

**ISPOR****International Society for Pharmacoeconomics and Outcomes Research**3100 Princeton Pike, #3-E, Lawrenceville, NJ 08648 USA • Tel: 1-609-219-0773 • Fax: 1-609-219-0774  
Email: info@ispor.org Internet: www.ispor.org

April 4, 2006

Division of Dockets Management (HFA-305)  
Food and Drug Administration  
5630 Fishers Lane, Room 1061  
Rockville, MD 20852

**Re: Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [Docket No. 2006D-0044]**

Dear Captain Burke:

Thank you for the opportunity to comment on the proposed FDA Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) is an international organization promoting the science of pharmacoeconomics and health outcomes research and is organized to act as a scientific leader relevant to research in pharmacoeconomics, health outcomes assessment including patient-reported outcomes, and related issues of public policy. The Society represents healthcare researchers and practitioners including pharmacists, physicians, economists, nurses and researchers from academia, pharmaceutical industry, government, managed care, health research organizations, and purchasers of healthcare.

More than 700 ISPOR members are interested in patient-reported outcomes (PRO); and nearly 50 members provided comments and suggestions to this Patient-reported Outcomes Guidance. The ISPOR members commenting on this guidance came from academia (36%), research organizations (22%), and pharmaceutical/medical device/diagnostic/biotech industry (42%). Comments were received from 12 countries, indicating the world-wide effect of this FDA PRO guidance on PRO research. This summary was developed by the ISPOR PRO Special Interest Group (SIG).

Overall, The Guidance is a comprehensive document reflecting the current state of the art of PRO questionnaire development, validation, and implementation in clinical research and analysis. The FDA PRO Guidance makes a major policy statement concerning the role of the patient, and patient reported outcomes, in the drug approval process. ISPOR applauds the FDA's definition of "treatment benefit" as "An improvement in how a **patient** survives, feels, or functions as a result of treatment." This focus will provide a balance between the traditional biologic and physiologic markers and outcomes that directly measure impact to the patient, as reported by the patient.

However, the "gold standard" proposed by the guidance for questionnaire development may be very difficult to achieve for every instrument, particularly multi-dimensional instruments. The desire to comprehensively collect patient reported outcomes in clinical trials, through the use of developed and validated instruments, may be thwarted when the previous validation procedures do not meet the proposed FDA "gold standard" level. The intent of the guidance to encourage more comprehensive and careful inclusion of PRO may paradoxically result in a reliance on more limited patient input through the use of one-dimensional, symptomatic measures.

As the FDA clarifies its standards for PRO measures, ISPOR strongly suggests that it also examine its requirements for the psychometric performance of other endpoint measures such as clinical and biologic markers and physician/caregiver reports. In this way consistency of criteria will be ensured across measurement methods.

A summary of comments by section are given in the attached document, which includes specific comments by section, line or page, and person commenting as well as institution.

The following are key issues to be addressed in the FDA PRO Guidance:

- **Issue One:** the guidance states ideal requirements which may be difficult to meet in every case. Practical or non-consensual methodological issues such as the determination of the MID are not discussed. We strongly recommend inclusion of additional language clarifying that review of PRO data and supporting information about PRO measures will be based on the body of evidence provided and that the FDA recognizes that an acceptable level of evidence may include less than the full set of criteria outlined as best practice in the Guidance.
- **Issue Two:** considering any modification of an instrument as a new instrument may be too strict in the majority of cases. Minor modifications or translation of instruments should not be considered as new and therefore should not require a complete “revalidation. We recommend inclusion in the Guidance of examples of the types of evidence, short of full validation, that could be acceptable for modifications of existing instruments (e.g., cognitive interviews with patients).
- **Issue Three:** the determination of the MID is given prominence in the discussion of interpretation of results. The current lack of consensus in the field regarding the value of MID and the best methods for establishing it make emphasis on MID within the Guidance inappropriate at this time. We recommend inclusion of language to indicate that MID is just one of many methods to aid interpretation and that choice of methods should match individual submission requirements.

Finally, we would urge that FDA consider the addition of language to emphasize that the document reflects ideal requirements which have to be evaluated in light of methodological and practical issues.

ISPOR would be interested in providing a forum to address these specific issues.

Thank you again for this opportunity. If you have any questions regarding our comments, please contact Marilyn Dix Smith, Ph.D., ISPOR Executive Director at [mdsmith@ispor.org](mailto:mdsmith@ispor.org).

Sincerely,



Patrick Marquis MBA, MD, *Chair, ISPOR PRO / QOL Information in Regulatory and Health Decisions Working Group, PRO SIG & Managing Director, Mapi Values, Boston, MA, USA*



Judith Barr MEd, ScD, *Chair, ISPOR PRO SIG & Associate Professor and Director, National Education and Research Center for Outcomes Assessment, Northeastern University, Boston, MA, USA*

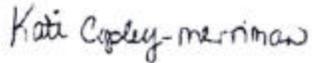


Stephen Joel Coons PhD, *Chair, ISPOR ePRO Working Group, PRO SIG & Professor, College of Pharmacy, The University of Arizona, Tucson, AZ, USA*



Diane Wild MSc, *Chair, ISPOR Cross-Cultural & Translational Adaptation (CTA) Working Group, PRO SIG & Partner, Oxford Outcomes, Oxford, UK*

**ISPOR PRO / QOL Information in Regulatory and Health Decisions Working Group, PRO SIG members:**



Kati Copley-Merriman MS, MBA, *Group Leader, WWOR Oncology, Pain and Inflammation, Pfizer Global Pharmaceuticals, Ann Arbor, MI, USA*



Lori Frank PhD, *Senior Research Scientist, Deputy Director, Center for Health Outcomes Research United BioSource Corporation, Bethesda, MD, USA*



Ari Gnanasakthy MBA, MSc, *Head of Patient Reported Outcomes Research, Novartis Pharma, East Hanover, NJ USA*



Clarice Hayes, *Health Outcomes Researcher, Eli Lilly & Company, Indianapolis IN USA*



Ruslan Horblyuk MBA, *Health Outcomes Manager, GlaxoSmithKline, Philadelphia, PA USA*



Ravishankar Jayadevappa, PhD, *Research Assistant Professor, University of Pennsylvania, Philadelphia, PA, USA*



Karen Lee, MA, *Health Economist, Canadian Agency for Drugs and Technologies in Health (CADTH), On, Canada*

[The signers of this letter represent the ISPOR PRO SIG and do not necessarily reflect the views of all of the ISPOR membership]

cc: Marilyn Dix Smith RPh, PhD, ISPOR Executive Director [Email: [mdsmith@ispor.org](mailto:mdsmith@ispor.org)], and ISPOR Board of Directors

**ISPOR Comments on:  
FDA Draft Guidance for Industry – Patient-Reported Outcome Measures:  
Use in Medical Product Development to Support Labeling Claims**

**ISPOR COMMENTS SUMMARY**

**I. INTRODUCTION – Specific language suggestions given on page 7 of this response; no general comments**

**II. BACKGROUND – Specific language suggestions given on page 7 of this response; no general comments**

**III. PATIENT-REPORTED OUTCOMES – REGULATORY PERSPECTIVE – Specific language suggestions given on page 7 of this response; no general comments**

**IV. EVALUATING PRO INSTRUMENTS**

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

1. It should be noted that, with the PRO Guidance Document, the FDA has provided a “gold standard” for instrument development. However, somewhere in the document, the FDA should acknowledge that it is unrealistic for any PRO measure to achieve this standard. Moreover, the FDA should not encourage the development of new instruments, but should encourage standardization of instruments. For existing instruments, however, the FDA should specifically delineate what changes (e.g., recall periods, cultural adaptation) require a new validation process. A requirement of a validation study for minor changes to an existing instrument should not be necessary.
2. All figures, tables, and diagrams should be accompanied with references.
3. The discussion of the use of a single item PRO is confusing because it appears to be an academic discussion without a clear guidance. Further on, as the FDA discusses multi-domain instruments, some PRO-relevant concepts like social functioning may be measured appropriately by a multi-domain instrument without a targeting of specific domains.
4. The designation of the expected relationship among PRO items and domains, such as is depicted in Figure 2, may be difficult prior to the validation process and question the FDA’s use of such information.
5. We acknowledge the importance of identification of the intended population, but the comparison of the population used for instrument development to the study populations on all of the suggested variables (age, sex, ethnic identity, and cognitive ability) to be too strict. With respect to cognition, the FDA should specify the test that it would deem appropriate for determining the cognitive abilities of patients. In addition, disease severity should be added to the list of variables compared.

