

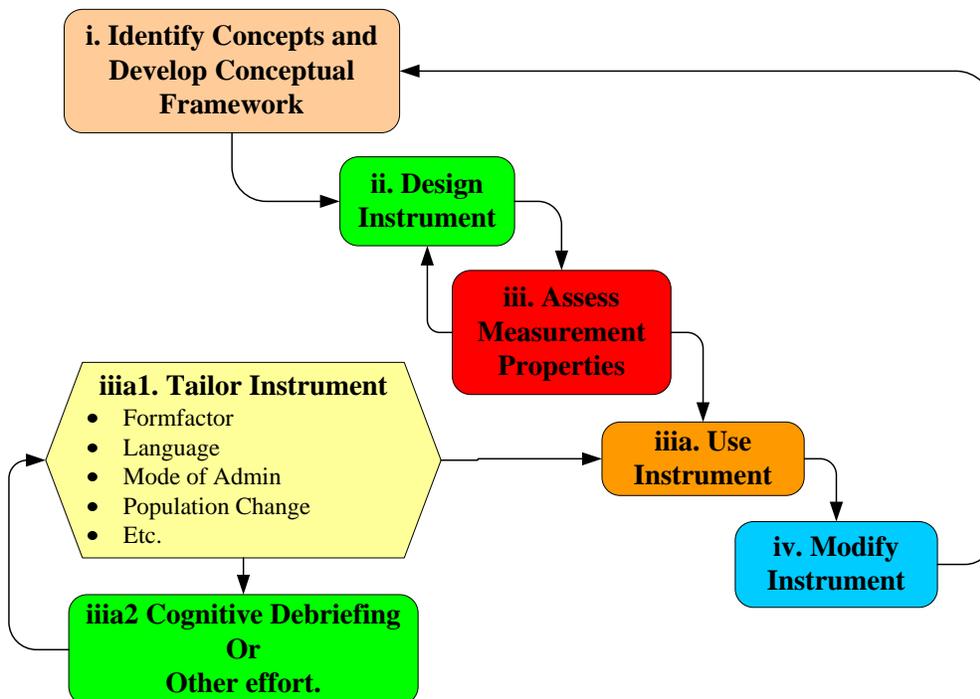
Greg Gogates [greg.gogates@crfhealth.com]  
CRF Inc.  
Suite 140  
1690 Sumneytown Pike  
Lansdale, PA 19446

## Comments to the draft PRO Guidance. Docket #2006D-0044

### Comment #1

Line 191 Figure 1 shows a PRO Instrument development and Modification process. It is shown as a circular lifecycle and while accurate could be more detailed to differentiate the difference between the initial development and the subsequent modifications as noted in section IV.D starting on line #579.

I propose that the lifecycle for development is fine but the modification should follow a TAILORING process as outlined in the SEI Institute <http://www.sei.cmu.edu/about/about.html> . This concept separates the design of an existing product from configurations (Tailoring) of the product which is synonymous to the Guidance section IV.D. The below diagram outlines this thought to make it clear that you have the full validation of an instrument along with a separate Tailoring effort to adapt it to the different target population, format, language, or mode of administration.



This would aid users to understand the difference between the two distinct processes.

Subsequent Comments are noted (highlighted in yellow) in the succeeding markup text.

## TABLE OF CONTENTS

<b>I.</b>	<b>INTRODUCTION</b> .....	<b>1</b>
<b>II.</b>	<b>BACKGROUND</b> .....	<b>2</b>
<b>III.</b>	<b>PATIENT-REPORTED OUTCOMES — REGULATORY PERSPECTIVE</b> .....	<b>3</b>
<b>A.</b>	<b>Why Use Patient-Reported Outcome Instruments in Medical Product Development?</b> .....	<b>3</b>
1.	<i>Some Treatment Effects Are Known Only to the Patient</i> .....	3
2.	<i>Patients Provide a Unique Perspective on Treatment Effectiveness</i> .....	4
3.	<i>Formal Assessment May Be More Reliable Than Informal Interview</i> .....	4
<b>B.</b>	<b>A Taxonomy of PRO Instruments</b> .....	<b>4</b>
<b>IV.</b>	<b>EVALUATING PRO INSTRUMENTS</b> .....	<b>6</b>
<b>A.</b>	<b>Development of the Conceptual Framework and Identification of the Intended Application</b> .....	<b>7</b>
1.	<i>Identification of Concepts and Domains That Are To Be Measured</i> .....	7
2.	<i>Identification of the Intended Application of the PRO Instrument</i> .....	9
3.	<i>Identification of the Intended Population</i> .....	9
<b>B.</b>	<b>Creation of the PRO Instrument</b> .....	<b>9</b>
1.	<i>Generation of Items</i> .....	9
2.	<i>Choice of the Data Collection Method</i> .....	10
3.	<i>Choice of the Recall Period</i> .....	10
4.	<i>Choice of Response Options</i> .....	11
5.	<i>Evaluation of Patient Understanding</i> .....	12
6.	<i>Development of Format, Instructions, and Training</i> .....	12
7.	<i>Identification of Preliminary Scoring of Items and Domains</i> .....	13
8.	<i>Assessment of Respondent and Administrator Burden</i> .....	13
9.	<i>Confirmation of the Conceptual Framework and Finalization of the Instrument</i> .....	14
<b>C.</b>	<b>Assessment of Measurement Properties</b> .....	<b>15</b>
1.	<i>Evaluation of Reliability</i> .....	18
2.	<i>Evaluation of Validity</i> .....	18
3.	<i>Evaluation of Ability to Detect Change</i> .....	18
4.	<i>Choice of Methods for Interpretation</i> .....	19
a.	<i>Defining a minimum important difference</i> .....	19
b.	<i>Definition of responders</i> .....	20
<b>D.</b>	<b>• Modification of an Existing Instrument</b> .....	<b>20</b>
1.	<i>Revised Measurement Concept</i> .....	20
2.	<i>• Application to a New Population or Condition</i> .....	21
3.	<i>Changed Item Content or Instrument Format</i> .....	21
4.	<i>Changed Mode of Administration</i> .....	21
5.	<i>Changed Culture or Language of Application</i> .....	21
6.	<i>• Other Changes</i> .....	22
<b>E.</b>	<b>• Development of PRO Instruments for Specific Populations</b> .....	<b>22</b>
1.	<i>Children and Youth</i> .....	22
2.	<i>• Patients Cognitively Impaired or Unable to Communicate</i> .....	22
<b>V.</b>	<b>STUDY DESIGN</b> .....	<b>23</b>

<b>A.</b>	<b>General Protocol Considerations .....</b>	<b>23</b>
1.	<i>Blinding and Randomization .....</i>	23
2.	<i>Clinical Trial Quality Control .....</i>	24
3.	<i>Designing the Trial to Avoid Data Missing Due to Withdrawal From Exposure .....</i>	24
<b>B.</b>	<b>Frequency of Measurements .....</b>	<b>24</b>
<b>C.</b>	<b>Duration of Study .....</b>	<b>24</b>
<b>D.</b>	<b>Design Considerations for Multiple Endpoints .....</b>	<b>25</b>
<b>E.</b>	<b>Planning for Study Interpretation .....</b>	<b>25</b>
<b>F.</b>	<b>Specific Concerns When Using Electronic PRO Instruments .....</b>	<b>25</b>
<b>VI.</b>	<b>DATA ANALYSIS .....</b>	<b>27</b>
<b>A.</b>	<b>General Statistical Considerations .....</b>	<b>27</b>
<b>B.</b>	<b>Statistical Considerations for Using Multiple Endpoints .....</b>	<b>27</b>
<b>C.</b>	<b>Statistical Considerations for Composite Measures .....</b>	<b>28</b>
<b>D.</b>	<b>Statistical Considerations for Patient-Level Missing Data .....</b>	<b>29</b>
1.	<i>Missing Items Within Domains.....</i>	29
2.	<i>Missing Entire Domains or Entire Measurements .....</i>	29
<b>E.</b>	<b>Interpretation of Study Results .....</b>	<b>30</b>
<b>GLOSSARY .....</b>		<b>31</b>



1  
2  
3  
4  
5  
6  
7 8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19 **I. INTRODUCTION**  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36

## Guidance for Industry<sup>1</sup> Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims

This draft guidance, when finalized, will represent the Food and Drug Administration’s (FDA’s) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

### I. INTRODUCTION

This guidance describes how the FDA evaluates patient-reported outcome (PRO) instruments used as effectiveness endpoints in clinical trials. It also describes our current thinking on how sponsors can develop and use study results measured by PRO instruments to support claims in approved product labeling.<sup>2</sup> It does not address the use of PRO instruments for purposes beyond evaluation of claims made about a drug or medical product in its labeling. By explicitly addressing the review issues identified in this guidance, sponsors can increase the efficiency of their endpoint discussions with the FDA during the product development process, streamline the FDA’s review of PRO endpoint adequacy, and provide optimal information about the patient’s perspective of treatment benefit at the time of product approval.

A PRO is a measurement of any aspect of a patient’s health status that comes directly from the patient (i.e., without the interpretation of the patient’s responses by a physician or anyone else). In clinical trials, a PRO instrument can be used to measure the impact of an intervention on one or more aspects of patients’ health status, hereafter referred to as PRO concepts, ranging from the purely symptomatic (response of a headache) to more complex concepts (e.g., ability to carry out activities of daily living), to extremely complex concepts such as *quality of life*, which is

---

1 This guidance has been prepared by the Office of New Drugs and the Office of Medical Policy in the Center for Drug Evaluation and Research (CDER) in cooperation with the Center for Biologics Evaluation and Research (CBER) and the Center for Devices and Radiological Health (CDRH) at the Food and Drug Administration.

2 *Labeling*, as used in this guidance, refers to the medical product description and summary of use, safety, and effectiveness that must be approved by the FDA. See 21 CFR 201.56 and 201.57 for regulations pertaining to prescription drug (including biological drug) labeling. For medical device labeling, see 21 CFR 801. For blood and blood products for transfusion, see 21 CFR 606.122 Instruction Circular.

## ***Contains Nonbinding Recommendations***

*Draft — Not for Implementation*

37 widely understood to be a multidomain concept with physical, psychological, and social  
38 components. Data generated by a PRO instrument can provide evidence of a treatment benefit  
39 from the patient perspective. For this data to be meaningful, however, there should be evidence  
40 that the PRO instrument effectively measures the particular concept that is studied. Generally,  
41 findings measured by PRO instruments may be used to support claims in approved product  
42 labeling if the claims are derived from adequate and well-controlled investigations that use PRO  
43 instruments that reliably and validly measure the specific concepts at issue.

44  
45 The Glossary defines many of the terms used in this guidance. In particular, the term *instrument*  
46 refers to the actual questions or items contained in a questionnaire or interview schedule along  
47 with all the additional information and documentation that supports the use of these items in  
48 producing a PRO measure (e.g., interviewer training and instructions, scoring and interpretation  
49 manual). The term *conceptual framework* refers to how items are grouped according to  
50 subconcepts or *domains* (e.g., the item *walking without help* may be grouped with another item,  
51 *walking with difficulty*, within the domain of *ambulation*, and *ambulation* may be further  
52 grouped into the concept of *physical ability*).

