content validity by determining if subjects are interpreting items and using response scales as
intended. In the current context, cognitive interviews are not intended to revisit the content validity of the original instrument.

One approach to conducting cognitive interviews involves having subjects complete the instrument on the original mode and new mode of data collection and determining if there are items for which the responses differ between the two modes. A distraction task can be included between completion of the modes to reduce potential memory/carryover effects yet allow a short interval between administrations to reduce subject burden. The interview then focuses on those items individually to determine if the different responses were random (i.e., “I could go either way”) or systematic due to a difference in the meaning or interpretation of the item by the subject on the alternative modes.

If the latter is the case with a substantial number of subjects in the cognitive interview sample, the changes made in migrating those items to the new mode need to be re-visited to determine if a successful migration of those items is possible. It should be noted that, other than using the responses on the two modes to identify where differences exist, this approach is not quantitative; the responses are not used for any descriptive or inferential statistical analyses. Further, if such discrepancies between modes occur and changes are made, the cognitive interviewing must be replicated. If there are still discrepancies in subject qualitative reporting due to format differences, a quantitative equivalence study should be considered (see below).

A second approach involves having subjects complete the instrument on the new mode and asking them how they interpret what each item is asking them. This can be accomplished by asking the subject to repeat the question being asked in their own words (i.e., paraphrasing) or through a think-aloud task which involves the subject talking through how he or she arrives at the response (Willis 2005). The subject’s interpretation of the item is then compared to the item definition or concept elaboration document prepared by the instrument developer to determine if there is concordance. This approach more closely parallels cognitive interviewing during the instrument development process. It assumes that documentation of the intended meaning/interpretation of the items is available. If the instrument had been translated for use in other languages/cultures, such documentation should exist since it is essential for linguistic validation. If it does not exist, it should be able to be constructed in conjunction with the instrument’s developer.

A third approach is to ask subjects only about instructions and/or items that were modified during the migration process to reduce subject burden and interview length. This enables a more focused investigation of the potential impact of those changes and a potentially shorter interview. Subjects are asked to read both versions of the instructions or items on the two modes and identify any perceived differences in the self-report task or in the interpretation/meaning of modified items. However, if most or all of an instrument’s items required modification during the migration process, then this approach does not necessarily decrease the amount of time required to conduct the cognitive interview because all such items would still require debriefing.
At the present time there is no consensus regarding the optimal approach to cognitive interviewing during the migration process. A combination or hybrid of two or more of the above is a viable option if it makes sense for a particular study.

**Common Quantitative Study Designs**

Quantitative equivalence studies are recommended for moderate modifications between the modes (see Table 1) when migrating and for mixing modes that involve visual vs. auditory use (IVR), use of Web at subjects’ homes, and for paper vs. electronic diary studies. All of these scenarios present greater risks for differences in response between modes and therefore a greater need to demonstrate that they provide sufficiently equivalent results.

Within quantitative equivalence study types, Coons et al. (2009) mentions randomized cross-over and randomized parallel groups as the typical options for evaluating equivalence. Randomized cross-over designs have become the preferred study design for migration equivalence studies because subjects serve as their own controls and therefore the sample size is significantly reduced. Within the randomized cross-over approach, study designs may be either single visit when evaluating if the migration changed interpretation or multi-visit when evaluating if migration changed interpretation and how items were completed in the context of implementation. The multi-visit study design is most useful for evaluating field-based assessments which are intended to be completed on a daily basis over a period of time and scores are typically averaged.

Figure 4 depicts the most common study designs for quantitative equivalence studies. The single visit study design is appropriate for a site-based or field-based assessments and involves a randomized cross-over in which each study subject completes both modes of data capture but is randomized to the order of completion (i.e., randomized to which mode is completed first to control for order effects).

The multi-visit study field evaluation design is appropriate for diary or field based instruments, if there is a need to establish that the two modes of data collection are equivalent in the context of a simulated study design. The multi-visit study involves a randomized cross-over in which each subject is randomized to order and completes the first mode for one to two weeks and then crosses over to complete the second mode for the same length of time. This approach is recommended in cases where the two modes are intended to be mixed in future studies.