**B. Creation of the PRO Instrument - Comments**

1. The FDA should encourage standardization and discourage a proliferation of instruments. The Agency should view instruments developed ad hoc for clinical trials with extreme suspicion.
2. The idea that PROs will be assessed for their relevance to the study population (age, sex, ethnic identity and cognitive ability) as well as for minor changes (e.g. recall period) is too strict. Validated instruments should not have to be re-created in light of this Guidance and minor changes and inappropriate use of instruments only inflate the Type II error.
3. Somewhere in the Guidance should be a section on existing instruments and standards for assessing their use in labeling.

### **C. Assessment of Measurement Properties-Comments**

1. The concept and practice of establishing MID causes many concerns. One key point is that MID is just one of many concepts useful for aiding interpretation and its mention in the guidance document gives the impression that it is a required property to establish. As there is not consensus in the field for methods for determining MID (e.g., anchor-based vs. distribution-based methods), the emphasis given to establishing MID seems premature relative to other key aspects of psychometric performance to demonstrate. In addition, the mention of MID should include distinction between patient-based MID determination and clinician-based MID determination. Similar concerns were raised regarding the concept of "responder." Respondents look forward to increased specificity.
2. While there was agreement that assessment of measurement properties is necessary, a listing of measurement properties (e.g., Table 4) leaves the impression that the burden is on the applicant to demonstrate each of these for each PRO. This is not feasible and not consistent with requirements for acceptance of non-PRO measurement data. Addition of language to emphasize that this section reflects an ideal rather than a practical requirement would be helpful, as would language indicating that actual practice will usually involve collection of a core subset of data on psychometric performance.
3. Further guidance on requirements regarding establishing measurement properties of translations would be also welcomed.
4. Public access to newly developed instruments will aid optimizing collection of psychometric data.
5. The meaning of validity could be explored further within the guidance. Validity and reliability of broad QOL domains is likely lower than for more specific concepts.
6. Language could be added to further encourage concept specificity. The meaning of validity could be explored further within the guidance. Mention of scale properties and consequences for statistical testing should be included in this section.
7. It is desirable to encourage usage of commonly used instruments when applicable. The FDA should encourage standardization of instruments, process of administration (self-report vs. proxy) and methods (e.g., handling of missing data)
8. Identification of targeted populations is crucial for any outcome assessments so the focus on measurement specificity within the document is appreciated.
9. The document implies that test-retest reliability is preferable to obtain relative to internal consistency reliability. Language should clarify the distinct purposes of each type of reliability and the importance hierarchy should be reexamined.

### **D. Modification of an Existing Instrument-Comments**

1. Modified instruments: the fact that a modified instrument would be considered as a different instrument is a concern. A minor change should not require a new complete "validation".
2. Translated versions: the evidence that measurement properties for translated versions are comparable raised the same concern. A proper translation process based on accepted methods including a cognitive testing should provide sufficient evidence of the conceptual equivalence.
3. Instruments included in a battery: when stand alone instruments are included in a battery, an order effect can be observed but as long as the order of the instruments remains the same within the clinical trial additional validation should not be required.

### **E. Development of PRO Instruments for Specific Populations-Comments**

1. An important point that has not been addressed in this section is that the PRO measures/instruments should be culturally sensitive and validated across targeted ethnic, gender and age groups.

## V. STUDY DESIGN-Comments

1. More detail is needed regarding methods for handling missing data and sample size determination, particularly for the instrument development process. Perhaps, the section on electronic data capture can be reduced and instead reference existing FDA guidances on this topic.

## VI. DATA ANALYSIS-Comments

1. Multiplicity - There are concerns regarding whether secondary endpoints are sufficient to support a claim and if superiority should be the objective. FDA should clarify whether superiority and powering of the study should be based on MID, taking potential multiplicity into account when needed.
2. Composite Measures - FDA stated that patients enrolled in a clinical study should be impaired in all domains of a multi-domain instrument being used in the trial: This point is valid for composite measures, but not if the *a priori* intent is to report selected domains scores and use them for label claim. This concept should be clarified.
3. Missing data - One member provided extensive comments regarding missing data (see appendix). In summary, it was felt that several missing data issues should be more completely addressed, such as imputation of missing data at baseline. In addition FDA should supply recommendations for the primary analysis, such as based on “missing at random” (MAR) assumptions with sensitivity analysis based on “missing not at random” (MNAR) assumptions using pattern mixture models. It is agreed that multiple imputation methods are preferable to single imputation methods or worst case/best case scenarios.
4. Minimal important differences (MID)-FDA should clarify whether MID superiority is needed to obtain a label claim, only statistical significance or both. Substantial clinical experience is needed to establish interpretation rules and this may evolve over time for new instruments.

## OTHER CONSIDERATION TO THE FDA PRO GUIDANCE:

FDA should consider the value of establishing a separate body comprised of representatives from government, academia, and industry, who could assist with review of supporting data for PRO measures and ensure consistency of evidence review. This group could issue a “seal of approval” for use of specific PROs with specific populations. In some cases the group could aid with instrument development documentation submitted to the FDA. For specific regulatory submissions, the burden of evaluating endpoint evidence, whether from PROs or not, would still rest with the FDA.

**ISPOR Member Responses to:  
Draft FDA PRO Guidance for Industry Patient-Reported Outcome Measures:  
Use in Medical Product Development to Support Labeling Claims**

Sub-Section /Line/Page	COMMENT
<b>I. INTRODUCTION</b>	
31	PRO instruments are used to support the effectiveness of drugs and medical products because they can incorporate the patients' perspective alongside traditional physiological and functional assessments, assessing both observable and un-observable health benefits.
31-32	Proxy measures should be mentioned (children, cognitive impaired patients etc.)
45-52	The use of a concrete instrument would help to explain the point in this paragraph
	Instead of a somewhat vague background information, real precise definitions and examples would be more useful; the definition is hided within the introductory section not speaking out clearly what is given in Glossary for example at the very end of the document the position statement of the FDA is missing; do they wish / are they interested in the sponsors to increase the use of PRO instruments? and when ? ; a section givin, for example details of the FDA's experience with the advantage of using PRO instruments would be helpful if the idea is really to guide the industry, positive example would be helpful for both sides
<b>II. BACKGROUND</b>	
78-79	This sentence would exclude any instruments used to measure direct preferences such as standard gamble since these instruments are not specific to any population or to characteristic of any condition/disease treated. I guess this sentence needs to be reworked to include ALL PRO instruments.
89	Add a comment that is it possible to get a symptom claim without having to show how that translates to other specific endpoints. It should not be a requirement that this translation is measured. Suggested alternative language (if applicable): It is not required that symptoms improvements are translated into other benefits however, in order to get a symptom label claim.
<b>III. PATIENT-REPORTED OUTCOMES-REGULATORY PERSPECTIVE</b>	
A. 99	Should read “ ....;or/and (3) ....”
A.2 114-118	Another very good example in this case is hypertension. Typically patients “feel” worst/sicker when taking their medication to control their blood pressure than when not taking their medication.
A. 3 123	The term “well being” would be more appropriate than the word “function”
154	The statement that PRO instruments that are used in clinical trials to support effectiveness claims should measure adverse consequences separately should be deleted. In some cases, a concept (e.g. fatigue) could be either a disease to treatment consequence and this would be impossible to separate. In addition, there are well validated tools that measure these together and they should still be allowed to support claims. Suggestion: Delete sentence
(line 166-7)	as events occur? At least at baseline and end of treatment. Crucial here that frequency of instrument administration is not leading to an unacceptable burden to the patient.
@ III. A	Why use PRO→ would it be possible for the FDA to give more a detailed support to develop and use PROs? eg by giving positive examples, links to useful internet pages, real recommendations. This section is more like a second part of a definition, it's explaining what a PRO is able to provide versus other methods of data collection in clinical trial about patients, but it is not giving a real reason why sponsors should use PROs.
<b>IV. EVALUATING PRO INSTRUMENTS</b>	
A	
	In general it was felt that the guidance provides clarity as to what can be expected from the FDA in relation to label claims. Throughout the document it is stated what the FDA plans to review, though very often words as "usually and generally" are used which introduces uncertainty as to what is important in a specific case
	It should be noted that there are a number of generic PRO instruments, such as the EuroQOL, Quality of Wellbeing, Health Utilities Index, and SF-36 (and related instruments) that have been subject to years of validation and verification in myriad populations. The FDA should encourage standardization and discourage a proliferation of instruments. Many of these generic instruments also have disease specific components. The use of standard measures assists in building a large database of values associated with specific health states and lends to comparability of assessments across studies. The Agency should view instruments developed ad hoc for clinical trials with extreme suspicion and so state.
	In summary, the reporting of outcomes directly from patients is very valuable. The draft guidance on patient-