53  
54 FDA's guidance documents, including this guidance, do not establish legally enforceable  
55 responsibilities. Instead, guidance documents describe the Agency's current thinking on a topic  
56 and should be viewed only as recommendations, unless specific regulatory or statutory  
57 requirements are cited. The use of the word *should* in Agency guidance documents means that  
58 something is suggested or recommended but not required.

## 61 **II. BACKGROUND**

62  
63 PRO instruments provide a means for measuring treatment benefits by capturing concepts related  
64 to how a patient feels or functions with respect to his or her health or condition. The concepts,  
65 events, behaviors, or feelings measured by PRO instruments can be either readily observed or  
66 verified (e.g., walking) or can be non-observable, known only to the patient and not easily  
67 verified (e.g., feeling depressed). Although an assessment of symptom improvement or pertinent  
68 function depends on patient perception, historically these assessments were often made by  
69 physicians who observed and interacted with patients (depression scales, heart failure severity  
70 scales, activities of daily living scales). Increasingly, such assessments are based on PRO  
71 instruments. The purpose of this guidance is to explain how the FDA evaluates such instruments  
72 for their usefulness in measuring and characterizing the benefit of medical product treatment.  
73

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

74 The amount and kind of evidence that the FDA expects to support a labeling claim measured by  
75 a PRO instrument is the same as that required for any other labeling claim.<sup>3</sup> As with other  
76 labeling claims, the determination of whether the PRO instrument supports an effectiveness  
77 endpoint includes an assessment of the ability of the PRO instrument to measure the claimed  
78 treatment benefit and is specific to the intended population and to the characteristics of the  
79 condition or disease treated. Endpoints measured by PRO instruments are most often used in  
80 support of claims that refer to a patient's symptoms or ability to function.

81  
82 Note, however, that PRO instruments that measure a simple concept may not be adequate to  
83 substantiate a more complex claim. For example, PRO-based evidence of improved symptoms  
84 alone generally is not sufficient to substantiate a claim related to improvement in a patient's  
85 ability to function or the patient's psychological state. Rather, to substantiate such a general  
86 claim, a sponsor should develop evidence to show not only a change in symptoms, but how that  
87 change translates into other specific endpoints such as ability to perform activities of daily  
88 living, or improved psychological state. Accordingly, many PRO instruments are specifically  
89 designed to assess both symptoms and other possible consequences of treatment.

90

91

### 92 **III. PATIENT-REPORTED OUTCOMES — REGULATORY PERSPECTIVE**

93

#### 94 **A. Why Use Patient-Reported Outcome Instruments in Medical Product** 95 **Development?**

96

97 PRO instruments are included in clinical trials for new medical products because (1) some  
98 treatment effects are known only to the patient; (2) there is a desire to know the patient  
99 perspective about the effectiveness of a treatment; or (3) systematic assessment of the patient's  
100 perspective may provide valuable information that can be lost when that perspective is filtered  
101 through a clinician's evaluation of the patient's response to clinical interview questions.

102

There should be a 4<sup>th</sup> reason added. It would be the timeliness (contemporaneousness) of the data entry. This is only valid for ePRO but should be mentioned as a reason.

#### 103 *1. Some Treatment Effects Are Known Only to the Patient*

104

105 For some treatment effects, the patient is the only source of data. For example, pain intensity  
106 and pain relief are the fundamental measures used in the development of analgesic products.

107 There are no observable or physical measures for these concepts.

108

---

<sup>3</sup>For drugs, section 505(d) of the Federal Food, Drug, and Cosmetic Act (the Act) establishes *substantial evidence* as the evidence standard for making conclusions that a drug will have a claimed effect and states that reports of adequate and well-controlled investigations provide the basis for determining whether there is *substantial evidence* to support claims of effectiveness for new drugs. See 21 CFR 314.126 for a description of the characteristics of an adequate and well-controlled investigation. See the guidance for industry *Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products* for considerations concerning the quantity of evidence necessary to meet the *substantial evidence* standard (<http://www.fda.gov/cder/guidance/index.htm>).

For medical devices, the Medical Device Amendments of 1976 to the Act established the assurance of safety and

effectiveness of medical devices intended for human use. See 21 CFR 860.7 for the evidence used in the determination of safety and effectiveness of a medical device.

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

### 2. *Patients Provide a Unique Perspective on Treatment Effectiveness*

PRO instruments can be developed to measure what patients want and expect from their treatment and what is most important to them. When used to measure study endpoints, PRO instruments can augment what is known about the product based on the clinician perspective or physiologic measures. This is important because improvements in clinical measures of a condition may not necessarily correspond to improvements in how the patient functions or feels. For example, clinically meaningful improvements in lung function as measured by spirometry may not correlate well with improvements in asthma-related symptoms and their impact on a patient's ability to perform daily activities.

### 3. *Formal Assessment May Be More Reliable Than Informal Interview*

Seeking information from patients about their symptoms and the impact of those symptoms on function is not new. In clinical practice, to obtain information known only to the patients, clinicians often assess patient status by informally asking questions such as, "How many pillows do you sleep on?" or, "Do you cough at night?" In clinical trials, clinical assessments are formalized using specific questions because a structured interview technique minimizes measurement error and ensures consistency. Self-completed questionnaires that are given directly to patients without the intervention of clinicians are often preferable to the clinician-administered interview and rating. Self-completed questionnaires capture directly the patient's perceived response to treatment, without a third party's interpretation, and may be more reliable than observer-reported measures because they are not affected by interobserver variability (which usually can be reduced only by extensive training of observers). On the other hand, PRO measures may be affected by interpatient variability if the instrument is not easily understood and completed by patients. Despite these concerns, well-developed and adequately validated PRO instruments have been shown to give answers that match the results obtained by the most expert assessors (indeed, that is the usual way their validity is assessed), and they appear to be particularly suitable in studies involving many investigators.

## **B. A Taxonomy of PRO Instruments**

PRO instruments measure concepts ranging from the state of discrete symptoms or signs (e.g., pain severity or seizure frequency) to the overall state of a condition (e.g., depression, heart failure, angina, asthma, urinary incontinence, or rheumatoid arthritis), where both specific symptoms and the impact of the condition (e.g., on function, activities, or feelings) can be measured, to feelings about the condition or treatment (e.g., worry about getting worse, having to avoid certain situations, feeling different from others). PRO concepts can be general (e.g., improvement in physical function, psychological well-being, or treatment satisfaction) or specific (e.g., decreased frequency, severity, or how bothersome the symptoms are). PRO concepts can also be generic (i.e., applicable in a broad scope of diseases or conditions as in the case of physical functioning), condition-specific (e.g., asthma-specific), or treatment-specific (e.g., measures of the toxicities of a class of drugs such as interferons or opioids).

***Contains Nonbinding Recommendations***  
*Draft — Not for Implementation*

153 Some PRO instruments (e.g., health-related *quality of life* instruments) attempt to measure both  
 154 the effectiveness and the side effects of treatment. PRO instruments that are used in clinical trials  
 155 to support effectiveness claims should measure the adverse consequences of treatment separately  
 156 from the effectiveness of treatment.

157  
 158 The specific attributes of a PRO instrument will affect the way it is developed, tested, and  
 159 incorporated into a study protocol to support conclusions of treatment benefit. Table 1 lists some  
 160 of the ways that PRO instruments can vary in their objectives, uses, and characteristics. When the  
 161 FDA reviews a PRO instrument, our goal is to determine whether its characteristics are  
 162 appropriate and adequate to support the study objectives.

163

164 **Table 1: Taxonomy of PROs Used in Clinical Trials**

Attribute	Types
Intended use of the measure	<ul style="list-style-type: none"> <li>• To define entry criteria for study populations</li> <li>• To evaluate efficacy</li> <li>• To evaluate adverse events <b>THIS IS NOT TRUE AS PRO DOES NOT COLLECT AE's. THIS SHOULD BE CHANGED TO STATE ONLY AE TRIGGERS</b></li> </ul>
Concepts measured	<ul style="list-style-type: none"> <li>• Overall health status</li> <li>• Symptoms/signs, individually or as a syndrome associated with a medical condition</li> <li>• Functional status (physical, psychological or social)</li> <li>• Health perceptions (e.g., self-rating of health or worry about condition)</li> <li>• Satisfaction with treatment or preference for treatment</li> <li>• Adherence to medical treatment</li> </ul>
Number of items	<ul style="list-style-type: none"> <li>• Single item for single concept</li> <li>• Multiple items for single concept</li> <li>• Multiple items for multiple domains within a concept</li> </ul>
Intended measurement population or condition	<ul style="list-style-type: none"> <li>• Generic</li> <li>• Condition-specific</li> <li>• Population-specific</li> </ul>
Mode of data collection	<ul style="list-style-type: none"> <li>• Interviewer-administered</li> <li>• Self-administered, with or without supervision</li> <li>• Computer-administered or computer-assisted</li> <li>• Interactively administered (e.g., interactive voice response systems or Web-based systems)</li> </ul>

165

*continued*

**Contains Nonbinding Recommendations**  
**Draft — Not for Implementation**

166

*Table 1, continued*

Attribute	Types
Timing and frequency of administration	<ul style="list-style-type: none"> <li>• As events occur</li> <li>• At regular intervals throughout a study</li> <li>• Baseline and end of treatment</li> </ul>
Types of scores	<ul style="list-style-type: none"> <li>• Single rating on a single concept (e.g., pain severity)</li> <li>• Index — single score combining multiple ratings of related domains or independent concepts</li> <li>• Profile — multiple uncombined scores of multiple-related domains</li> <li>• Battery — multiple uncombined scores of independent concepts</li> <li>• Composite — an index, profile, or battery</li> </ul>
Weighting of items or concepts	<ul style="list-style-type: none"> <li>• All items and domains are equally weighted</li> <li>• Items are assigned variable weights</li> <li>• Domains are assigned variable weights</li> </ul>
Response options	<ul style="list-style-type: none"> <li>• See Table 2 for examples of response options (types of PRO scales)</li> </ul>

167

168

169 **IV. EVALUATING PRO INSTRUMENTS**

170

171 The adequacy of a PRO instrument as a measure to support medical product claims depends on  
 172 its developmental history and demonstrated measurement properties. Sponsors are encouraged  
 173 to identify all endpoint measurement goals early in product development, before studies are  
 174 initiated, to provide the basis for product approval or claim substantiation, allowing adequate  
 175 time for PRO instrument identification, modification, or if necessary, new instrument  
 176 development. A new PRO instrument can be developed or an existing instrument can be  
 177 modified if sponsors determine that none is available, adequate, or applicable to their product  
 178 development program. When considering an instrument that has been modified from the  
 179 original, the FDA generally plans to evaluate the modified instrument just as it would a new one.  
 180 Therefore, in such instances, we encourage sponsors to document the original development  
 181 processes, all modifications made, and updated assessments of its measurement properties.