Due to the longer study duration, it is important to also assess whether the subject’s condition has changed to ensure that the comparisons made are within those subjects who have not changed during the course of the study. This approach allows for procedural differences between the modes to be tested in setting similar to what the subject would experience in the clinical trial. However it has an increased risk of demonstrating a lack of equivalence because of the potential for larger response differences between the two modes due to the manner in which the subjects are completing data entry.
For example, if a protocol has subjects entering data within a specific time window per day, data entry with an electronic data collection device may be confined to that time window whereas with a paper-based data collection, data may be entered by the patients at any time. (This study design functionally includes the concept of feasibility, examining performance of the system in the context of a specific study design, as discussed above. Therefore, the situations in which a multi-visit equivalence study design is incorporated should be carefully considered.)

Single visit studies answer a different question from multi-visit study designs. Single visit studies focus on equivalence in interpretation at a point in time, which is sufficient when moving away from the paper-based data collection, whether at the site or in the field. Multi-visit studies are needed to address equivalence between modes in a field-based context of a specific study design, and are needed if intending to mix modes in the future. More specifically, if one intends on mixing paper and electronic diaries in a trial, the multi-visit feasibility study is needed to establish equivalence of these two modes of PRO data collection in a real-life setting, given the procedural differences between the two modes. (Again, this will also functionally accomplish the goal of feasibility testing of the electronic data collection mode.)

In either case, similar statistical methods such as those proposed in Coons et al. (2009) and McEntegart (2010), are used to evaluate the equivalence between responses.
Common Cross-Over Equivalence Study Designs

**Single Visit**
- Site- or Field-based Assessment
- (1 visit; N=60)

**Multi-Visit Field Evaluation**
- Field-based Assessment
- (3 visits; N=60)

**Single Visit**
- Randomized to order
- Complete both modes within same visit session
- Distraction task in between
- Time between completions varies
  - Few minutes – 2 hours
- Results are compared statistically

**Multi-Visit Field Evaluation**
- Randomized to order
- Visit 1: Provide 1\textsuperscript{st} mode, training if ePRO
- 1\textsuperscript{st} mode completed between visit 1 and 2\textsuperscript{nd} visit
- Visit 2: Provide 2\textsuperscript{nd} mode
- 2\textsuperscript{nd} mode completed between visit 2 and 3
- Time between visits varies
  - 1 week – 2 weeks
- Results are compared statistically
Table 2 displays the two common study design types and some considerations for each.

<table>
<thead>
<tr>
<th>Instrument Type</th>
<th>Study design type</th>
<th>Pros</th>
<th>Cons</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRO instruments completed at site; Field-based assessments where mixing is not intended</td>
<td>Single Visit – randomized cross-over</td>
<td>Statistical equivalence level between modes can be established</td>
<td>Assesses format differences but not procedural differences</td>
<td>Comparison with original mode test-retest reliability may be limited; doesn't reflect performance of paper diary in clinical trial setting</td>
</tr>
<tr>
<td>Field-based assessments, especially frequent or episodic assessments per day, where mixing is intended although not recommended</td>
<td>Multi-visit field evaluation randomized cross-over</td>
<td>Statistical equivalence level between modes can be established; real world setting for diary</td>
<td>Studies difficult to operationalize because target concepts are variable, need to control for change; high likelihood that equivalence won't be found</td>
<td>Comparison with original mode test-retest reliability may be limited;</td>
</tr>
</tbody>
</table>

In cases where little evidence of the test-retest reliability of the original version of the instrument is available, it may be helpful to conduct a “double cross” study in which each subject crosses between modes and then back to the original mode so that test-retest reliability can be obtained and compared for both within and between modes. See Appendix B for details on the “double cross” study design.

When planning subject recruitment for qualitative and/or quantitative equivalence studies, it is important to consider potential overlap with recruitment for clinical trials with the same patient population, especially in rare disease where the population for trials is limited. It is acceptable to recruit subjects for equivalence studies who may then go on to participate in a clinical trial in which the modes tested in the equivalence studies are to be used, in order not to reduce the pool of potential clinical trial participants. Unless it is absolutely necessary, it is not recommended to recruit participants for equivalence studies who have already participated in a clinical trial or validation study using one of the modes in question because they have already experienced the mode and PRO instrument being studied and may have a biased response during the equivalence study.