	reported outcomes acknowledges and describes the benefits well.
	This report is a technically excellent report about PRO instruments. PRO instruments are used to support the effectiveness of drugs and medical products because they can incorporate the patients' perspective alongside traditional physiological and functional assessments, assessing both observable and un-observable health benefits.
	<p>As a participant at the meeting this past weekend I am heartened that the FDA may actually heed the concerns voiced. HOWEVER, it is important to note that the draft document is written like a text book on how to create the perfect PRO instrument and how to conduct the perfect PRO analysis. Other measures, like physical measures (BP or weight) or toxicity measures (NCI Common Toxicity Criteria), are not held to the SAME rigor as this document suggests PROs should be held. It was the consensus of the 400 people in the room (as assessed by a show of hands) that this document is out of balance with standards or guidance for the measure of ANY OTHER outcome instrument. At the bottom of this email is a small sample of abstracts documenting the low to moderate reliability of physical measures that are generally accepted for outcomes measurement without being held to the rigor the draft document proposes for PROs.</p> <p>There is no known PRO measure, in oncology or in any other field, that has been developed using every method listed in the guidance. Not one of the drugs that have to date gained approval based on PRO claims would have received approval if this document was in place and used as a "checklist" for approval.</p> <p>Further, although the document specifically states "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." it is clear by the attendance of so many NCI folks, including almost all of DCP, that this document will inform how NCI (including DCP and CTEP) review PRO study aims.</p>
	It would be helpful to have a statement early on in the guidance that the document applies only to regulatory application of PRO research, and to acknowledge that there are broader issues for use of PROs in clinical and research contexts that are not expressly addressed in this document.
	<p>An important objective of this report is to address the comparability between studies and to achieve a "gold standard" in using the PRO measures. However, this has received lesser attention, especially, uniformity, reproducibility and transparency of methods and results need more emphasis.</p> <p>Also, it would be helpful to identify if PRO is measured as a primary or secondary endpoint? This identification has important implications for product development and analysis.</p>
	References need to be provided for all definitions, diagrams, figures and tables. The concepts, definitions and processes presented are not novel and have originated from the field of psychometric/survey research and have been widely documented within measurement theory, sociology, psychology and education literature and it would be ethical to provide references of the various sources wherever appropriate.
	Why is this guidance limited to labeling claims? It would be worthwhile to include a section related to what is required (to demonstrate) to get appropriate wording in the clinical trial section of the product labeling.
	In general it was felt that the guidance provides clarity as to what can be expected from the FDA in relation to label claims. Throughout the document it is stated what the FDA plans to review, though very often words as "usually and generally" are used which introduces uncertainty as to what is important in a specific case.
p. 7:	The section on development of the conceptual framework is extremely important and an important contribution for the document. See also Glossary, line 1049 -- confirm vs. support - as it is always empirical and non-observable, the definition should state that the validation process supports rather than confirms the conceptual framework.
1, p. 8, lines 227 etc:	I do not fully agree with the FDA position on mixing constructs within an instrument. The acceptable match between claim and measure used to support it is an important topic and specificity is always desirable. However, some PRO-relevant concepts like social functioning or psychological well-being may be appropriately measured by a multidomain instrument and the relative importance of component domains may vary across individuals.
Line 179	Evaluation of instrument modifications as new instruments is not always appropriate, as others have noted.
	For determination of what constitutes "good enough" evidence -- A SEPARATE BODY SHOULD BE FORMED -- for recommendations and evidence review. In the past AHRQ has been considered as a potential locus for this type of review – this should be considered anew.
Line 227 (pg 8)	Multi-domain [was: multidomain] PRO instruments.
Line 229 (pg 8)	...The complex nature of multi-domain [was: multidomain] PRO instruments...
Line 235 (pg 8)	...all evidence based on multi-domain [was: multidomain] PRO instruments...
Line 277 (pg 9):	...appropriate to these populations [was: that population] with respect to...
	References need to be provided for all definitions, diagrams, figures and tables. The concepts, definitions and processes presented are not novel and have originated from the field of psychometric/survey research and have been widely documented within measurement theory, sociology, psychology and education literature and it would be ethical to provide references of the various sources wherever appropriate.

	Not clear as to what types of documentation (e.g. publication etc) are acceptable to demonstrate proven validity for a pre-developed instrument. It would help provide clarity by providing a list of acceptable documentation types.
(e.g. 178-181, 579-670)	At several places it is mentioned that small changes to the instrument or instructions are viewed by the agency as new versions and will lead to a complete review of all evidence on psychometric properties. Given that translation of instruments and instructions often lead to slight revision of these documents, it is not realistic to require full validation of revised instruments.
(275-279, 478-480)	Much attention is paid to the validation population and final population to justify a claim, focusing on age, sex, ethnicity, cognitive ability and socioeconomic status. Also here only relevant characteristics (e.g. disease severity) should be mentioned.
	It should be noted that there are a number of generic PRO instruments, such as the EuroQol, Quality of Wellbeing, Health Utilities Index, and SF-36 (and related instruments) that have been subject to years of validation and verification in myriad populations. The FDA should encourage standardization and discourage a proliferation of instruments. Many of these generic instruments also have disease specific components. The use of standard measures assists in building a large database of values associated with specific health states and lends to comparability of assessments across studies. The Agency should view instruments developed ad hoc for clinical trials with extreme suspicion and so state.
172-176	I cannot agree more!
178	Considering any modification to be held to the same standards as a new instrument is too strict. Some modifications are very minor and should not require extensive revalidation. Suggestion: Delete sentence- next sentence is more appropriate.
275	The idea that FDA will compare the patient populations used in the PRO instrument development process to the study populations with respect to <i>age, sex, ethnic identity and cognitive ability</i> is far too strict. Study trials themselves are a proxy for the eventual population who will be using our drugs. It would be impossible to comply with this requirement for other endpoints. Suggestion: The FDA plans to compare the patient populations used in PRO instrument development process to the study populations enrolled in clinical trials to ensure that they are similar.
Ad3. p.9, ll275-279:	Identification of intended population-very important point could be critical for many PRO analysis, especially “cognitive ability” mentioned in line 277-278. How we should determine the cognitive ability of the patient? Specific test? Who should create that kind of test in different countries or for different cultural groups? I think it is a serious problem and should be resolved.
Part B-p.9-items	This is very delicate operation and really needs the detailed population studies in the pragmatic and empathic way. My work in prenatal sector during 30 years was enormous source.
→ Hence: @ section IV. A – D	on the other hand: a very detailed information what the FDA will review and question if PRO instruments are used is given ⇒ the result of this guidance to industry might be a clear a tendency by any sponsor to avoid the use of any PRO instrument, especially in the current early phase of developing PRO instruments and integrating them into CUTS this section is nearly prohibitive to the use of PRO instruments.
page 6; line 178-181	How to determine what size of the modification will require a further validation?
Page 9; line 275-278.	This could become an issue when more and more studies are being conducted in e.g. Asia, while the PRO was developed in e.g. UK. Even though new instruments today are being simultaneously developed in several countries, not all parts of the world will be included in the development phase.
#1	A lack of correlation with a single item can correctly identify that the single item is not capturing all elements of the domain.
#2	PROs are not to be used for adverse event reporting.
III (line 166-7)	As events occur? At least at baseline and end of treatment. Crucial here that frequency of instrument administration is not leading to an unacceptable burden to the patient
IV (page 8):	What is the FDA perspective on a built in adherence subscale?
IV (line 277):	What the FDA perspective on use of PROs in cognitive disabled patient populations?
Lines 212 to 225	Present a rather hypothetical discussion of single item PROs. Lines 218-219 talk about “single item questions” – a phrase could be improved. The paragraph seems strange since it starts off talking about a reliable, validated single question PRO instrument, but then proceeds to discuss when a single item isn’t enough. The latter discussion seems academic. These two concepts need separation.
Lines 249 to 256	Suggest that expected relationships among PRO items and domains need to be diagrammed before the validation process. That may be difficult to do at the least, and the need and utility are unclear.
B	
	It looks like to FDA is too ambitious in reviewing the distribution of item responses to verify the response options (413/4) and the weighing of items (424/5).
	It should be noted that there are a number of generic PRO instruments, such as the EuroQol, Quality of