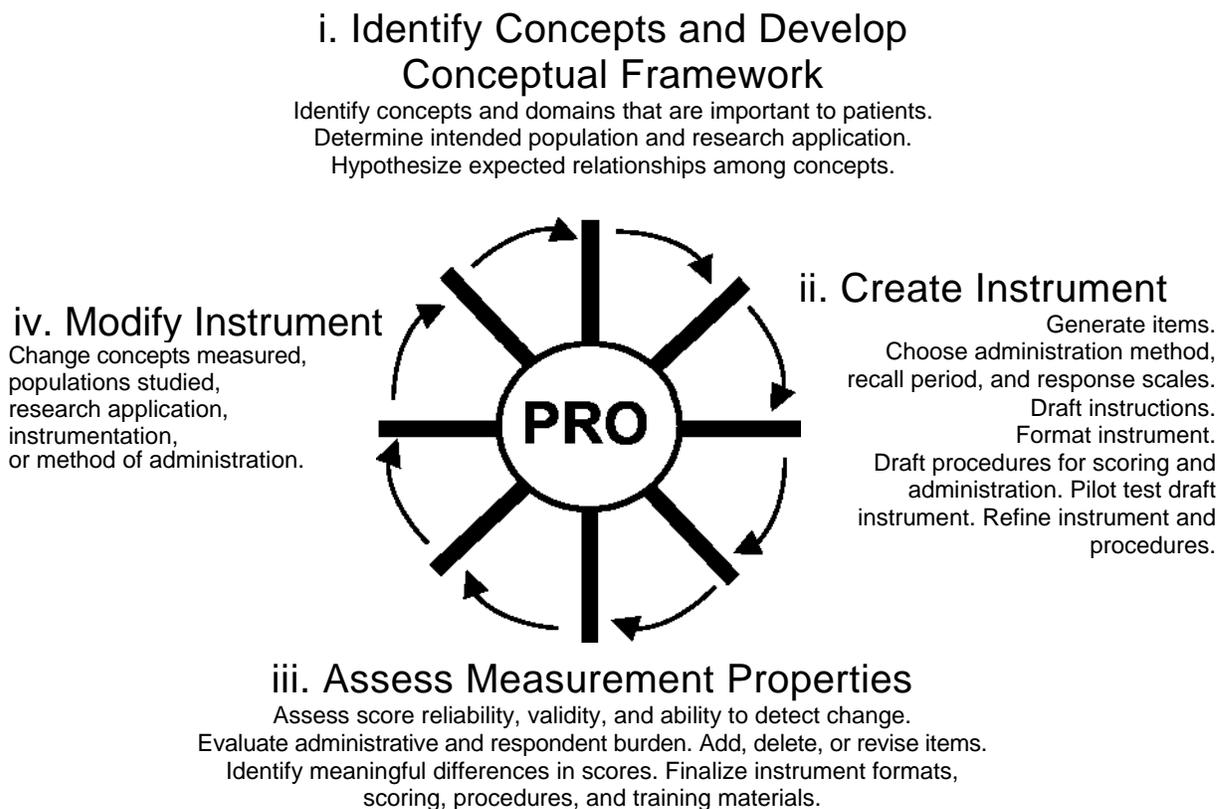
182

183 PRO instrument development, modification, and validation usually occur in a nonlinear fashion  
 184 with a varying sequence of events, simultaneous processes, or iterations. This iterative process  
 185 is presented as a *wheel and spokes* diagram, shown in Figure 1, and discussed in detail in  
 186 Sections IV.A. – IV.D. One or more parts of the original process may be repeated in new PRO  
 187 instrument development, modification, or change in application of an existing instrument. The  
 188 following five sections describe the steps usually taken in instrument development.

189

190  
191

**Figure 1: The PRO Instrument Development and Modification Process**



192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211

**A. Development of the Conceptual Framework and Identification of the Intended Application**

During the planning of clinical development programs, the FDA encourages sponsors to specify what claims they seek, determine what concepts underlie those claims, and then determine whether an adequate PRO instrument exists to assess and measure those concepts. If it doesn't, a new PRO instrument can be developed. The typical steps involved in the selection or development of PRO instruments for endpoints for clinical trials are described in the following sections.

*1. Identification of Concepts and Domains That Are To Be Measured*

One fundamental consideration in the development and use of a PRO instrument is whether the instrument's conceptual framework is appropriate and clearly defined. In some cases, of course, the question of what to measure may be obvious given the nature of the condition being treated. Generally, however, instrument developers choose the concepts and domains to be measured based on patient interviews along with reviews of the literature and expert opinion.

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

212 If documentation exists that a single item is a reliable and valid measure of the concept of interest  
213 (e.g., pain severity), a one-item PRO instrument may be a reasonable measure to support a claim  
214 concerning that concept. If the concept of interest is general (e.g., physical function), a single-  
215 item PRO instrument is usually unable to provide a complete understanding of the treatment’s  
216 effect because a single item cannot capture all the domains of the general concept. For this reason,  
217 single-item questions about general concepts that imply multiple domains rarely provide sufficient  
218 evidence to support claims about that general concept. However, single-item questions about  
219 general concepts can be useful to help interpret multi-item measures of the same concept and to  
220 determine whether important items or domains of a general concept are missing (e.g., when results  
221 using single general questions do not correlate with results using a multi-item questionnaire, this  
222 may be evidence that the questionnaire is not capturing all the important domains of the concept  
223 contained in the claim). Evidence from the patient cognitive debriefing studies (i.e., the interview  
224 schedule, transcript, and listing of all concepts elicited by a single item) can be used to determine  
225 when a concept is adequately captured by a single item.

226  
227 Multidomain PRO instruments can be used to support claims about a general concept if the PRO  
228 instrument has been appropriately developed and validated to measure the important and relevant  
229 domains of the general concept. The complex nature of multidomain PRO instruments, however,  
230 often raises significant questions about how to interpret and report results in a way that is not  
231 misleading. For example, if improvements in a score for a general concept (e.g., physical function)  
232 is driven by a single responsive domain (e.g., symptom improvement) while other important  
233 domains (e.g., physical abilities and activities of daily living) did not show a response, a general  
234 claim about improvements in physical function would not be supported. The FDA intends to  
235 review all evidence based on multidomain PRO measurements with particular attention to the  
236 precise claim that is supported by the results in the measured concepts or domains.

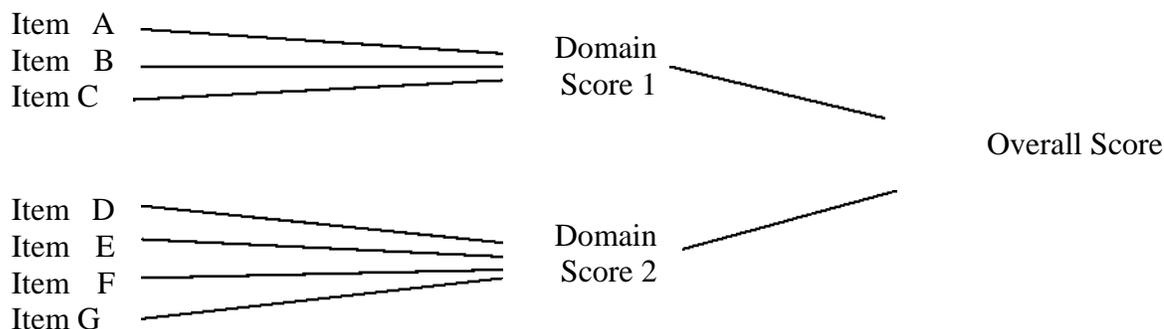
237  
238 Documentation of the instrument development process should reveal the means by which the  
239 domains were identified and named. This helps substantiate the adequacy of the measure to  
240 support both the general concept and the named domains. If a sponsor desires to support a claim  
241 based on a portion of a multi-item instrument (a domain or an item), the development and  
242 validation process should ensure that the instrument supports the measurement of the claimed  
243 concept. For example, some broad health status measures include item lists of symptoms that are  
244 summed in an overall score. Individual items that contribute to the overall score (e.g., dyspnea)  
245 generally would not support a dyspnea claim unless the items were developed to measure the  
246 claimed concept (e.g., the items validly and reliably capture the impact of treatment on dyspnea).

247  
248 For measures of general concepts, the FDA intends to review how individual items are associated  
249 with each other, how items are associated with each domain, and how domains are associated  
250 with each other and the general concept of interest. A diagram of the expected relationships  
251 among the PRO items and domains can help reviewers evaluate these relationships. The diagram in  
252 Figure 2 depicts a generic example of a conceptual framework where Domain Score 1, Domain  
253 Score 2, and Overall Score each represent related but separate concepts. Items in this diagram are  
254 aggregated into domains. In some measures, domains can be aggregated into an overall score.  
255 These expectations should be specified before the validation process begins.

256

257  
258  
259

**Figure 2: Diagram of a Conceptual Framework**



260  
261  
262  
263

2. *Identification of the Intended Application of the PRO Instrument*

264 It is also important to consider whether the development and demonstrated measurement  
265 properties of a PRO instrument provide an adequate basis for its planned use in the study to  
266 support a claim. This is best established before the study commences, but would in any case be  
267 part of the FDA’s application review. This is true whether the PRO instrument is generic,  
268 intended for use across multiple applications and populations, or specific, developed for a certain  
269 condition or population. The PRO instrument can be developed for a variety of roles, including  
270 defining trial entry criteria, including excessive severity, evaluating treatment benefit, or  
271 monitoring adverse events.

272  
273  
274

3. *Identification of the Intended Population*

275 The FDA plans to compare the patient population used in the PRO instrument development  
276 process to the study populations enrolled in clinical trials to determine whether the instrument is  
277 appropriate to that population with respect to patient age, sex, ethnic identity, and cognitive  
278 ability. Specific measurement considerations posed by pediatric, cognitively impaired, or  
279 seriously ill patients are discussed in Section IV.E.

280  
281  
282

**B. Creation of the PRO Instrument**

283 When developing a PRO instrument, sponsors are encouraged to assess its adequacy in the  
284 context of the following development processes.

285  
286  
287

1. *Generation of Items*

288 It is important to consider the procedures used to identify the set of items selected to measure a  
289 specific concept. PRO instrument items can be generated from literature reviews, transcripts  
290 from focus groups, or interviews with patients, clinicians, family members, researchers, or other  
291 sources. Depending on the conceptual framework, the FDA may review whether appropriate

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

292 individuals and sources were used and how information gleaned from those sources was used in  
293 the PRO instrument development process.

294  
295 PRO instrument item generation is incomplete without patient involvement. Item generation  
296 generally incorporates the input of a wide range of patients with the condition of interest to  
297 represent appropriate variations in severity and in population characteristics such as age or sex.  
298 The FDA plans to review instrument development (e.g., results from patient interviews or focus  
299 groups) to determine whether adequate numbers of patients have supported the opinion that the  
300 specific items in the instrument are adequate and appropriate to measure the concept.