Qualitative study designs are acceptable for demonstrating measurement equivalence for minor modifications and for migrations in which the original and alternative data collection mode are not intended.
to be combined in clinical trials. These studies do not statistically test measurement equivalence for mixed modes and are insufficient for mixed paper and electronic field-based assessment (e.g., daily diary) to be used within the same trial. Within quantitative study designs, if a field-based assessment is tested in a clinic-based single visit design, it does not reflect the actual trial setting and is unlikely to assess the true performance of the instrument.

It is also critical for field-based assessment studies that the subject population is stable and unchanging to limit true change in response in equivalence studies, but clinical trial use assumes that subjects will change over time due to the treatment. Therefore it may be impossible to distinguish what is driving change in scores when mixed paper and electronic field-based assessments are used in a treatment setting. The result of the equivalence studies may be to conclude that the potential differences between paper and electronic field-based assessments are too great to allow mixing modes within a clinical trial, and in these cases the default should be the electronic data collection mode only.

As the term migration in and of itself merely refers to the transfer of an instrument from one mode or format to another, it carries no implication of what will be done with either mode in the future. In many cases, the migration results in a new mode that will replace the original mode in future studies, while mixing involves using both old and new modes within or between studies and then pooling the data from different modes for analysis. Therefore, when migrating permanently it is only necessary to demonstrate equivalence for prospective use, while when migration results in mixing modes, it is necessary to demonstrate equivalence for concurrent use. In the former case, a qualitative study may suffice, while in the latter case it is necessary to conduct a quantitative equivalence study. Therefore, it is strongly recommended that the potential for mixed modes be considered at the start of the equivalence study planning process so that the appropriate approach to evaluating equivalence, qualitative or quantitative, will be used in the most timely and efficient way.

VII. Mixing

While mixed modes can and do occur in all research settings, the primary focus of this paper is on clinical trials in which the PRO endpoints are intended to support labeling claims. The discussion thus far regarding mixing modes has focused on mixing within a given trial. The general recommendation is to avoid, where possible, such mixing due to the increase in measurement error associated with introducing any variable into a study.

We note that mixing, as used in this report, refers to the administration of the same instrument via different data collection modes in a single clinical trial; it does not refer to the administration of different instruments via different data collection modes in a single clinical trial. The latter does not pose a threat to measurement error as discussed in this report.
Our recommendations thus far for mixing have focused on determining the need to establish measurement equivalence when mixing occurs. It could be argued that randomization of subjects into groups is sufficient to account for mixing. As long as the pattern of mixing modes is the same in the treatment and control groups, any potential measurement error introduced by the mixed modes will be comparable across the two groups. However, even the balanced introduction of measurement error across treatment arms has the potential to put the trial at risk of not showing a treatment effect if the signal to noise ratio is decreased. Any change during the trial (after randomization) that leads to different data capture mode patterns across the treatment and control patients (or within treatment or control patients) has the potential to differentially introduce measurement error.

There are a number of ways in which mixing of modes can occur in development of medical products, including mixing:

1) Between product development programs
2) Between clinical trials within a program
3) Within a single clinical trial, such as
   a) countries within a trial
   b) sites within a country
   c) subjects within a site
   d) within a subject
   e) timepoints within a trial (e.g., start with one mode and change to another mode).

We now turn to a discussion of the various ways in which modes might be mixed.

Mixed modes occurring between medical product development programs or clinical trials within a program are often the result of evolving technology. New and better methods of PRO data collection emerge or regulatory requirements change and modes may change during the product development program. The implementation of mixed modes in these situations can be carefully planned and executed with supporting studies to demonstrate the equivalence of modes.

In this manner, if the old or original mode is abandoned, a demonstration of measurement equivalence supports consistency of interpretation of the data between trials, and no additional activities are needed in new trials because only a single mode of PRO data collection is being incorporated. In some development programs, there may not be a need to compare current trials to previous ones with respect to the PRO data. If that is the case, then there is no need to establish measurement equivalence between the new mode, and
the one used in previous trials. The need for measurement equivalence will be driven by the need to compare data across trials within the program.

The remaining types of mixing are within a given clinical trial. The first that we will consider is mixing modes across countries within a trial. When conducting multi-country studies, not all countries may have access to the technology being implemented for PRO data collection (e.g., Internet access for Web-based version). In such cases, specific countries within multi-country studies may need to collect PRO data using one mode while the remaining countries use another. Measurement equivalence should be demonstrated across mode. Historically, researchers considered mixing technology-based solutions with paper-based solutions if a specific country is not able to support the selected technology.