	Wellbeing, Health Utilities Index, and SF-36 (and related instruments) that have been subject to years of validation and verification in myriad populations. The FDA should encourage standardization and discourage a proliferation of instruments. Many of these generic instruments also have disease specific components. The use of standard measures assists in building a large database of values associated with specific health states and lends to comparability of assessments across studies. The Agency should view instruments developed ad hoc for clinical trials with extreme suspicion and so state.
	We have reviewed the document and congratulate the staff at the Food and Drug Administration and its consultants for producing and circulating the Draft Guidance document. The content of the Draft Guidelines represents an important step in promoting the use of reliable, valid, responsive, and interpretable measures of health-related quality of life for generating evidence about the effects of pharmaceuticals to assist in making evidence-based decisions on the appropriate use of pharmaceutical products. We also thank ISPOR for giving us the opportunity to voice our comments.
275	The idea that FDA will compare the patient populations used in the PRO instrument development process to the study populations with respect to <i>age, sex, ethnic identity and cognitive ability</i> is far too strict. Study trials themselves are a proxy for the eventual population who will be using our drugs It would be impossible to comply with this requirement for other endpoints. Suggestion: The FDA plans to compare the patient populations used in PRO instrument development process to the study populations enrolled in clinical trials to ensure that they are similar.
296	Item generation is frequently done with a fairly small number of patients and it would be very burdensome to expect a wide range of ages and severity and population variations to be studied. Suggestion: Delete the word "wide".
351	VAS- editorial comments should be deleted Suggestion: Delete: These scales often produce a false sense of precision.
416	Implies that we have to determine weights for all domain scores. This paragraph does not add anything to the Guidance- it's philosophical. Suggestion: Delete paragraph from 416-422
470	Somewhere in the Guidance should be a section on existing instruments and standards for assessing their use in labeling. I don't know if that should come at the end of this section, or somewhere else. Validated instruments should not have to be re-created in light of this PRO guidance.
480	Table 4 should be described as taxonomy. Not all of these measurement properties will be examined for every instrument (e.g. predictive validity. Suggestion: Change sentence on 478 to read: "Table 4 outlines the taxonomy of measurement properties that could be studied for a new instrument."
B.4	Table 2. Under the "Description" of the "Visual Analog Scale (VAS)", the last sentence refers to opinion and therefore should be removed.
p.9	Items ----this is very delicate operation and really needs the detailed population studies in the pragmatic and empathic way. My work in perinatal sector during 30 years was enormous source.
p.10	-Recall period-should be determined from the clinical point of view and regarding the scientific data about the drug or method efficiency and efficacy expression.
p.11	I do not agree that the VAS scale often produce a false sense of precision. From my research it gives a very good result-re-test analyze*.
p.12	Cognitive debriefing is interesting but some "control" questions should be done to test the patients "credibility".
P.14	FDA examines the development history of the instrument??? The creative capacity and correct psychometric application is not measurable!
	There is no known PRO measure, in oncology or in any other field that has been developed using every method listed in the guidance. Not one of the drugs that have to date gained approval based on PRO claims would have received approval if this document was in place and used as a "checklist" for approval. It is fine to list the "text book" version on how to develop PRO instruments and how to study PROs HOWEVER language is needed that qualifies how to continue sound research with the current state of the science. The current state of PRO science does not provide instruments developed using all of the criteria mentioned in the guidance, nor could this be done any time in the near future. If any entity could develop instruments with such rigor, it would only be pharma, since they are the only ones with such resources, but even with pharma, this would take tremendous time and resources and research should not cease until the perfect world of PRO instrumentation is attained (if it could ever be). The problem is that once pharma is forced to put the resources into instrument validation, there is no mandate for them to share the tremendous work. Pharma should be in some way encouraged and REWARDED for putting such work in the public domain. The NCI sponsored cooperative groups and individual investigators do not now, nor will they ever, have the resources for instrument development and validation put forth in this document. This in no way should be a barrier to sound PRO research in the cooperatives or by individual

	investigators. When not in conflict with copyright laws, there should be public databases of PRO instruments (as there are now) but with the addition of a comprehensive compendium of validation work.
page 10, line 334-337	For paper diaries that are to be filled out at home, there is no way of determining when the items in the diary were answered. Should this be interpreted that only electronic diaries are applicable?
page 10; line 342;	there are several occasions when asking the patient to only describe their current status instead of referring to e.g. the past week, could result in loss of useful information .
.	This draft guidance by the FDA "... describes how the FDA evaluates patient-reported outcome (PRO) instruments used as effectiveness endpoints in clinical trials" and "... describes [the FDAs] current thinking on how sponsors can develop and use study results measured by PRO instruments to support claims in approved product labeling", but it does not at all give any statement of the FDA
IVB.4, line 378:	I agree that evaluation of patient understanding is important and would recommend inclusion of modifications to directions and response scale as possible outcomes of cognitive debriefing as well (not just item modification).
IVB.7, p.13: I	Appreciate that the FDA will consider if response choices represent appropriate intervals but basing this evaluation on review of item distributions is not entirely logical. Sample characteristics may inform distribution of responses and may not reflect true interval appropriateness.
IVB.7, p.13, line 429:	I recommend expanding discussion of the appropriate application of population-specific preference weights. Also, definitions of population "equivalence" or sufficient equivalence for acceptance of psychometric data requires additional attention. It would be helpful to have specific recommendations on psychometric data that are required from a new population when a PRO is used on a different population than the one(s) on which it was developed and validated. It may not always be necessary to match inclusion/exclusion criteria to the development sample and in some cases careful cognitive interviewing may suffice.
lines 63 etc and lines 305 etc	Perceptions of performance capacity vs. self-report of performance may both be useful ways to address valid concepts and recommending against the former is too restrictive. Recognition should be made that both types of measurement are distinct from performance-based measurement. To extend on the point made above, some PROs may incorporate performance-based measurement as well.
Line 330 (page 10):	...requested information [was: information requested] as proposed.
Line 333 (page 10):	...what steps are [was: were] taken to ensure...
Line 356 (page 12):	...the intended population (e.g., patients with to complete). <i>[just to be consistent with the other bullets]</i>
Line 448 (page 14):	After this bullet, please add new bullet for <i>Language unknown to patients</i> . This is a different factor from literacy level, because one patient can be highly literate and fluent in one or a few languages/dialects but not other languages.
Line 466 (page 14):	...as diagrammed in Figure 2) has [was: have] been...
C	
(275-279, 478-480),	Much attention is paid to the validation population and final population to justify a claim, focusing on age, sex, ethnicity, cognitive ability and socioeconomic status. Also here only relevant characteristics (e.g. disease severity) should be mentioned.
	A Counterpoint: Title: Inability to Ascertain the Validity of Broad Patient-Reported Outcomes (PRO) Measures By Steven Pashko, Ph.D. Without a specified contextual framework to guide them, patients utilize two viewpoints as they answer broad questions about life, happiness, subjective well-being and QOL. The ability of one person to use two distinctly different viewpoints as they answer PRO questions of a broader nature makes dubious the validity of these tests and the singular nature of these concepts. The dual viewpoints we all utilize in answering questions of a global and subjective nature are called "objective/ subjective", "relative/ absolute" or "ego/ Self" (C.G. Jung, <i>The Undiscovered Self</i> , 1957). ...As researchers, we need to remember this viewpoint exists and appropriately address the ramifications it presents. This may be especially true when PROs are asked of patients who are at the end of their lifetime but this viewpoint is also dominant at other times and in other circumstances. Many tests currently in use require subjective reports of a subjective nature (i.e., assessments of depression, pain, anxiety). Of those, most have undergone validity and reliability testing and performed well. It's clear that patients can be assessed