301  
302 Items that ask patients to respond hypothetically or that give patients the opportunity to respond  
303 on the basis of their desired condition rather than on their actual condition are not recommended.  
304 For example, in assessing the concept *performance of daily activities*, it is more appropriate to  
305 ask whether or not the respondent performs specific activities (and if so, with how much  
306 difficulty) than whether or not he or she can perform daily activities (because patients may report  
307 they are able to perform a task even when they never do so). Of course, it would be critical to  
308 know that each item refers to something that patients actually do.

309  
310 It is also important to consider all of the item generation techniques used, including any  
311 theoretical approach used, the populations studied, sources of items, selection and reduction of  
312 items, cognitive debriefing interviews, pilot testing, importance ratings, and quantitative  
313 techniques for item evaluation such as factor analysis and item-response analysis.

314  
315 2. *Choice of the Data Collection Method*

316  
317 Sponsors should consider the method of data collection and all procedures and protocols  
318 associated with instrument administration, including instructions to interviewers, instructions for  
319 self-administration, instructions for supervising self-administration, case report forms or  
320 examples of electronic PRO instruments, and other special considerations specific to the mode of  
321 administration including data quality control procedures. Modes of administration include  
322 interview, paper-based, electronic, Web-based, and interactive voice response formats. The  
323 FDA intends to review the comparability of data obtained when using multiple modes of  
324 administration to determine whether pooling of results from the multiple modes is appropriate.

I would like the above sentence expanded to state the psychometric issues surrounding mixing data  
collected electronically and via paper.

325  
326 3. *Choice of the Recall Period*

327  
328 Sponsors should also evaluate the rationale and the appropriateness of the recall period for a  
329 PRO instrument. To this end, it is important to consider patients' ability to accurately recall the  
330 information requested as proposed. The choice of recall period that is most suitable depends on  
331 the purpose and intended use of the instrument, the characteristics of the disease/condition, and  
332 the treatment to be tested. When evaluating PRO-based claims, the FDA intends to review the  
333 study protocol to determine what steps were taken to ensure that patients understand the  
334 appropriate recall period. If a patient diary or some other form of unsupervised data entry is

335 used, the FDA plans to review the protocol to determine what measures are taken to ensure that

## ***Contains Nonbinding Recommendations***

*Draft — Not for Implementation*

336 patients make entries according to the study design and not, for example, just before a clinic visit  
337 when their reports will be collected.

338  
339 PRO instruments that require patients to rely on memory, especially if they must recall over a  
340 period of time, or to average their response over a period of time may threaten the accuracy of  
341 the PRO data. It is usually better to construct items that ask patients to describe their current  
342 state than to ask them to compare their current state with an earlier period or to attempt to  
343 average their experiences over a period of time.

344

### 345 4. *Choice of Response Options*

346

347 It is also important to consider whether the response options are consistent with the purpose and  
348 intended use of the PRO instrument. Table 2 describes the types of response options that are  
349 typically used in clinical trials.

350

351

**Table 2: Types of Response Options**

<b>Type</b>	<b>Description</b>
Visual analog scale (VAS)	A line of fixed length (usually 100 mm) with words that anchor the scale at the extreme ends and no words describing intermediate positions. Patients are instructed to place a mark on the line corresponding to their perceived state. These scales often produce a false sense of precision.
Anchored or categorized VAS	A VAS that has the addition of one or more intermediate marks positioned along the line with reference terms assigned to each mark to help patients identify the locations (e.g., half-way) between the ends of the scale.
Likert scale	An ordered set of discrete terms or statements from which patients are asked to choose the response that best describes their state or experience.
Rating scale	A set of numerical categories from which patients are asked to choose the category that best describes their state or experience. The ends of rating scales are anchored with words but the categories do not have labels.
Event log	Specific events are recorded as they occur using a patient diary or other reporting system (e.g., interactive voice response system)
Pictorial scale	A set of pictures applied to any of the other types of response options. Pictorial scales are often used in pediatric questionnaires but also have been used for patients with cognitive impairments and for patients who are otherwise unable to speak or write.
Checklist	Checklists provide a simple choice between a limited set of options, such as <i>Yes</i> , <i>No</i> , and <i>Don't know</i> . Some checklists ask patients to place a mark in a space if the statement in the item is true. Checklists are reviewed for completeness and nonredundancy.

352

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

353 Response choices are generally considered appropriate when:

- 354 • Wording used in responses is clear and appropriate (e.g., anchoring a scale using the term  
355 *normal* assumes that patients understand what is normal).
- 356 • Responses are appropriate for the intended population. For example, patients with visual  
357 impairment may find the VAS difficult to complete.
- 358 • Responses offer a clear distinction between choices (e.g., patients may not distinguish  
359 between *intense* and *severe* if both are offered as response choices to describe their pain).
- 360 • Instructions to patients for completing the questionnaire and selecting response options  
361 are adequate.
- 362 • The number of response options is justified.
- 363 • Response options are appropriately ordered and appear to represent equal intervals.
- 364 • Response options avoid potential ceiling or floor effects (e.g., introducing more  
365 categories to capture worsening or improvement so that fewer patients respond at the top  
366 or bottom of the response continuum).
- 367 • Response options do not bias the direction of responses (e.g., offering one negative  
368 choice, one neutral choice, and two or more positive choices on a scale makes it more  
369 likely for patients to respond that they feel or function better).

370

### 371 5. *Evaluation of Patient Understanding*

372

373 Sponsors are encouraged to examine the procedures used with patients to determine readability  
374 and understanding of the items included in the PRO instrument. The FDA's evaluation of these  
375 procedures is likely to include a review of a cognitive debriefing report containing the  
376 readability test used, the script used in patient cognitive debriefing interviews, the transcript of  
377 the interviews, the analysis of the interview results, and the actions taken to delete or modify an  
378 item in response to the cognitive debriefing interview or pilot test results.

379

### 380 6. *Development of Format, Instructions, and Training*

381

382 PRO study results can vary according to the instructions to patients or the training given to the  
383 interviewer or persons supervising PRO data collection. Sponsors should consider all PRO  
384 instrument instructions and procedures contained in publications and user manuals provided by  
385 developers, including procedures for reviewing completed questionnaires and re-administration  
386 to avoid missing data or clarify responses. Other important considerations include the format of  
387 the questionnaire, the final wording of PRO instruments as implemented in clinical trials, and  
388 any potentially important changes in presentation or format. Examples of changes that can alter  
389 the way that patients respond to the same set of questions include:

- 390 • Changing an instrument from paper to electronic format
- 391 • Changing the timing of or procedures for PRO instrument administration within the clinic  
392 visit
- 393 • Changing the order of items or deleting portions of a questionnaire
- 394 • Changing the instructions or the placement of instructions within the PRO instrument

395

396 It is important that the PRO instrument format used in the clinical trial be consistent with the  
397 format that is used in the instrument validation process. *Format* refers to the exact appearance of

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

398 the instrument. Instrument format is specific to the mode of administration, including paper and  
399 pencil, interviewer-administered or supervised, or electronic data collection. The FDA plans to  
400 review the PRO instrument in the format used in the clinical trial case report forms, including the  
401 order and numbering of items, the presentation of response options in single response or grid  
402 formats, the grouping of items, patterns for skipping questions that are not applicable, and all  
403 instructions to patients in the interview schedule or on the questionnaire.

404  
405 The FDA recommends that the PRO instrument development process includes the generation of a  
406 user manual that specifies how to incorporate the instrument into a clinical trial in a way that  
407 minimizes administrator burden, patient burden, missing data, and poor data quality.

### 408 7. *Identification of Preliminary Scoring of Items and Domains*

409  
410  
411 For each item, numerical scores are generally assigned to each answer category based on the most  
412 appropriate scale of measurement for the item (e.g., nominal, ordinal, interval, or ratio scales).  
413 The FDA intends to consider whether a PRO measure conforms to assumptions that the response  
414 choices represent appropriate intervals by reviewing distributions of item responses.

415  
416 A scoring algorithm creates a single score from multiple items. Equally weighted scores for each  
417 item are appropriate only when the responses to the items are relatively uncorrelated. Otherwise,  
418 the assignment of equal weights will overweight correlated items and underweight independent  
419 items. Even when items are uncorrelated, assigning equal weights to each item may overweight  
420 certain items if the number of response options or the values associated with response options  
421 varies by item. The same weighting concerns apply with added complexity when combining  
422 domain scores into a single overall score.

423  
424 When empirically determined patient preference ratings are used to weight items or domains, the  
425 FDA also intends to review the composition of samples and the process used to determine the  
426 preference weights. Because preference weights are often developed for use in resource allocation  
427 (e.g., as in cost-effectiveness analysis that may use predetermined community weights), it is  
428 tempting to use those same weights in the clinical trial setting to demonstrate treatment benefit.  
429 However, this practice is discouraged unless the relationship of the preference weights to the  
430 intended study population is known and found adequate and appropriate.

### 431 8. *Assessment of Respondent and Administrator Burden*

432  
433  
434 Undue physical, emotional, or cognitive strain on patients are burdens that will generally decrease  
435 the quality and quantity of PRO data. Factors that can contribute to respondent burden include the  
436 following:

- 437 • Length of questionnaire or interview
- 438 • Formatting
- 439 • Font size too small to read easily
- 440 • New instructions for each item
- 441 • Words or sentence structures that require a technical knowledge or developmental level
- 442 • beyond that of the patients in the trials

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

- 443 • Requirement that patients consult records to complete responses
- 444 • Privacy of the setting in which the PRO is completed (e.g., not providing a private space
- 445 for patients to complete questionnaires containing sensitive information about their sexual
- 446 performance or substance abuse history)
- 447 • Inadequate time to complete questionnaires or interviews
- 448 • Literacy level too high for population
- 449 • Questions that patients are unwilling to answer
- 450 • Perception by patients that the interviewer wants or expects a particular response

451  
452 The degree of respondent burden that is acceptable for instruments in clinical trials depends on  
453 the frequency and timing of PRO assessments in a protocol and on the severity of the illness or  
454 toxicity of the treatment studied. For example, if the questionnaire contains instructions to skip  
455 one or more questions based on responses to a previous question, respondents may fail to  
456 understand what is required and make errors in responding or find the assessment too  
457 complicated to complete. Sponsors should consider missing data and the refusal rate as possible  
458 indications of unacceptable patient burden or inappropriate items or response options.

459  
460 9. *Confirmation of the Conceptual Framework and Finalization of the Instrument*

461  
462 The FDA intends to examine the final version of an instrument in light of its development  
463 history, including documentation of the complete list of items generated and the reasons for  
464 deleting or modifying items, as illustrated in Table 3. It will be important to determine from  
465 empirical data submitted whether the conceptual framework (e.g., the expected relationships  
466 between items, domains, and measurement concepts as diagrammed in Figure 2) have been  
467 demonstrated.