However, with careful planning, the modes that are mixed may have only minor differences (e.g., both are screen-based systems), and thus, a lower likelihood of introducing measurement error resulting from measurement inequivalence. For example, if a Web-based PRO data collection system is the default for the trial, perhaps a handheld or tablet can be used for countries that do not have internet access that can support the Web-based system.

Within a country, modes can also be mixed between participating clinical sites within a trial, again, because of access to the specific technology, or possibly the site’s ability. If the issue is site ability, then the potential mixing of electronic solutions becomes more challenging. The likely case is that the investigative site does not believe they can implement the technology-based solution. It may be possible to have more similar technology solutions that a site can implement, which, again, will minimize the potential for introducing measurement error.

Alternatively, the sponsor may choose to mix paper and electronic solutions. If the PRO data will be field-based assessment, our general recommendation still prevails; it is unlikely that equivalence can be established. In such cases, it may be prudent for the sponsor to consider other options such as not including the specific site or region in question. (This also applies to the country case described above.) In contexts where the subject sample is extremely difficult to recruit, such as for rare diseases, the sponsor may be faced with a significant dilemma between mixing modes and increasing measurement error, versus obtaining the subjects for the trials. In such cases, the sponsor will need to make this decision based upon the specific issues facing them.

Subjects within a given investigative site may have the need for various modes of PRO data collection because of subject ability, preference, state of health, or site preference for a given subject. Such decisions to mix subjects within a given site should have been anticipated and planned for in a similar manner to the above cases of mixing across countries or sites. If mixing of modes within a site is anticipated, then appropriate upfront equivalence should be established.
The additional challenge, however, is deciding that a subject needs one mode or another. We can anticipate that in these cases, the request will be for a familiar mode of PRO data collection, likely paper. If a potential subject has never previously used an electronic method of data collection, then his or her initial preference may be to select paper, if that is the other option presented in the trial. Such subject preference would not be based upon actual ability to utilize the electronic solution, but a subject’s impression or belief. We recommend that rather than letting the subjects decide, a more objective method of evaluating individual subject’s ability be used if a sponsor wants to provide options to subjects. Such an evaluation can take the form of having the subject attempt to use the electronic method at the investigative site.

If this evaluation demonstrates that the subject cannot use the primary electronic solution, possibly because of subject ability or state of health, then the investigative site should establish that the subject can use the alternative mode. It should be emphasized to the investigative site that this evaluation is mandatory so that the site does not attempt to use its own, idiosyncratic evaluation of subject ability/competence to make this decision. We recommend that mixing modes of PRO data collection within an investigative site be used rarely and approached cautiously.

In the final situation for consideration, subjects may begin a study using one PRO data collection and finish with another. Some trials have built in a paper back-up solution for situations where the technology fails, given that one potential issue for any electronic PRO data collection system is failure of the technology. Thus, some subjects may begin in one mode but switch to another as a backup in cases of device loss or failure, or inability to access the electronic version. In such situations, it will be important to note where such a switch took place within the duration of the trial.

Further, more recent advances in both technology, as well as delivery of technology, may facilitate a more reliable technology back-up (e.g., a Web-based system as a temporary replacement for a stand-alone system), or quick recovery back to the same technology. The primary issue is one of the potential for missing data versus the introduction of measurement error through mixed modes of data collection. The sponsor will need to consider a back-up solution, and what solution most appropriately balances those two considerations. The sponsor may decide that a backup solution that minimizes missing data takes priority regardless of the nature of the backup. In addition, a low level of mixing of modes (e.g., <10%) may not have an impact on the overall result, but sensitivity analyses are recommended to verify whether it did or not, especially if equivalence has not been shown a priori.