	<p>relatively well (i.e., with good validity) on a variety of subjective states. That such assessments lead to changes in treatment practices or formulary decisions seems reasonable because the assessments are limited to fairly specific conditions. During such testing, patients more naturally utilize the ego viewpoint because of how the questions are framed. This permits relatively high levels of test reliability and validity to be obtained. When it comes to the relationship of testing and policy decisions, however, this author is less comfortable with the claims of validity for broad concepts such as QOL and subjective well-being. Because of the relatively broad, subjective nature of the concepts there is probably an increased propensity for both viewpoints to be used. Each viewpoint has an inherent validity of its own, yet each is remarkably different from the other. For policy decision-making, it seems appropriate to understand the strengths and weaknesses of both viewpoints prior to settling on the usefulness of either one. Then, tests can be constructed that assess QOL and subjective well-being from either one or the other or both viewpoints. Policies based on the ego viewpoint may tend to over-value physical health. There is value to each viewpoint but researchers must not confuse or try to combine the two. In summary, the reporting of outcomes directly from patients is very valuable. The draft guidance on patient-reported outcomes acknowledges and describes the benefits well. History has shown that reliable and valid tests can be constructed and utilized for use when patients are asked to provide subjective reports on circumscribed outcomes. However, when the outcomes broaden to such concepts as overall QOL and (total) subjective well-being, concern should be heightened. The egoic viewpoint typically used when answering questions of a limited nature may be exchanged for the Self viewpoint when questions of a more global nature are asked. Those who construct global PRO assessments need to understand the value and appropriateness of including the appropriate respondent viewpoint in the construction of tests so that test concepts can undergo appropriate validity testing.</p>
475	<p>The comment in brackets should be either more explicit on what minimum important difference refers to or should be deleted.</p>
480	<p>Table 4 should be described as taxonomy . Not all of these measurement properties will be examined for every instrument (e.g. predictive validity. Suggestions: Change sentence on 478 to read: “Table 4 outlines the taxonomy of measurement properties that could be studied for a new instrument.”</p>
527	<p>It places too big a burden on the sponsor to examine every possible race, age or ethnicity when looking at responsiveness. This requirement should not be implied in the PRO Guidance. Suggestions: Delete sentence (The extent...) The last sentence covers this concept.</p>
	<p>In particular I have difficulty with the present position of most researchers and clinicians in regard to Minimum Important Difference (MID), responsiveness measures and other metrics to express relevant changes (page 19). In my view, science is characterized by the fact that it uses standardized and well-defined measures. In the case of MID and responsiveness measures there seems to be a whole bunch of different constructs and measures. In my view, if there is no agreement how to measure such constructs or measurement properties, this may indicate two things: 1) one of the measures is correct the others not, or they are all derivatives from each other and therefore are (more or less) all equal, 2) all measures are wrong, there is not such a thing like MID or responsiveness that can be measured. The latter option is probably more likely than the first option, because relevance and meaningfulness are subjective aspects that have to be ultimately determined by the research/medical community and can not be determined by mathematical procedures. Therefore, the approach mentioned at the first bullet may be optional. The second bullet seems to me an approach with some characteristics of a perpetuum mobile.</p> <p>Bullet 3 represents the technical, mathematical approach that in my view is not offering what we are looking for. Bullet 4 is indeed arbitrary. The last advice is that it is generally helpful to use a variety of methods to confirm MID is not very scientific and in my view is even showing that we don't know what we are doing. Maybe it is better not to mention the MID because there is simply no general agreement about its concept and the way to measure it.</p>
p.18	<p>Reliability-I do agree that test-retest is the most important type of reliability testing For Pro Instrument Used In Clinical Trials.</p>
	<p>As a participant at the meeting this past weekend I am heartened that the FDA may actually heed the concerns voiced. HOWEVER, it is important to note that the draft document is written like a text book on how to create the perfect PRO instrument and how to conduct the perfect PRO analysis. Other measures, like physical measures (BP or weight) or toxicity measures (NCI Common Toxicity Criteria) are not held to the SAME rigor as this document suggests PROs should be held. It was the consensus of the 400 people in the room (as assessed by a show of hands) that this document is out of balance with standards or guidance for the measure of ANY OTHER outcome instrument. At the bottom of this email is a small sample of abstracts documenting the low to moderate reliability of physical measures that are generally accepted for outcomes measurement without being held to the rigor the draft document proposes for PROs.? There is no known PRO measure, in oncology or in any other field that has been developed using every method listed in the guidance. Not one of the drugs that have to date gained approval based on PRO claims would have received</p>

	<p>approval if this document was in place and used as a “checklist” for approval.</p> <p>Further, although the document specifically states “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.” It is clear by the attendance of so many NCI folks, including almost all of DCP that this document will inform how NCI (including DCP and CTEP) review PRO study aims. I think there are some important issues to consider: 1. Before this guidance is approved it should be in balance with the guidance for the measurement of outcomes using any other measure including physical or toxicity measures. Other measures should be held to the same rigor as this guidance holds PROS or the language in this guidance must be modified. 2. It is fine to list the “text book” version on how to develop PRO instruments and how to study PROs HOWEVER language is needed that qualifies how to continue sound research with the current state of the science. The current state of PRO science does not provide instruments developed using all of the criteria mentioned in the guidance, nor could this be done any time in the near future. 4. If any entity could develop instruments with such rigor, it would only be pharma, since they are the only ones with such resources, but even with pharma, this would take tremendous time and resources and research should not cease until the perfect world of PRO instrumentation is attained (if it could ever be). The problem is that once pharma is forced to put the resources into instrument validation, there is no mandate for them to share the tremendous work. Pharma should be in some way encouraged and REWARDED for putting such work in the public domain. 5. The NCI sponsored cooperative groups and individual investigators do not now, nor will they ever, have the resources for instrument development and validation put forth in this document. This in no way should be a barrier to sound PRO research in the cooperatives or by individual investigators. 6. When not in conflict with copyright laws, there should be public databases of PRO instruments (as there are now) but with the addition of a comprehensive compendium of validation work. As mentioned above, this is a small sample of abstracts documenting the low to moderate reliability of physical measures that are generally accepted for outcomes measurement without being held to the rigor the draft document proposes for PROs. Most have kappa’s below what we would accept with PROs.</p>
→ Hence: @ section IV. A – D	<p>on the other hand: a very detailed information what the FDA will review and question if PRO instruments are used is given ⇒ the result of this guidance to industry might be a clear a tendency by any sponsor to avoid the use of any PRO instrument,</p> <p>Especially in the current early phase of developing PRO instruments and integrating them into Cuts this section is nearly prohibitive to the use of PRO instruments.</p>
@ IV. C. 4. a row 566/567:	<p>The need for MID is really questionable – if a treatment is leading to a very small change within the treated group this is an additional effect (provided sample size is huge enough), who should be able to judge whether this is “important enough”. Obviously for the patients it is, reflected by the PRO. For otherwise measured clinical effects (traditionally, by representatives of the patients like the GP) the clinical relevance or minimum important difference is in fact the one which is leading to a relevance to the patient. However, if a PRO instrument is leading to a difference, than it is already relevant, otherwise the patients would not have recognised the difference and the PRO would not lead to a difference.</p> <p>The remaining question is the willingness to pay for such (a small) difference in PRO.</p>
@ IV. C. 4. b row 576 f:	<p>The definition of responder – however it is defined – remains an artificial demarcation. PRO has the advantage to go beyond the concept of responder definition. Hence, these concepts should be used complementary and one replacing the other. That means the comparison eg of different percentages of responders in different study groups, might give additional information, but it should not be taken as a real measure of difference on which to base and decisions about medical products (which is done a lot indeed, as a result of lacking better methods.) PRO instruments should be able to give new insights on the effects of medical interventions by assessing any changes in more detail.</p>
	<p>... Actually what I am missing is the discussion of scale property (ordinal etc.) and statistical calculations in more depth. What I haven't done, is to cross-check for consequences to industry.</p>
p.20	<p>The long first sentence at the top of p. 20 should be broken into at least two sentences.</p>
Table 4, p.17	<p>The FDA is looking for comment on MID and responder definitions. One general point is that the focus on the MID may obscure the more important general issue of setting <i>a priori</i> criteria for establishing “important” differences – statistically or via another method. This point warrants additional attention.</p>
IVC.1, p. 18, line 495 etc	<p>The document would benefit from recognition of the distinct purposes of internal consistency reliability data and test-retest reliability data. As presented here, internal consistency reliability is not as important as test-retest data. We disagree with the implied evidence hierarchy.</p>
IVC.4.a, p 19, line 545	<p>"If PRO instruments are to be considered more sensitive than past measures..." - that's a high bar and although increased sensitivity is often a motivator for creation of PROs, it's not always. Sometimes the goal of a PRO is to obtain the patient perspective, and optimizing all psychometric performance is always part of good instrument creation. Rewording would mitigate concern about an “unlevel playing field.”</p>
Line 508 (page 18):	<p>the FDA would be [was: is] interested in...</p>