468  
469

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

470 **Table 3: Common Reasons for Chan<sup>g</sup>in<sup>g</sup> PRO Instruments Durin<sup>g</sup> Initial Develo<sup>p</sup>ment**

<b>Item Property</b>	<b>Reason for Change or Deletion</b>
Clarity or relevance	<ul style="list-style-type: none"> <li>• Reported as not relevant by a large segment of the population of interest</li> <li>• Generates an unacceptably large amount of missing data points</li> <li>• Generates many questions or requests for clarification from patients as they complete the PRO instrument</li> <li>• Patients interpret items and responses in a way that is inconsistent with the conceptual framework</li> </ul>
Response range	<ul style="list-style-type: none"> <li>• A high percent of patients respond at the floor (worst end of the response scale) or ceiling (optimal end of the response scale)</li> <li>• Patients note that none of the response choices apply to them</li> <li>• Item means are highly skewed</li> </ul>
Variability	<ul style="list-style-type: none"> <li>• All patients give the same answer (i.e., no variance)</li> <li>• Most patients choose only one of the response choices</li> <li>• Differences among patients are not detected when important differences are known</li> </ul>
Reproducibility	<ul style="list-style-type: none"> <li>• Unstable scores over time when there is no logical reason for variation from one assessment to the next</li> </ul>
Inter-item correlation	<ul style="list-style-type: none"> <li>• Item uncorrelated with other items in the same concept of interest</li> </ul>
Ability to detect change	<ul style="list-style-type: none"> <li>• Item is nonresponsive (i.e., does not change when there is a known change in the concepts of interest)</li> </ul>
Item discrimination	<ul style="list-style-type: none"> <li>• Item is highly correlated with measures of concepts other than the one it is intended to measure</li> </ul>
Redundancy	<ul style="list-style-type: none"> <li>• Item duplicates information collected with other items that have equal or better measurement properties</li> </ul>

471  
472 **C. Assessment of Measurement Properties**

473  
474 The FDA generally intends to review a PRO instrument for: reliability, validity, ability to detect  
475 change, and interpretability (e.g., minimum important difference). The FDA plans to review the  
476 measurement properties that are specific to the documented conceptual framework, confirmed  
477 scoring algorithm, administration procedures, and questionnaire format in light of the study  
478 population, study design, and statistical analysis plan. The sociodemographic and medical  
479 characteristics of any sample used to develop or validate a PRO instrument determine its  
480 appropriateness for future clinical study settings. (See Table 4.)

***Contains Nonbinding Recommendations***  
*Draft — Not for Implementation*

481 **Table 4: Measurement Properties Reviewed for PRO Instruments Used in Clinical Trials**

<b>Measurement Property</b>	<b>Test</b>	<b>What is Assessed</b>	<b>FDA Review Considerations</b>
Reliability	Test-retest	Stability of scores over time when no change has occurred in the concept of interest	Does the PRO instrument reliably measure the concepts it was designed to measure? Were appropriate reliability tests conducted? What was the quality of the evidence of reliability?
	Internal consistency	Whether the items in a domain are intercorrelated, as evidenced by an internal consistency statistic (e.g., coefficient alpha)	
	Inter-interviewer reproducibility (for interviewer-administered PROs only)	Agreement between responses when the PRO is administered by two or more different interviewers	
Validity	Content-related	Whether items and response options are relevant and are comprehensive measures of the domain or concept	Do items in the verbatim copy of the PRO instrument appear to measure the concepts they are intended to measure in a useful way? Have patients similar to those participating in the clinical trial confirmed the completeness and relevance of all items?
	Ability to measure the concept (also known as construct-related validity; can include tests for discriminant, convergent, and known-groups validity)	Whether relationships among items, domains, and concepts conform to what is predicted by the conceptual framework for the PRO instrument itself and its validation hypotheses.	Do observed relationships between the items and domains confirm the hypotheses in the conceptual framework? Do results compare favorably with results from a similar but independent measure? Do results distinguish one group from another based on a prespecified variable that is relevant to the concept of interest?
	Ability to predict future outcomes (also known as predictive validity)	Whether future events or status can be predicted by changes in the PRO scores	Do PRO scores predict subsequent events or outcomes accurately?

482

*continued*

***Contains Nonbinding Recommendations***  
*Draft — Not for Implementation*

483 *Table 4, continued*

<b>Measurement Property</b>	<b>Test</b>	<b>What is Assessed</b>	<b>FDA Review Considerations</b>
Ability to detect change	Includes calculations of effect size and standard error of measurement among others	Whether PRO scores are stable when there is no change in the patient, and the scores change in the predicted direction when there has been a notable change in the patient as evidenced by some effect size statistic. Ability to detect change is always specific to a time interval.	Has ability to detect change been demonstrated in a comparative trial setting, comparing mean group scores or proportion of patients who experienced a response to the treatment? Has ability to detect change been assessed for the time interval appropriate to study?
Interpretability	Smallest difference that is considered clinically important; this can be a specified difference (the minimum important difference (MID)) or, in some cases, any detectable difference. The MID is used as a benchmark to interpret mean score differences between treatment arms in a clinical trial	Difference in mean score between treatment groups that provides convincing evidence of a treatment benefit. Can be based on experience with the measure using a distribution-based approach, a clinical or nonclinical anchor, an empirical rule, or a combination of approaches. The definition of an MID using a clinical anchor is sometimes called an MCID.	The FDA is specifically requesting comment on appropriate review of derivation and application of an MID in the clinical trial setting.
	Responder definition — used to identify responders in clinical trials for analyzing differences in the proportion of responders between treatment arms	Change in score that would be clear evidence that an individual patient experienced a treatment benefit. Can be based on experience with the measure using a distribution-based approach, a clinical or nonclinical anchor, an empirical rule, or a combination of approaches.	The FDA is specifically requesting comment on appropriate review of derivation and application of responder definitions when used in clinical trials.

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

484           1.           *Evaluation of Reliability*

485

486 Because clinical trials involve change over time, the adequacy of a PRO instrument for use in a  
487 clinical trial depends on its reliability. Because clinical trials are intended to provide unbiased  
488 estimates of true treatment impact, systematic and/or other changes in measurement methods  
489 may undermine the purpose of the trial.

490

491 Test-retest reliability is the most important type of reliability for PRO instruments used in clinical  
492 trials. Test-retest is most informative when the time interval chosen between the test and retest is  
493 appropriate for identifying stability in reference to the clinical trial protocol.

494

495 Internal consistency reliability, in the absence of test-retest reliability, does not generally constitute  
496 sufficient evidence of reliability for clinical trial purposes. When PRO instruments are  
497 interviewer-administered, inter-interviewer reproducibility is critical.

498

499           2.           *Evaluation of Validity*

500

501 The FDA recognizes that the validation of an instrument is an ongoing process and that validity  
502 relates to both the instrument itself and how it is used. Sponsors should consider a PRO endpoint  
503 for evidence of content-related validity, the instrument’s ability to measure the stated concepts,  
504 and the instrument’s ability to predict future outcomes, as illustrated in Table 4.

505

506 If instrument developers expected the instrument to give results for the measured concept similar to  
507 those measured by existing PRO or non-PRO measures (e.g., physical or physician-based  
508 measures), the FDA is interested in documented demonstration of those relationships to determine  
509 whether the instrument convincingly measures that concept and can therefore support a claim  
510 about that concept. If developers expected the instrument to discriminate between patient groups  
511 (e.g., between patients with different levels of severity), the FDA is interested in evidence that  
512 shows the instrument meaningfully discriminates.

513

514 In some cases, some types of validity testing are not possible due to the nature of the concept to be  
515 measured. In such instances, the FDA generally plans to review the cumulative evidence for the  
516 appropriate use of the measure and apply it to the interpretation of clinical study results.

517

518           3.           *Evaluation of Ability to Detect Change*

519

520 When a concept is expected to change, the values for the PRO instrument measuring that concept  
521 should change. If there is clear evidence that patient experience relative to the concept has  
522 changed, but the PRO scores do not change, the validity of the PRO instrument should be  
523 questioned. If there is evidence that PRO scores are affected by changes that are not specific to the  
524 concept of interest, the validity of the PRO instrument should be questioned.

525

526 The ability of an instrument to detect change influences the sample size needed to evaluate the  
527 effectiveness of treatment. The extent to which the PRO instrument’s ability to detect change  
528 varies by important patient subgroups (e.g., sex, race, age, or ethnicity) can affect clinical trial

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

529 results. It is important to identify any important subgroup differences in ability to detect change  
530 so that these differences can be taken into account in assessing results.

531

532 4. *Choice of Methods for Interpretation*

533

534 The following sections describe some of the methods that have helped sponsors and the FDA  
535 interpret clinical trial results based on PRO endpoints.

536

537 a. Defining a minimum important difference

538

539 Many PRO instruments are able to detect mean changes that are very small; accordingly it is  
540 important to consider whether such changes are meaningful. Therefore, it is appropriate for a  
541 critical distinction to be made between the mean effect seen (and what effect might be  
542 considered important) and a change in an individual that would be considered important, perhaps  
543 leading to a definition of a *responder*. For many widely used measures (pain, treadmill distance,  
544 HamD), the ability to show *any* difference between treatment groups has been considered  
545 evidence of a relevant treatment effect. If PRO instruments are to be considered more sensitive  
546 than past measures, it can be useful to specify a minimum important difference (MID) as a  
547 benchmark for interpreting mean differences. An MID is usually specific to the population  
548 under study.

549

550 The FDA has reviewed MIDs derived in many ways. Examples include:

- 551 • Mapping changes in PRO scores to clinically relevant and important changes in non-PRO  
552 measures of treatment outcome in the condition of interest (e.g., when PRO measures of  
553 asthma or COPD are mapped to spirometry scores).
- 554 • Mapping changes in PRO scores to other PRO scores to arrive at an MID that is  
555 appreciable to patients (e.g., when multi-item PROs are mapped to a single question  
556 asking the patient to rate his or her global impression of change since the start of  
557 treatment). A problem with this approach is that it uses individual rates to reach a  
558 conclusion about mean effects. It may be more useful to look at the distribution of  
559 individual effects in treatment and control groups.
- 560 • Using a distribution-based approach (e.g., defining the MID as 0.5 times the standard  
561 deviation). This, of course, may bear no relation to the patient's assessment and is  
562 usually inadequate in isolation.
- 563 • Using an empirical rule (e.g., 8 percent of the theoretical range of scores). Again, this  
564 arbitrary approach does not take into account patient preferences or assessment.

565

566 If an MID is to be applied to clinical study results, it is generally helpful to use a variety of  
567 methods to discover whether concordance among methods confirms the choice of an MID.<sup>4</sup>

568

---

<sup>4</sup>The FDA is specifically asking for comment on the need for, and appropriate standards for, MID definitions applied to PRO instruments used in clinical studies.