Any one of these types of mixing can yield differences in the data and introduce measurement error into the trial or clinical program results. We therefore recommend that the need for data comparability and the impact of introducing measurement error be assessed for each situation. Our recommendation is that if data are to be compared or pooled at any level, quantitative evidence of equivalence is necessary. Table 3 provides a presentation of the risk of not having equivalence at each level at which mixing may occur.
Table 3. Mixing data collection modes and risk of not having equivalence

<table>
<thead>
<tr>
<th>Level of Mixing</th>
<th>Risk to Equivalence</th>
<th>Comments</th>
</tr>
</thead>
</table>
| Between product development programs         | Varies              | If there is no need to compare or pool current results with prior products developed, then risk is low. If the new product is in the same therapeutic area, the risk may become higher.  
Analytical techniques can be used to evaluate any error introduced due to mixing; see Post Trial section. |
| Clinical trials within a program             | Varies              | The risk will vary based upon stage of product development and the need for trial comparability or pooling.  
Analytical techniques can be used to evaluate any error introduced due to mixing; see Post Trial section. |
<p>| Countries within a clinical trial            | High                | If data are to be pooled across countries, as with most trials, comparability is very important and differences between countries should be evaluated. |
| Sites within a trial                         | High                | Difficult to establish between site comparability of data if mixing occurs, therefore it is                                             |</p>
<table>
<thead>
<tr>
<th>Subjects within a site</th>
<th>Very High</th>
<th>It is very difficult to assess a site’s performance if the data are collected with different modes by subjects at the same site. It may be very difficult to determine if changes are due to response to treatment or to difference in mode. This is not recommended.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within a subject</td>
<td>Extremely High</td>
<td>Can potentially compromise usability of subject data. Difficult to demonstrate that mixing did not have impact. Strongly discourage this level of mixing except in extreme cases in which data would otherwise be missing. This is not recommended.</td>
</tr>
</tbody>
</table>

Table 3 above suggests situations that vary in risk when using mixed modes for measurement of PRO endpoint(s) between and within a clinical trial. This could be the difference between success and failure for the trial if it is the primary endpoint. As mentioned in the introduction, not having measurement equivalence between the modes could increase measurement error, in turn attenuating the ability to identify a treatment effect within a given trial. Such measurement error could then result in a non-significant difference in the primary efficacy PRO endpoint for a new medical product.

**Types of Modes Being Mixed**

Once it is determined that mixing will be done either between or within a trial, there are considerations for which types of modes to mix. We now turn to issues to consider with mixing various types of modes of PRO data collection across the situations just described. This includes mixing paper and electronic modes and different electronic modes of data collection.
Mixing Paper and Electronic

Mixing paper and electronic modes is the most risky combination because of the differences in how a subject interacts with paper, having little to no restriction on how he or she responds to questions, in comparison with electronic modes, which verify responses via edit checks and restrict the subject in how they may respond to questions. Our general recommendation is to avoid mixing paper and electronic modes of data collection to the extent possible. There is less risk in mixing site-based instruments because they are completed under supervision and corrections to invalid responses can be made, and if equivalence between these modes has been previously demonstrated.

Empirical evidence is emerging to demonstrate moderate to good correlations between paper and Web-based data collection for site-based instruments (Bennett 2013a; Bennett 2013b). We strongly discourage the mixing of paper and electronic field-based assessments because of the significant potential equivalence issues, the significant procedural change between these two modes, and the likelihood they will not generate equivalent responses.

The FDA clearly discourages field-based PRO data collection using paper because of the inability to know when the data are entered. Specifically, the PRO Guidance (2009, p. 14) states “If a patient diary or some other form of unsupervised data entry is used, we plan to review the clinical trial protocol to determine what steps are taken to ensure that patients make entries according to the clinical trial design and not, for example, just before a clinic visit when their reports will be collected.” This quote specifically addresses the ill-advised use of paper as the single data collection mode, but it also underscores the impracticality of mixing paper and electronic diaries in a clinical trial setting.

Mixing Electronic Modes

Mixing visual modes, such as tablet with the Web or a smartphone device, is less risky because it is potentially easier to demonstrate equivalence between these modes and implement them in a similar fashion so that differences in format are minimized. However, the use of the Web without restrictions on screen size and resolution is potentially risky because it is not possible to control all elements of the visual presentation to ensure that all subjects see the questions and responses the same way.

We recommend caution when using Web-based data collection for this reason and consider it a mixed modes situation because of the degree of differences. Mixing visual and auditory modes, such as Web and IVR, requires that quantitative equivalence be demonstrated to ensure that the moderate difference between modes does not impact interpretation and response. Some studies have demonstrated equivalence between Web and IVR modes (Bennett et al. 2013a; Bennett et al. 2013b). There may also be implementation challenges with such disparate modes that need to be considered.
Operational and Statistical Considerations for Mixing Modes

At this point, once a decision to incorporate mixed modes within a trial, and which modes will be mixed has been made, there are a number of operational and statistical issues to be considered as implementation occurs.