Line 511 (page 18):	...the FDA would be [was: is] interested in...
Lines 571-577 (page 20) & 692-699 (page 22 & 23): Comments to Definition of responders, & Special Populations (Sections IV.C.4.b & E.2)	---- Will a proxy be regarded as a responder from the start of a PRO study? The document mentions in Section E.2 about inclusion of proxy reports for patients that become cognitively impaired or unable to communicate "over the course of study". I know that CMMS has studies (the one I know is not for supporting medical product) with PRO data that came entirely from proxies when some patients were too ill to self-report from the start of study, i.e., there was no patient self-report in parallel. Does FDA accept this to support a claim? If so, it will be helpful to reflect and clarify in this draft in both Sections IV.C.4.b & E.2 that proxy can be a responder and can be one from the beginning of a PRO study. If need further definition, it can be e.g.: A responder is a person other than the patient who understands very well the patient's feelings, attitudes, and abilities in physical, psychological and cognitive terms and can represent a patient for his or her outcomes.
#502-504:	what is meant?
#514	example
	References need to be provided for all definitions, diagrams, figures and tables. The concepts, definitions and processes presented are not novel and have originated from the field of psychometric/survey research and have been widely documented within measurement theory, sociology, psychology and education literature and it would be ethical to provide references of the various sources wherever appropriate....Not clear as to what types of documentation (e.g. publication etc) are acceptable to demonstrate proven validity for a pre-developed instrument. It would help provide clarity by providing a list of acceptable documentation types....Since 'paper' was the most widely used previously, most instruments are in paper form. It is not very clear as to what type of validity studies would be considered adequate. Content validity is very subjective, and there is no widely used objective measure. Further clarity would help. Evaluation of Reliability. Test Retest between visits may not be appropriate. Test-retest is used to test reproducibility, given no other change. Since in clinical trials there could be a trt benefit involved, this might not be appropriate across all studies. Sample size considerations for developing a new instrument need to be discussed within the document. For translations, clarification is required as to if redoing the validity analyses performed during original instrument development and providing consistent results would be sufficient as evidence of measurement properties of translated versions being preserved
	I think this is a very interesting document, however it needs to be worded carefully as the current draft leads to a literal interpretation that is far too strict. Interpretability Also, regarding interpretability of results - section IV C - the text does not distinguish between MID - a patient based meaningfulness; and MCID - a clinical based meaningfulness. This is a useful distinction to draw as meaning is often different when viewed in patient or clinical terms. It would also be useful to have some guidance on the way to decide upon THE MID level when various levels using different methods have been calculated.
D	
At several places (e.g. 178-181, 579-670)	it is mentioned that small changes to the instrument or instructions are viewed by the agency as new versions and will lead to a complete review of all evidence on psychometric properties. Given that translation of instruments and instructions often lead to slight revision of these documents, it is not realistic to require full validation of revised instruments.
470	Somewhere in the Guidance should be a section on existing instruments and standards for assessing their use in labeling. I don't know if that should come at the end of this section, or somewhere else. Validated instruments should not have to be re-created in light of this PRO guidance.
527	It places too big a burden on the sponsor to examine every possible race, age or ethnicity when looking at responsiveness. This requirement should not be implied in the PRO Guidance. Suggestion: Delete sentence (The extent...) The last sentence covers this concept.
610	To require modification or re-validation if a population is "different" is far too broad and onerous. It would be hard to lay out a guideline around this. Suggestion: Change to "if a population is determined from the literature or other avenues to be substantially different, re-validation of an existing instrument should be considered.
618-629	Same comment as above- small changes in working, etc should not require re-validation. Suggestion: Delete the examples and make a more general statement that significant change in content or format should be examined and re-validation considered..
667	Adding an instrument as a battery should not require extra work. Suggestion: Delete the second bullet,.
page 21; line 621:	To require additional validation if "wording or placement of instructions" has changed is too strict
page 22; line 660	"The evidence that measurement properties for translated versions are comparable". Should this sentence be interpreted that a psychometric validation is needed for each new translation? It is not very practical/feasible to perform separate psychometric validations for each new translation with respect to budget/time

	lines/resources. Could a more extensive cognitive debriefing be usable? How to handle "old" translations that have been used for for a long time, but no formal psychometric validation has been performed for each translation?
page 22; line 666.	Several PRO instruments have been developed in clinical practice and then incorporated into clinical trials.
Line 582-667	<p>Although this section begins (Line 582) with the statement that “When a PRO instrument is modified, additional studies may be needed to confirm the adequacy of the modified instrument’s measurement properties,” language later in the section (i.e., Line 590: “The FDA intends to consider a modified instrument as a different instrument from the original...”) suggests a more stringent requirement than “additional studies may be needed.” Hence, there is a concern that even minor, non-substantive changes will require “validation” studies. However, the sentence starting on line 583 (i.e., For example, small nonrandomized studies may be adequate to assess the results of changing a response scale from vertical to horizontal”) does leave the door open for cognitive testing in small numbers of people.</p> <p>It appears that the issues addressed in D.1. Revised Measurement Concept (Lines 595-606) reflect accepted practice and that it is understood that additional evidence would be needed to support the measurement properties for a “new concept” in the situations listed.</p> <p>The language in the draft Guidance regarding Changed Culture or Language Adaptation is consistent with generally accepted practices regarding the cross-cultural adaptation process. However, the statement (Line 660) in the Guidance that “The evidence that measurement properties for translated versions are comparable” causes some concern. If the cross-cultural translation process is conducted in a manner consistent with accepted practices, sufficient measurement equivalence should be obtained. Cognitive testing in a small sample of the in-country target population should be considered acceptable.</p> <p>A number of the “modifications of an existing instrument” addressed in this section should not lead to the need for extensive “additional validation” (Line 592). For many of the potential modifications (e.g., Line 621: “Wording or placement of instructions”), cognitive testing or debriefing in a small sample of the intended/target population should suffice.</p> <p>The statement/bullet on Line 666 appears to be unnecessary. If the measurement properties of the PRO measure in the intended/target population are adequate, where it was developed should not matter. Many PRO instruments were not specifically developed for use in clinical trials but could be very appropriate for use in one. The same criteria for measurement properties should be required by the FDA regardless of the context in which the measure was developed.</p> <p>The statement/bullet starting on Line 667 seems unnecessary. Multiple measures are often combined into a battery of measures, but it unlikely to cause a problem in the clinical trial setting. There is evidence that order effects can result when measures are combined in a battery; however, as long as the order in which the measures are administered does not change within the trial, this should not be of concern</p>
E	
(275-279, 478-480	<p>Much attention is paid to the validation population and final population to justify a claim, focusing on age, sex, ethnicity, cognitive ability and socioeconomic status. Also here only relevant characteristics (eg disease severity) should be mentioned.</p> <p>The FDA plans to compare the patient populations used in PRO instrument development process to the study populations enrolled in clinical trials to ensure that they are similar.</p>
277-278	<p>Identification of intended population-very important point, could be critical for many PRO analysis, especially “cognitive ability” mentioned in line 277-278. How we should determine the cognitive ability of the patient?</p> <p>Specific test? Who should create that kind of test in different countries or for different cultural groups? I think it is a serious problem and should be resolved.</p>
. IVB.7, p.13, line 429:	I recommend expanding discussion of the appropriate application of population-specific preference weights. Also, definitions of population “equivalence” or sufficient equivalence for acceptance of psychometric data requires additional attention. It would be helpful to have specific recommendations on psychometric data that are required from a new population when a PRO is used on a different population than the one(s) on which it was developed and validated. It may not always be necessary to match inclusion/exclusion criteria to the development sample and in some cases careful cognitive interviewing may suffice.
	Cognitive status, education level, age group, ethnic group and response fatigue of patients may affect the outcome measures significantly and require better emphasis. It is important the developed instruments are sensitivity to cultural changes and validated across targeted population.
(page 8):	What is the FDA perspective on a built in adherence subscale?
line 277	what the FDA perspective on use of PROs in cognitive disabled patient populations?
line 329	recall period in patients with some cognitive disability
line 432 etc	non-distracting conditions facilitating completion of questionnaire material is of relevance especially in