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

569                   b. Definition of responders

570  
571   There may be situations where it is more reasonable to characterize the meaningfulness of an  
572   individual's response to treatment than a group's response, and there may be interest in  
573   characterizing an individual patient as a responder to treatment, based upon pre specified criteria  
574   backed by empirically derived evidence supporting the responder definition as a measure of  
575   benefit. Such examples include categorizing a patient as a responder based upon a prespecified  
576   change from baseline on one or more scales; a change in score of a certain size or greater (e.g., a 2-  
577   point change on an 8-point scale); or a percent change from baseline.<sup>5</sup>

578  
579                   **D.       Modification of an Existing Instrument**

580  
581   When a PRO instrument is modified, additional validation studies may be needed to confirm the  
582   adequacy of the modified instrument's measurement properties. The extent of additional  
583   validation recommended depends on the type of modification made. For example, small  
584   nonrandomized studies may be adequate to assess the results of changing a response scale from  
585   vertical to horizontal. On the other hand, if the PRO instrument is to be used in an entirely new  
586   population of patients, a small randomized study to ascertain the measurement properties in the  
587   new population may minimize the risk that the instrument will not perform adequately in a phase 3  
588   study.

589  
590   The FDA intends to consider a modified instrument as a different instrument from the original  
591   and will consider measurement properties to be version-specific. The FDA recommends  
592   additional validation to support the development of a modified PRO instrument when one or  
593   more of the following modifications occur.

594  
595                   1.       *Revised Measurement Concept*

596  
597   An instrument that is developed and validated to measure one concept is used to measure a  
598   different concept. For example:

- 599       • A single domain from a multiple domain PRO is administered without the other domains  
600       • Response options are changed to assess a different quality (e.g., frequency versus how  
601       bothersome)  
602       • An index or composite score is used to summarize multiple PRO concepts/domains when  
603       existing validation applies only to concept/domain-specific scores  
604       • Items from an existing PRO instrument are used to create a new instrument  
605       • One or more items from an existing instrument are used to support a claim for a concept  
606       the items were not developed to measure

607

---

<sup>5</sup>The FDA is specifically asking for comment on the appropriate review standards for the definition of a responder when applied to PRO instruments used in clinical studies to support medical product development.

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

608           2.           *Application to a New Population or Condition*

609  
610 An instrument developed for use in one population or condition is used in a different patient  
611 population or condition. For example:

- 612       • Patients in the proposed trial have a disease, condition, or severity level that is different
- 613       from that of the patient population used for instrument development and validation
- 614       • Patients in the proposed trial differ in age, gender, race, or developmental or life stage
- 615       from those for instrument development and validation

616  
617           3.           *Changed Item Content or Instrument Format*

618  
619 An instrument is altered in item content or format. This includes changes in the following:

- 620       • Number of items (more or fewer) used to assess a concept or domain
- 621       • Wording or placement of instructions
- 622       • Wording or order of the items
- 623       • Wording, scaling, ordering, or number of response options
- 624       • Recall period associated with an item
- 625       • Point of reference for comparison for an item or domain
- 626       • Weighting of items
- 627       • Scoring (including creation of summary scores, subdomain scores, or cut-points)
- 628       • Any changes that could alter the patient’s interpretation of the instructions, items, or
- 629       response options

630  
631           4.           *Changed Mode of Administration*

632  
633 An instrument’s data collection mode is altered. For example:

- 634       • An interviewer-administered or supervised questionnaire is modified for self-
- 635       administration (skip patterns can be a problem in this situation)
- 636       • Paper-and-pencil self-administered PRO is modified to be administered by computer or
- 637       other electronic device (e.g., computer adaptive testing, interactive voice response
- 638       systems, Web-based questionnaire administration, computer)
- 639       • Instructions or procedures for administration within a trial differ from those used in
- 640       validation studies (can alter the meaning of the responses from that of the original
- 641       version)

642  
643           5.           *Changed Culture or Language of Application*

644  
645 An instrument developed in one language or culture is adapted or translated for use in another  
646 language or culture. The FDA recommends that sponsors provide evidence that the methods and  
647 results of the translation process were adequate to ensure that the validity of the responses is not  
648 affected. Some examples include the following:

- 649       • PRO instruments are developed initially in one language, culture, or ethnic group and are
- 650       used subsequently in another
- 651       • PRO instruments developed and validated outside the United States are applied to the
- 652       U.S. population

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

653  
654 Sponsors should consider whether generally accepted standards for translation and cultural  
655 adaptation have been used to support the validity of data from a translated/adapted PRO  
656 instrument, including but not restricted to the following:

- 657 • The background and experience of the persons involved in the translation/adaptation
- 658 • The translation/adaptation methodology used
- 659 • The harmonization of different versions
- 660 • The evidence that measurement properties for translated versions are comparable

661  
662 6. *Other Changes*

663  
664 Other changes to the PRO instrument or the way in which it is assessed that may necessitate  
665 additional validation include:

- 666 • The PRO instrument was not developed and validated for use in a clinical trial
- 667 • A PRO instrument developed and previously used as a stand-alone assessment is  
668 included as a part of a battery of measures
- 669 • A PRO developed to measure a treatment benefit is subsequently used to measure a  
670 decrement as interpreted by a score change in the opposite direction

671  
672 **E. Development of PRO Instruments for Specific Populations**

673  
674 Measurement of PRO concepts in children and youth, and in patients who have cognitive  
675 impairment, introduces challenges in addition to those already mentioned. These are discussed in  
676 the following sections.

677  
678 1. *Children and Youth*

679  
680 In general, the review issues related to the development and validation of pediatric PRO  
681 instruments are similar to those detailed for adults. It is important that PRO instruments  
682 developed for adults are not used in pediatric populations unless the measurement properties are  
683 similar in all age groups tested. We recommend that instruments intended for use in pediatric  
684 populations be rigorously developed and validated according to the principles described earlier.  
685 Additional review issues for PRO instruments applied in children and youth include age-related  
686 vocabulary, language comprehension, comprehension of the health concept measured, and  
687 duration of recall. Instrument development and validation testing within fairly narrow age  
688 groupings is important to account for developmental differences and to determine the lower age  
689 limit at which children can understand the questions and provide reliable and valid responses that  
690 can be compared across age categories.

691  
692 2. *Patients Cognitively Impaired or Unable to Communicate*

693  
694 Over the course of some clinical trials, it can be anticipated that patients may become too ill to  
695 complete a questionnaire or to respond to an interviewer. In such cases, proxy reporting may  
696 help to prevent missing data. When this situation is anticipated, the FDA encourages the  
697 inclusion of proxy reports in parallel with patient self-report from the beginning of the study

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

698 (i.e., even before the patient is no longer able to answer independently) so that the relationship  
699 between the patient reports and the proxy reports can be assessed.

700  
701

702 **V. STUDY DESIGN**

703  
704 The same study design principles that apply to other endpoint measures apply to PROs. This  
705 section, therefore, focuses primarily on issues unique to PROs.

706

707 **A. General Protocol Considerations**

708

709 If the goal of PRO measurement is to support claims, we recommend that measurement of the  
710 PRO concept be clearly stated as a specific study objective. It is important that the protocol  
711 include the exact format and version of the specific PRO instrument to be administered. In the  
712 process of considering the NDA/BLA/PMA or NDA/BLA/PMA supplement, the FDA intends to  
713 compare both the planned and actual use of the PRO instrument and its analysis.

714

715 *1. Blinding and Randomization*

716

717 Because responses to PRO measures are subjective, representing a patient's impression, open-  
718 label studies, where patients and investigators are aware of assigned therapy, are rarely credible.  
719 Patients who know they are in an active treatment group may overestimate benefit while those  
720 who know they are not receiving active treatment may underreport any improvement actually  
721 experienced. Every effort should be made to assure that patients are masked to treatment  
722 assignment throughout the trial. If the treatment has obvious effects, blinding may be difficult.  
723 The impact of possible unblinding is important to consider in the interpretation of study results.

724

725 The importance of blinding can be determined, in part, by the characteristics of the PRO  
726 instrument used. For example, questions that ask how patients' current status compares to  
727 baseline seem likely to be more influenced by unblinding (optimism can readily be expressed as  
728 a favorable comparison) than questions that ask about current status (which requires a current  
729 assessment, not a statement about duration). Questions that ask for current status, or PRO  
730 instruments that ask many questions, are harder to answer in a biased way when previous  
731 answers are not available. For the same reasons, allowing patients access to previous responses  
732 can bias results when unblinding is a possibility. This is, however, an area that could benefit  
733 from rigorous study.

734

735 There are certain situations, particularly in the development of medical devices, where blinding  
736 is not feasible and other situations where there is no reasonable control group (and therefore no  
737 randomization). When a PRO instrument appears useful in assessing patient benefit in those  
738 situations, the FDA encourages sponsors to confer with the appropriate review division.

739

***Contains Nonbinding Recommendations***  
*Draft — Not for Implementation*

740           2.       *Clinical Trial Quality Control*

741  
742 Study quality can be optimized at the design stage by specifying procedures to minimize  
743 inconsistencies in trial conduct. Examples of standardized instructions and processes that may  
744 appear in the protocol include:

- 745       • Standardized training and instructions to patients for self-administered PRO instruments
- 746       • Standardized interviewer training and interview format for PRO instruments administered  
747           in an interview format
- 748       • Standardized instructions for the clinical investigators regarding patient supervision,  
749           timing and order of questionnaire administration during or outside the office visit,  
750           processes and rules for questionnaire review for completeness, and documentation of  
751           how and when data are filed, stored, and transmitted to or from the study site

752  
753           3.       *Designing the Trial to Avoid Data Missing Due to Withdrawal From Exposure*

754  
755 Sometimes patients fail to report for visits, fail to complete questionnaires that contain response  
756 endpoints, or withdraw from assigned treatment prior to planned completion of a clinical trial  
757 without contributing PRO information. The resulting missing data can introduce bias and  
758 interfere with the ability to compare effects in the test group with the control group because only  
759 a subset of the initial randomized population contributes, and these patient groups may no longer  
760 be comparable. Missing data is a major challenge to the success and interpretation of any  
761 clinical trial.

762  
763 The protocol can increase the likelihood that a trial will still be informative by establishing plans  
764 for gathering all treatment-related reasons for patients withdrawing from a trial and by trying to  
765 minimize patient dropouts prior to trial completion. We recommend the study protocol describe  
766 how missing data will be handled in the analysis. It could also establish a process by which PRO  
767 measurement is ascertained before or shortly after patient withdrawal from treatment exposure  
768 due to lack of efficacy or toxicity.

769  
770           **B.       Frequency of Measurements**

771  
772 The frequency of PRO assessment depends on the natural history of the disease and the nature of  
773 the treatment. Some diseases, conditions, or study designs may necessitate more than one  
774 baseline assessment and several PRO assessments during treatment. The frequency of PRO  
775 assessment should correspond with the demonstrated measurement properties of the instrument  
776 and with the planned data analysis.