Pre-trial Preparation

As discussed many times in this report, it is important in most situations to first evaluate equivalence between the modes. It is strongly recommended that the potential for mixed modes be considered at the start of the equivalence study planning process so that the appropriate approach to evaluating equivalence, qualitative (for minor and/or permanent migration) or quantitative (moderate and/or mixed modes in the future), will be used in the most timely and efficient way. It is also critical to assess the risks of certain types of mixing as described above. Assuming that measurement equivalence between the modes has been established, the results of this evaluation should be taken into account when determining the sample size for the study. This can be done by working with the appropriate biostatistician to adjust the presumed measurement error in the sample size calculation. The specifics of this computation are beyond the scope of this report.

Additionally, pre-trial planning should include the issues raised above in the discussion of mixing within a trial. That is, appropriate training for both modes will need to be established. Perhaps most important, the criteria for determining which countries, regions, sites, or subjects are permitted to mix will need to be established, documented, and then clearly conveyed to the investigative sites.

Trial Implementation

If mixing was pre-planned, then one of the key implementation challenges will be to manage where and when each mode is used. If the plan was for mixing across countries, regions, or sites, the challenges will be fewer than cases where mixing occurs within a site or within a patient because the data collection mode will be same within the country, region, or site. As mentioned above, the key issue will be to ensure the investigative site follows the sponsor’s procedures for mixing, rather than using idiosyncratic criteria.

Perhaps the most challenging situation will be the case where mixing was not planned, an electronic PRO data collection mode fails, and the sponsor defaults to a paper-based method. Our recommendation is that the sponsor should always have a contingency in case of technology failure. In the absence of such planning, and some determination of the potential impact of mixing in the case of electronic system failure, ad hoc implementation of a back-up may result in significant compromise to the study data. Because of the major risks in this type of ad hoc mixing in studies involving field-based data collection, options other than paper should be considered as a backup in case there is device loss or failure. Finally, whether mixing is planned or happens ad hoc, it will be necessary to develop the statistical analysis plan (SAP) to address the analysis of mixed modes a priori to evaluate if the treatment effect differs by mode.
In cases where there were mixed modes, the analytic methods applied post trial are those used to evaluate poolability of data. Other examples where poolability would be tested include translations (Wild et al., 2009), countries, regions, and sites. If there are sufficient data available per mode, it is recommended to assess the mode as a variable for analysis in order to perform mean comparisons and to test for a treatment by mode interaction. Finally, it may be helpful to conduct a sensitivity analysis to evaluate the impact on the data and treatment effect of including or excluding the alternate mode data, especially in the case of ad hoc mixing where only a small number of subjects or sites used the non-standard mode. All of these statistical techniques can be worked through in detail with a biostatistician.

VIII. Conclusion

One of the most important developments in the field of PRO measurement in clinical trials has been the emergence of technologies that enable electronic collection of data. With the increasing variety of data collection modes, mixing these modes between and within clinical trials in a medical product development program is possible. Although it has the potential to add measurement error if not planned and implemented properly, mixing of PRO data collection modes within trials does occur and must be addressed pragmatically.

This task force report provides an overview of important issues to consider in the process of migrating between modes of data collection and also in using multiple modes of data collection in clinical trials. The key drivers of this report are to address the FDA concern regarding measurement equivalence and the potential impact of using multiple modes of data collection or administration on the treatment effect in the trial and the choice to reduce measurement error where possible when insufficient information about measurement equivalence is available.

It is important to conduct a “faithful migration” from the original mode to a new mode or modes to ensure that subjects interpret and respond to the questions/items on the PRO instrument the same way regardless of data collection mode. The levels of evidence needed to establish equivalence have been expanded to address both format and procedural differences that occur between modes, and an overview of common study designs for both qualitative and quantitative equivalence studies has been presented.