	patients with mental illnesses; burden could be reduced to acceptable level by providing a time window for completion and by staggering questionnaire administration if multiple questionnaires applied
lines 643 and 692	The discussions of language changes and proxy respondents beginning on lines 643 and 692, respectively seem scanty. At least, the agency might note that family members or friends, rather than doctors and their staff members, should generally be proxy respondents on PRO questions. Some detail on when proxy responses could be the basis for labeling claims would be useful, e.g., pediatric patients with a condition lacking in hard clinical endpoints.
<b>V. STUDY DESIGN</b>	
lines 643 and 692	Overall, there is not enough of concrete examples. A list of commonly used instruments as an appendix would be great. I have a suggestion as a "step 2" to this review. What about a project/study that would provide an overview (comparison) of guidelines when it comes to PRO measurements in clinical trials from different countries? As we know, clinical studies more than often involve many countries and therefore the choice on what and how to measure PRO in trials is based on guidelines not only from the USA but also from other countries. This project/study could support or provide additional insight to FDA and other health authorities on why the PRO measures in certain trials do not entirely follow local guidelines. Just a suggestion!
711	CRFS are not included in protocols. Suggestion: Delete the words "exact format"
785	Requiring longer follow-up than for other endpoints is not practical. Suggestion: Delete sentence "In a trial..."
813-858	This entire section could be reduced and just refer to other Guidances on electronic data capture..
page 23; line 718	"rarely credible" is a very strong expression. This would exclude several treatment satisfaction assessments.
Line 1019:	Review document to see if other recommendations can be made for example in the section Interpretation of study results
	Highlight all recommendations in situ or produce a summary of recommendations
	What about recommendations on the handling of queries by respondents during the performance of a study?
p. 21	In the second bullet on p. 21, insert "used" between "those" and "for".
p. 26	At the top of p. 26, it might be good to mention back-translation as a useful method.
Page 27,	first line under "1. Blinding and Randomization": Insert "such measures used in" before "open-". As it is, the sentence says that open-label studies [in general] are rarely credible, but that's true for ones that don't use PRO measures as well (although single-arm, open-label studies can be useful for assessing long-term safety).
VA.2, p. 24:	Clinical trial QC and standardized instructions are extremely important and mention in the guidance is helpful.
#717	Should it be called 'patient's perception'? Impression is to negative.
	FDA PRO Use in Medical Product Development to Support Labeling Claims References need to be provided for all definitions, diagrams, figures and tables. The concepts, definitions and processes presented are not novel and have originated from the field of psychometric/survey research and have been widely documented within measurement theory, sociology, psychology and education literature and it would be ethical to provide references of the various sources wherever appropriate. Sample size considerations for developing a new instrument need to be discussed within the document.
P.23	General Protocol Considerations-I should add the possibility to diminish the patient bias in blinding and randomization- CORRELATION the selected items used in clinical and PRO trials. <b>THIS IS THE PROPOSAL FOR THE 2 LEVEL QUALITY CONTROL!</b> The time for the scientific study methods in PRO ANALYSIS IS COMING. We should have in mind that the patient is the most sensitive "instrument" and should be treated with the open and wise way.
<b>VI. DATA ANALYSIS</b>	
Line 912 (page 28):	
Lines 807-809-	talk about the need for pre-specifying the way in which results will be interpreted is important for new instruments
Lines 927-929-	show the agency know that substantial clinical experience is needed to establish interpretation rules. We assume the agency will work with sponsors on this topic when the instrument is new
Line 940 says	"...it is critical to ensure that patients enrolled in a clinical study are impaired in all domains" of a multi-domain instrument being used in the trial. We can see the point that if this rule is violated, a composite score using all the domains may not show a change or difference. We do not see why this rule needs to be followed if the a priori intent is to report selected domain scores and sue them as the basis for a label claim (assuming the domains and their scores are validated). We think FDA's comment here stems from the agency's predisposition to think that all the domain results must be similar before the data produced by the instrument can be used in any way. This concept should be revisited.
	how to account statistically for substantial imbalance in drop-out rates between treatment arms? Key in

	<p>guidance: (Co-) primary endpoint needed for a labeling claim or secondary endpoint would be also sufficient to support such a claim? No clarity at this point. Presumably superiority should be the objective. Superiority and powering of the study to be based on MID taking potential multiplicity into account when needed? What is required for getting some wording in the clinical trial section of the product labeling? Superiority or significance sufficient? In case of labeling claim/wording in clinical trial section, one or multiple trials with consistency in outcomes required? Some more detail with regard to development and validation of a new instrument in particular statistical tools to be applied and why;</p>
p.27-	Data Analysis -The Wright Answer Is-"There Is No Single Best Statistical Procedure"
p.30-	Interpretation of the study results-the logic and wise interpretation as in the good scientific work.
	The time for the scientific study methods in Pro Anlysis Is Coming.We should have in mind that the patient is the most sensitive "instrument" and should be treated with the open and wise way .
	<p>My comments are related to the missing data section.</p> <p>I think that the FDA can beef up their guidance should be more like a cookbook. I would like to address the issue of handling missing data when conducting analysis on data collected from a clinical trial. 1) should missing data at baseline be imputed? If the percent of patients who are missing a baseline visit do not differ from patients who have post-baseline measures, then could use multiple imputation methods to impute values at baseline. If there is a difference, then baseline measures should be set to missing. 2) In case of data missing at post-baseline, the following approach should be used : Prior to performing analyses, evaluate the amount, reasons for and patterns of missing data.. If the reason for missing data is not related to the endpoint of interest, then the missing data are considered to be missing completely at random (MCAR). If the missingness depends on previously observed outcomes, then the data are considered to be missing at random (MAR). If the probability of missingness depends on the unobserved measures, then missing data are considered non-ignorable, or missing not at random (MNAR). There may be all three types of missing data in any particular study. I suggest that a test for the pattern of missingness be conducted: To test for MCAR, we will compare the baseline characteristics of subjects with and without complete data, both within and between treatment groups. To evaluate the possibility of MAR, create an indicator variable to denote whether or not the assessment was completed at each visit, and then use logistic regression models to determine whether the previously obtained measure is associated with the probability of missingness at the next visit. Although the sponsor should prospectively document reasons for dropout, formally distinguishing between MAR and MNAR is not trivial and relies on assumptions that are themselves untestable. In which case, the primary analysis should be based on the MAR assumption with sensitivity analysis based on an MNAR assumption using pattern mixture models. Pattern-mixture techniques may be used to assist in the analysis of potentially nonignorable missing data. These methods may be implemented as a means of evaluating the robustness of the results obtained from the primary analysis. There are varied MNAR approaches that can be used, since these methods require assumptions that usually cannot be validated from the data at hand. Thus, one potential strategy that may be implemented is to classify subjects into two or three groups, using dropout status, reason and time of the last completed assessment as possible grouping factors. Separate models will be created within each stratum and the parameter estimates will be combined into a weighted average for the entire study population. Sensitivity analyses will be performed to compare the results of different analytic strategies to ensure that treatment effects are consistent across different analytical models and to evaluate the range of possible differences. 3 ) The impact of differential censoring on the analysis should also be examined. These analyses consist of evaluating if there are differences in the mean baseline scores of patients by study completion status. Study completion status is defined by the number of assessments completed by the patient. This analysis will be conducted for all patients grouped together, and stratified by treatment. Also, an analysis of PRO completion rates should be assessed. This entails an analysis within the questionnaire (how many items are missing) and across questionnaires over time. 4) The use of the best case/worse case scenario is not appropriate. Analysis of outcome data are often compromised by the lack of complete data. Missing information can result in biased estimates of the effects of the intervention of interest, and the loss of statistical power to detect group differences. The choice of which imputation method to use is largely based on the underlying mechanism of the missing data. However, multiple imputation methods are preferred to single imputation methods since the former approach reduces the random variation associated with missing data. Multiple imputation techniques are consistent with regulatory needs for a simple pre-specified analysis; are easy to implement with standard software, and is an improvement over single imputation methods.</p> <p>5) There are a variety of methods that can be used to impute for missing data. The goal of the imputation technique is to incorporate missing information into the analysis. The imputation method can either produce implicit or explicit data values that can be used in the analysis. These technique(s) use observed values to impute for missing values. A consideration when imputing is that the imputation method should not distort neither the distribution nor the variance of the data for which the imputation is being conducted. Generally, techniques which produce explicit values (values which complete the dataset) are</p>