777  
778           **C.       Duration of Study**

779  
780 It is also important to consider whether the duration of the study is of adequate length to support  
781 the proposed claim and assess a durable outcome in the disease or condition being studied.  
782 Generally, duration of follow-up with a PRO assessment should be at least as long as for other  
783 measures of effectiveness. It should be noted, however, that the study duration appropriate for  
784 the PRO-related study objective may not be the same as the study duration for other study

***Contains Nonbinding Recommendations***  
***Draft — Not for Implementation***

785 endpoints. In a trial for a progressive disease where the PRO concept of interest does not change  
786 until after the follow-up required for other clinical efficacy parameters, longer study duration can be  
787 indicated.

788

789 **D. Design Considerations for Multiple Endpoints**

790

791 The hierarchy of endpoints is determined by the stated objectives of the trial and the clinical  
792 relevance and importance of each specific measure independently and in relationship to each other.  
793 A PRO instrument could be the primary endpoint measure of the study, a co-primary endpoint  
794 measure in conjunction with other objective or physician-rated measurements, or a secondary  
795 endpoint measure whose analysis would be considered according to a hierarchical sequence. The  
796 FDA recommends that the study protocol define the study endpoint measures and the criteria for  
797 the statistical analysis and interpretation of results, including a clear specification of the conditions  
798 for a positive study conclusion.

799

800 **E. Planning for Study Interpretation**

801

802 The FDA recommends that sponsors discuss with the appropriate review division how best to plan  
803 for the interpretation of study findings. In some cases, the FDA may request an *a priori* definition  
804 of the minimum observed difference between treatment group means (i.e., MID) that will serve as  
805 a benchmark to interpret whether study findings are conclusive. In other cases, the FDA may  
806 request an *a priori* definition of a treatment responder that can be applied to individual patient  
807 changes over time. Prespecification of methods for interpretation is particularly important with  
808 new or unfamiliar instruments or when patient dropouts, withdrawals from exposure, or missing  
809 data are expected (e.g., in studies where repeated PRO measurement is planned). See Section VI.E.  
810 for guidance on interpretation considerations for a study's statistical analysis plan.

811

812 **F. Specific Concerns When Using Electronic PRO Instruments**

813

814 When electronic PRO instruments are used, sponsors should plan carefully to ensure that FDA  
815 regulatory requirements are met for sponsor and investigator record keeping, maintenance, and <sup>6</sup>  
816 access. These responsibilities are independent of the method used to record clinical trial data

817 and, therefore, apply to electronic PRO data. Sponsors are responsible for providing investigators  
818 with the information they need to conduct the investigation properly, for monitoring the  
819 investigation, for ensuring that the investigation is conducted in accordance with the  
820 investigational plan, and for permitting the FDA to access, copy, and verify records and reports  
821 relating to the investigation.

822

823 The principal record keeping requirements for clinical investigators include the preparation and  
824 maintenance of adequate and accurate case histories (including the case report forms and  
825 supporting data), record retention, and provision for the FDA to access, copy, and verify records  
826 (i.e., source data verification). The investigator's responsibility to control, access, and maintain  
827

---

<sup>6</sup>For the principal record keeping requirements for clinical investigators and sponsors, see 21 CFR 312.50, 312.58, 312.62, 312.68, 812.140, and 812.145.

## *Contains Nonbinding Recommendations*

*Draft — Not for Implementation*

828 source documentation can be satisfied easily when paper PRO instruments are used, because the  
829 subject usually returns the diary to the investigator who either retains the original or a certified  
830 copy as part of the case history. The use of electronic PRO instruments, however, may pose a  
831 problem if direct control over source data is maintained by the sponsor or the contract research  
832 organization and not by the clinical investigator. The FDA considers the investigator to have  
833 met his or her responsibility when the investigator retains the ability to control and provide  
834 access to the records that serve as the electronic source documentation for the purpose of an  
835 FDA inspection. The FDA recommends that the study protocol, or a separate document, clearly  
836 specify how the electronic PRO source data will be maintained.

837  
838 In addition, the FDA has previously provided guidance to address the use of computerized  
839 systems to create, modify, maintain, archive, retrieve, or transmit clinical data to the agency<sup>7</sup> and  
840 to clarify the requirements and application of 21 CFR part 11.<sup>8</sup> Because electronic PRO data  
841 (including data gathered by personal digital assistants or phone-based interactive voice recording  
842 systems) are part of the case history, the FDA expects electronic PRO data to be consistent with  
843 the data standards described in that guidance. Sponsors should plan carefully to establish  
844 appropriate system and security controls, as well as cybersecurity and system maintenance plans  
845 that address how to ensure data integrity during network attacks and software updates.

846  
847 Sponsors should also plan to avoid the following:<sup>9</sup>

- 848 • Direct PRO data transmission from the PRO data collection device to the sponsor (i.e.,  
849 the sponsor should not have exclusive control of the source document)
- 850 • The existence of only one database without backup (i.e., risk of data corruption or loss  
851 during the trial with no way to reconstitute or verify the data)

This is a standard Part 11 issue and do not think it is necessary to discuss a database and the need for backup. Remove this bullet!

- 852 • Removal of investigator accountability for confirming the accuracy of the data

This is an interesting point but even though the investigator is responsible per 312.62.b it is nearly impossible for the investigator to second guess how a remote patient answered a question about their wellbeing. This should be changed to state that the investigator should be looking for trends in patient responses that could indicate either a problem (AE trigger) or fraud.

- 853 • Loss of adverse event data

This should be revised to state AE Trigger data as PRO does not collect AE's

- 854 • Access to unblinded data

In addition to this a bullet should be added about maintaining patient confidentiality as now you will have technology helpdesk personnel having direct access to patients.

- 855 • Inability of an FDA investigator to inspect, verify, and copy the data at the clinical site  
856 during an inspection

This is straight out of Part 11 and can be removed. A general statement to ensure compliance to Part 11 would suffice.

857 • An insecure system that allows for easily alterable records.

This is straight out of Part 11 and can be removed.

858

---

7 See the draft guidance for industry *Computerized Systems Used in Clinical Trials*. When final, this guidance will supersede the guidance of the same name issued in April 1999 and will represent the FDA's current thinking on this topic. For the most recent version of a guidance, check the CDER guidance Web page at <http://www.fda.gov/cder/guidance/index.htm>.

8 See the guidance for industry *Part 11, Electronic Records; Electronic Signatures — Scope and Application* (<http://www.fda.gov/cder/guidance/index.htm>)

9 The FDA specifically welcomes comment and additional information that will inform these policies as new electronic PRO technology is developed and used in the medical product development setting.

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

859 **VI. DATA ANALYSIS**

860  
861 Incorporating PRO instruments as study endpoint measures introduces challenges in the analysis  
862 of clinical trial data. Some of these challenges are discussed in the following sections.

863  
864 **A. General Statistical Considerations**

865  
866 The statistical analysis considerations for PRO endpoints are not unlike statistical considerations  
867 for any other endpoint used in drug development.<sup>10</sup> We recommend that the principal features of  
868 the planned statistical analysis of the data be described in the statistical section of the protocol  
869 and in a detailed elaboration of the analysis often called the Statistical Analysis Plan (SAP). The  
870 FDA intends to determine the adequacy of study data to support claims in light of the  
871 prespecified method for endpoint analysis. Unplanned or post hoc statistical analyses are usually  
872 viewed as exploratory and, therefore, unable to serve as the basis of a claim of effectiveness.

873  
874 **B. Statistical Considerations for Using Multiple Endpoints**

875  
876 It is important that the study protocol specify all endpoints that will be considered, including  
877 each domain score targeted to support a specific claim. The SAP should describe the planned  
878 primary analysis in detail, noting whether the endpoint will be analyzed as a continuous variable  
879 (mean scores), dichotomous variable (success/failure), or some graded response, the primary and  
880 secondary endpoints, corrections for multiplicity, and the specific statistical methods planned.

881  
882 In some situations, the SAP can specify that two or more variables must be statistically  
883 demonstrated to be superior to control group findings to support a claim. This may be the case,  
884 for example, when a clinician-reported endpoint and a patient-reported endpoint both need to be  
885 shown better than the control. Control for multiplicity (i.e., adjustment of the Type I error)  
886 generally is not a concern when all endpoints are shown to be superior to those of the  
887 comparison group, but we recommend carefully considering the impact of choosing multiple  
888 primary endpoints on Type II error and sample size. The sample size of the trial may be affected  
889 by how many endpoints are measured, the overall strategy planned to integrate all endpoints in  
890 the SAP, and the decision rule for declaring a successful study outcome.

891  
892 Because each PRO item or domain often can represent an endpoint that could imply a distinct  
893 claim on its own, we recommend careful planning to avoid substantial increases in Type I error  
894 from multiple endpoints. If it is important in a study to demonstrate that PROs have the same  
895 directional effect as other measures of treatment benefit, then statistical procedures can be  
896 considered to minimize the impact of multiple endpoint comparisons.

897  
898 There is no single best statistical procedure for multiplicity adjustment because the choice of  
899 procedure depends upon the study objectives, the most important endpoints among the  
900 collection, and other considerations. Some of the statistical procedures that can be useful for a  
901 more efficient analysis approach include methods that prespecify a sequence or order of the

---

<sup>10</sup> See the ICH guidance for industry *E9 Statistical Principles for Clinical Trials*  
(<http://www.fda.gov/cder/guidance/index.htm>)

## ***Contains Nonbinding Recommendations***

*Draft — Not for Implementation*

902 testing or that have a hierarchy of comparisons that first need to be satisfied before others are  
903 considered for testing (i.e., closed testing procedures, gatekeeper strategies). Generally, these  
904 statistical methods are less conservative than the classical Bonferroni or other statistical  
905 multiplicity adjustments that are used to control false positive conclusions from a family of  
906 eligible hypotheses. Another reason to consider less conservative methods is to adjust for what  
907 are often strong correlations among the endpoints (causing a Bonferroni adjustment to be too  
908 conservative). These strategies reduce the need for more stringent statistical tests for the  
909 subsequent endpoints, but do not allow statistical testing for endpoint combinations not  
910 prespecified.

911  
912 A multidomain PRO measure can successfully support a claim based on one or a subset of the  
913 domains measured if an *a priori* analysis plan prespecifies the domains that will be targeted as  
914 endpoints for the study. However, demonstration that only a subset of domains is affected by  
915 treatment (e.g., the physical function domain) generally will not support a general claim (e.g., a  
916 claim of *improved HRQL*) because such a claim implies improvement on all domains that are  
917 important to the general concept. Use of domain subsets as study endpoints presupposes that the  
918 PRO instrument was adequately developed and validated to measure the subset of domains  
919 independently from the other domains.

920  
921 The FDA recommends that the sponsor discuss with the FDA in advance of the study the  
922 appropriateness of the statistical strategies proposed in the SAP.

923

### 924 **C. Statistical Considerations for Composite Measures**

925

926 Understanding the usefulness and measurement properties of a composite endpoint (i.e., an  
927 index, profile, or battery of scores) is an iterative process that evolves over time. Rules for  
928 interpretation of composite measures depend on substantial clinical experience with the measure  
929 in the clinical trial setting. Development of a composite endpoint at the time the confirmatory  
930 clinical study protocol is generated is discouraged unless there is substantial prior empirical  
931 evidence of the value of the chosen components of the composite. Though one reason for use of  
932 a composite is to reduce the multiplicity problems associated with multiple separate endpoints,  
933 composites can do so only if it is agreed that treatment impact on each of the endpoints is of  
934 value and if the endpoints move in the same direction.