In the absence of documented evidence of measurement equivalence, it is strongly recommended that a quantitative equivalence study be conducted prior to mixing modes in a trial, to ensure that sufficient equivalence can be demonstrated to have confidence in pooling PRO data collected by the different modes. However, we also strongly discourage the mixing of paper and electronic field-based instruments and suggest that only mixing of electronic modes be considered for clinical trials and only after equivalence has been established.
If proceeding with mixing modes, it is important to implement data collection carefully in the trial itself in a planned manner at the country level or higher and minimize ad hoc mixing by sites or individual subjects. Finally, when mixing occurs it must be addressed in the statistical analysis plan for the trial and the ability to pool the data must be evaluated in order to then evaluate treatment effects with mixed modes data. A successful mixed modes trial requires a “faithful migration,” measurement equivalence established between modes, and carefully planned implementation to minimize the risk of increased measurement error.
References


Lundy JJ, Coons SJ, Aaronson NK. Testing the measurement equivalence of paper and interactive voice response system versions of the EORTC QLQ-C30. Qual Life Res. In press. (Published online: June 2013)


## APPENDIX A.

**Checklist of Recommended Steps to Conduct a “Faithful Migration”**

<table>
<thead>
<tr>
<th>“Faithful Migration” Step</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. □ Contact the instrument’s developer/copyright holder to obtain licensing and/or permission to conduct the migration and to determine if there are requirements for the migration process.</td>
<td>Some instrument developers are now restricting which modes are suitable for migration, and in some cases, which vendors are permitted to conduct such migrations. It is essential to obtain all information related to conducting a migration well before beginning the process to ensure the appropriate procedures are followed.</td>
</tr>
<tr>
<td>2. □ Review the original version of the instrument to identify necessary changes that need to be made to suit the new mode.</td>
<td>Common changes:</td>
</tr>
<tr>
<td>3. □ Contact instrument developer/copyright holder for approval of wording and formatting changes needed to migrate the instrument to the chosen platform.</td>
<td>The copyright holder’s approval is necessary to ensure that the proposed changes are considered appropriate and do not threaten the validity of the instrument. The copyright holder may have preferred solutions to the changes proposed so that consistency can be maintained across migrations conducted by separate ePRO vendors.</td>
</tr>
<tr>
<td>4. □ Conduct the migration process and generate draft screens or script.</td>
<td>- Retain exact wording of item (questions/stem and responses) where possible</td>
</tr>
<tr>
<td></td>
<td>- Retain the order of response options</td>
</tr>
<tr>
<td></td>
<td>- Keep question and response options together on the same screen</td>
</tr>
<tr>
<td></td>
<td>- Evaluate need for instructions on the same or different screens due to space constraints</td>
</tr>
<tr>
<td></td>
<td>- Consider aesthetic elements such as spacing between question and responses, space between response options, and equal spacing of response options to reduce bias</td>
</tr>
<tr>
<td></td>
<td>- Ensure font size and screen resolution are legible</td>
</tr>
<tr>
<td></td>
<td>- Keep text length in mind for future translations</td>
</tr>
<tr>
<td></td>
<td>- Use navigation buttons that are easily understood</td>
</tr>
<tr>
<td>5. □ Send draft screens or IVR script to instrument developer/copyright holder for review and approval.</td>
<td>There may be several iterations of draft screens/scripts and rounds of review to meet the needs of the instrument developer for the migrated format. This should be taken in to consideration when developing the timeline.</td>
</tr>
<tr>
<td>6. □ Finalize electronic version on electronic mode.</td>
<td>- Develop the requirements documentation</td>
</tr>
<tr>
<td></td>
<td>- Program in the mode of data capture</td>
</tr>
<tr>
<td></td>
<td>- Conduct user acceptance testing</td>
</tr>
</tbody>
</table>
Appendix B

Experimental Equivalence Study Designs “Double Cross”

In cases where little evidence of the test-retest reliability of the original version of the instrument is available, it may be helpful to conduct a “double cross” study in which each subject crosses between modes and then back to the original mode so that test-retest reliability can be obtained and compared both within and between modes. The reason this is important is that test-retest reliability between the new and old mode cannot reasonably be expected to be higher than test-retest reliability within either the old mode or new mode. In other words, if the double cross shows that test-retest reliability within the old mode is poor, then we would adjust our expectations for the value of any comparison between the new and old modes. Figure 5 depicts single visit and multi-visit options for the “double cross” depending on whether the instrument is site based or field based.