935

936 Establishing benefit is difficult if only one component of a composite endpoint responds to the  
937 treatment. For example, a treatment may relieve certain symptoms or improve functioning but  
938 this benefit may not be detected using a composite score that includes other endpoints (e.g.,  
939 psychological or emotional well-being) that fail to improve with the treatment. In any such  
940 composite, it is critical to ensure that patients enrolled in a clinical study are impaired in all  
941 domains (e.g., psychological or emotional well-being) because they cannot improve in domains  
942 if they are not impaired in whatever concept the domain measures.

943

944 Multiplicity problems arise when the multiple individual components of a composite endpoint  
945 are intended as possible claims. In general, individual components of a composite measure will  
946 not be adequate to support a claim unless the components are prespecified in the SAP as separate

## ***Contains Nonbinding Recommendations***

*Draft — Not for Implementation*

947 endpoints, either sharing overall study alpha (co-primary endpoints) or identified in a sequential  
948 analysis, and the study results are found statistically and clinically meaningful in the context of  
949 the total composite and other individual component results.

950  
951 In general, if analysis of scores for the individual component endpoints of a composite shows the  
952 improvement is driven primarily by a single domain (e.g., performance of a specific activity), the  
953 findings for the composite score would not support a general claim (e.g., psychological or  
954 emotional benefit, or even general physical state if all that is shown is symptom improvement).

955

### **D. Statistical Considerations for Patient-Level Missing Data**

956

957  
958 The FDA recommends that the SAP address plans for how the statistical analyses will handle  
959 missing data when evaluating treatment efficacy and when considering patient success or patient  
960 response.

961

#### *1. Missing Items Within Domains*

962

963  
964 At a specific patient visit, a domain measurement may be missing some, but not all, items.  
965 Defining rules that specify the number of items that can be missing and still consider the domain  
966 to have been measured is one approach to handling this type of missing data. Rules for handling  
967 missing data should be specific to each PRO instrument and should usually be determined during  
968 the instrument development and validation process. The FDA recommends that all rules be  
969 specified in the SAP. For example, the SAP can specify that a domain will be treated as missing  
970 if more than 25 percent of the items are missing; if less than 25 percent of the items are missing,  
971 the domain score can be taken to be the average of the nonmissing items.

972

#### *2. Missing Entire Domains or Entire Measurements*

973

974  
975 When the amount of missing data becomes large, study results can be inconclusive. As  
976 described earlier, the FDA encourages prespecified procedures in the study protocol, particularly  
977 when patients discontinue study treatment. Because missing data may be due to the treatment  
978 received or the underlying disease and can introduce bias in the analysis of treatment differences  
979 and conclusions about treatment impact, the FDA encourages sponsors to obtain data on each  
980 patient at the time of withdrawal to determine the reason for withdrawal. When available, this  
981 information can be taken into account in the analysis.

982

983 A variety of statistical strategies have been proposed in the literature and applications to the  
984 FDA to deal with missing data due to patient withdrawal from assigned treatment exposure prior  
985 to planned completion of the trial. No single method is generally accepted as preferred. One  
986 used in the past was to exclude subjects from the analyses if they did not complete the study (i.e.,  
987 *completers' analysis*). This strategy is generally inadvisable because the reason for missing data  
988 can be treatment-related and these patients may not adequately represent the study population.

989

990 Another common, albeit problematic, strategy is to use the last observation available as the *final*  
991 evaluation — usually referred to as last observation carried forward (LOCF). Even though

*Contains Nonbinding Recommendations*  
*Draft — Not for Implementation*

992 LOCF enables every patient randomized to contribute some observation to the analysis, it can be  
993 problematic for the following reasons:

- 994 • If the objective of the trial is to detect a treatment effect after a certain duration of  
995 treatment (e.g., at 8 weeks), then a comparison that includes only measurements on  
996 patients at earlier times or visits is not addressing the original trial objective. The  
997 average of patient responses, many of which are at different times or visits, may be  
998 uninterpretable.
- 999 • LOCF makes an implicit assumption that the patient would sustain the same response  
1000 seen at an early study visit for the entire duration of the trial. This assumption is  
1001 untestable and potentially unrealistic.

1002 Some other approaches involve imputation of missing data on a per-patient basis. These strategies  
1003 try to predict missing outcomes for a patient who has withdrawn from the trial using data from  
1004 subjects who stayed in the trial and for whom all data have been collected. All of these strategies  
1005 are imperfect, as they involve strong or weak assumptions about what caused data to be missing,  
1006 assumptions that usually cannot be verified from the data. If missing data are associated with  
1007 treatment effect in ways that cannot be predicted from measurements on subjects with complete  
1008 data, analyses using imputation procedures will be biased. When there are few patients with  
1009 missing measurements and the frequency of missing data or proportion of patients with missing  
1010 data is comparable across treatment groups, most approaches will yield similar results. When a  
1011 higher proportion of patients have missing data, the FDA recommends the use of several different  
1012 imputation methods (including a worst-case scenario in which missing data are assumed to be  
1013 unfavorable for those on the investigational treatment and favorable for those in the control group)  
1014 and an assessment of the consistency of the study results using each method. These analyses will  
1015 demonstrate the sensitivity of the conclusions to the assumptions made by the different methods.  
1016

1017 **E. Interpretation of Study Results**

1018  
1019 Because statistical significance can sometimes be achieved for very small changes if a study is  
1020 large enough, it is tempting to identify an MID as a benchmark for interpreting the clinical  
1021 importance or relevance of study results. If the MID is truly to be the smallest effect considered  
1022 meaningful, however, it would be logical to establish the null hypothesis to rule out a difference  
1023 less than or equal to the MID. This is rarely done, and would have major implications for sample  
1024 size.  
1025

1026  
1027 When clinical trials show small mean effect sizes, rather than considering results in terms of an  
1028 MID, it may be more informative to examine the distribution of responses between treatment  
1029 groups to more fully characterize the treatment effect and examine the possibility that the mean  
1030 improvement reflects very different responses in subsets of patients. When only a modest fraction  
1031 of people respond to a treatment, that fraction may experience meaningful change in the face of a  
1032 mean effect that is very small. When defining a meaningful change on an individual patient basis  
1033 (i.e., a responder), that definition is generally larger than the minimum important difference for  
1034 application to group mean comparisons.  
1035  
1036

**GLOSSARY**

1037  
1038

1039 **Claim** — A statement of treatment benefit or comparative safety advantage. A claim can appear 1040  
in any section of a medical product’s FDA-approved label or in advertising of prescription drugs. 1041

1042 **Cognitive debriefing** — A qualitative research tool used to determine whether concepts and  
1043 items are understood by patients in the same way that instrument developers intend. Cognitive  
1044 debriefing interviews involve incorporating follow-up questions in a field test interview to gain a  
1045 better understanding of how patients interpret questions asked of them.  
1046

1047 **Concept** — The specific goal of measurement (i.e., the *thing* that is to be measured by a PRO  
1048 instrument).  
1049

1050 **Conceptual framework** — The expected relationships of items within a domain and of domains  
1051 within a PRO concept. The validation process confirms the conceptual framework. When used  
1052 in a clinical trial, the observed relationships among items and domains will again confirm the  
1053 conceptual framework.  
1054

1055 **Domain** — A domain is a discrete concept within a multidomain concept. All the items in a  
1056 single domain contribute to the measurement of the domain concept.  
1057

1058 **Health-related quality of life (HRQL)** — A multidomain concept that represents the patient’s  
1059 overall perception of the impact of an illness and its treatment. An HRQL measure captures, at a  
1060 minimum, physical, psychological (including emotional and cognitive), and social functioning.  
1061 Claiming a statistical and meaningful improvement in HRQL implies: (1) that the instrument  
1062 measures all HRQL domains that are important to interpreting change in how the study  
1063 population feels or functions as a result of treatment; and (2) that improvement was  
1064 demonstrated in all of the important domains. An HRQL instrument is a particular type of PRO  
1065 instrument.  
1066

1067 **Instrument** — A means to capture data (e.g., questionnaire, diary) plus all the information and  
1068 documentation that supports its use. Generally, that includes clearly defined methods and  
1069 instructions for administration or responding, a standard format for data collection, and well-  
1070 documented methods for scoring, analysis, and interpretation of results.  
1071

1072 **Item** — An individual question, statement, or task that is evaluated by the patient to address a  
1073 particular concept.  
1074

1075 **Minimum important difference (MID)** — The amount of difference or change observed in a  
1076 PRO measure between treatment groups in a clinical trial that will be interpreted as a treatment  
1077 benefit.  
1078

1079 **Patient-reported outcome (PRO)** — Any report coming directly from patients (i.e., study  
1080 subjects) about a health condition and its treatment.  
1081

**Contains Nonbinding Recommendations**  
*Draft — Not for Implementation*

1082 **Quality of life** — A general concept that implies an evaluation of the impact of all aspects of life  
1083 on general well-being. Because this term implies the evaluation of nonhealth-related aspects of  
1084 life, it is too broad to be considered appropriate for a medical product claim.

1085

1086 **Questionnaire** — A set of questions or items shown to a respondent in order to get answers for  
1087 research purposes.

1088

1089 **Scale** — The system of numbers or verbal anchors by which a value or score is derived.

1090 Examples include visual analogue scales, Likert scales, and rating scales.

1091

1092 **Score** — A number derived from a patient's response to items in a questionnaire. A score is  
1093 computed based on a prespecified, validated scoring algorithm and is subsequently used in  
1094 statistical analyses of clinical study results. Scores can be computed for individual items,  
1095 domains, or concepts, or as a summary of items, domains, or concepts.

1096

1097 **Treatment benefit** — An improvement in how a patient survives, feels, or functions as a result of  
1098 treatment. Measures that do not directly capture the impact of treatment on how a patient  
1099 survives, feels, or functions are surrogate measures of treatment benefit.

1100

1101 **Validation** — The process of assessing a PRO instrument's ability to measure a specific concept

This Term should be changed to **Psychometric Validation** as the Validation term is well defined within ISO and should not be re-used.

1102 or collection of concepts. This ability is described in terms of the instrument's measurement  
1103 properties that are derived during the validation process. At the conclusion of the process, a set  
1104 of measurement properties is produced that are specific to the specific population and the  
1105 specific form and format of the PRO instrument tested. The validation process involves:

- 1106 • Identifying the concept to be measured
- 1107 • Assessing the content validity (i.e., being sure the items in the questionnaire cover all  
1108 important aspects of the concept from the patient perspective)
- 1109 • Evaluating the proposed scores to be obtained from the instrument
- 1110 • Defining *a priori* hypotheses of the expected relationships between PRO concepts and  
1111 other measures
- 1112 • Testing the hypotheses by reporting the observed correlations among scores