mostly applicable to cross-sectional models. Explicit imputation for missing longitudinal (panel) data is complicated by the fact that in these data unobserved heterogeneity is particularly important because respondents are assessed on more than one occasion leading to the probability that stable person specific effects are likely present. Data can be imputed either by single or multiple imputation. Maximum likelihood methods are used to implicitly impute for missingness. a. Common Single Imputation Methods: 1) Mean replacement 2) Hot-Deck 3) Last Observation Carried Forward (LOCF) 4) Predicted Mean b. Common Multiple Imputation Methods: 1) Predictive Model-(Bayesian based) 2) Propensity Score (approximate Bayesian bootstrap) C. Common Maximum Likelihood Methods; 1) Maximum Likelihood or restricted maximum likelihood (EM) 2) Full Information maximum likelihood Single imputation is easy to employ with a single value imputed for the missing data. Once missing data are imputed, standard analysis can be used to estimate the parameters. In panel data, LOCF uses person specific information to impute for missing data under restrictive conditions. Using person specific information to impute for missing data is preferred to say, mean imputation, in disease states that decline over time. However, LOCF will not capture improvement in patient outcomes as a result of an intervention. Multiple imputation to impute missing values uses a model that incorporates random variation. The model is replicated K times (usually 3-5) times producing K complete data sets. The desired analysis is performed on each data set using standard complete-data methods. The values of the parameter estimates are averages across the M samples to produce a single point estimate. Standard errors are calculated by averaging the squared standard errors of the K estimates.. This approach, however, is cumbersome as multiple datasets have to be created and is more appropriate for cross-sectional datasets. Random effects models (mixed effect model where the covariance structure is modeled and the subjects are consider random effects) as estimated by SAS (Proc Mixed procedure) is a maximum likelihood based approach from which to develop longitudinal analyses under the MAR assumption. If the missing data mechanism is MAR, then the analysis are more robust to potential bias from missing data than say, LOCF. In these analyses, observed data are used to provide information about the missing data, but missing data are not explicitly imputed. This approach uses person specific information to implicitly impute for missing data and is easier to implement. These models also allow for unbalanced data.

The full information maximum likelihood approach has been used in association with structural modeling.. Missing data imputation is based on partitioning the dataset into levels with each level containing similar missing data characteristics. Thus, if the division of the database results in many cells, this approach could become cumbersome. LISREL and AMOS are two software packages that are available to estimate missing data within this framework. Clearly, there are number of methods that can be used to impute for missing data. The question of course is which to method of imputation to use in light of the fact that there is no "gold standard". Single imputation is not as good as multiple imputation because it does not reflect the extra uncertainty associated with missing data. In general, single imputation methods do not perform as well as other methods and some may produce biased coefficients and/or standard errors. Further, though LOCF is the method of choice for regulatory authorities, it produces conservative estimates and only performs well if the underlying missing data mechanism is MCAR, and if the patients' response is constant from the last observed value to the end of the trial. In practice, these conditions seldom hold and carrying the last observation forward may confound treatment with time leading to a bias in the difference between estimates of treatment. There are commercially available programs, e.g., SOLAS, SAS and NORM, that can be used to employ multiple imputation strategies to generate a complete dataset. The algorithm used by NORM is predictive model based and uses the Bayesian method for augmenting datasets. SOLAS algorithm is propensity score based. There are however, several conditions that must hold in order for this strategy to produce unbiased parameter estimates (means and standard errors). First, the underlying missing data mechanism must be MAR; Second, the model used to generate the imputed values must be "correct"; Third, the model must match up with the model used in the imputation. Satisfying these conditions in practice may not be completely straightforward. Additionally, monte carlo studies have suggested that the parameter estimates are sensitive to the multiple imputation method employed.. Under general conditions, predictive modeling performs better than the propensity score approach in producing unbiased parameter estimates.. In contrast to the multiple imputation strategies that require additional data manipulation and modeling choices to impute for missing values, and to take into account uncertainty in missing data, the likelihood-based mixed-effects approach is easy to implement; is dictated entirely by the design of the study, and does not need additional steps to accommodate missing data. Further, likelihood-based models are based on reasonable assumptions regarding missing data, and are robust to violations of those assumptions, and provide an appropriate general analytic framework for assessing response profiles in longitudinal clinical trials. Summary The choice of which imputation method to use is largely based on the underlying mechanism of the missing data. To test for MCAR, baseline characteristics of subjects with and without complete data, both within and between groups are evaluated.. The possibility of MAR is tested by creating an indicator variable to denote whether or not the data point for each subject was available at each visit. Once

	<p>the indicator variable is created, a logistic regression model is used to determine whether the previously obtained measure is associated with the probability of missingness at the next visit. While the test to distinguish MCAR from MAR is straightforward, the test to formally distinguish MAR from MNAR is not trivial and relies on assumptions that are themselves untestable. Pattern-mixture techniques are typically used to assist in the analysis of potentially nonignorable missing data. However, there are varied MNAR approaches that can be used, because these methods require assumptions that usually cannot be validated from the data at hand. Thus, the analyst is left with no choice but to test for whether the missingness is MCAR or MAR and if the missingness is not MCAR, then to assume that the missingness is at least MAR, and to use MNAR methods as a means of evaluating the robustness of the results. Since, in most cases the missingness mechanism is MAR, the choice of which imputation method to use is based on considerations related to reducing random variation. In this respect, multiple imputation methods are preferable to the single imputation method because it reduces the uncertainty associated with missing data. I.e., multiple imputation reduces the variability between the true subject score and that which is imputed. A rule of thumb for panel data, is that when the missing data mechanism follows a MAR process, and if there is a likelihood that missingness introduces an imbalance in the groups that are being evaluated, then using maximum likelihood (EM) methods to impute for missing data would be a good choice. A rule of thumb for cross-sectional data, under the same conditions is to use either an EM or the Bayesian based approaches. These approaches, for example, are consistent with regulatory needs for a simple pre-specified analysis; are easy to implement with standard software, and is an improvement over single imputation methods such as LOCF and worst/best case scenario.</p>
<b>GLOSSARY</b>	
@ Glossary	<p>very good, giving some definitions and explanations which are missing in the text, but incomplete to be considered to include the definitions of the glossary in the text as well, the glossary should just give an overview, but the body of the text should give any details without needing to check within the glossary,</p> <p><i>Line 1037:</i> An expanded glossary-with real not theoretical examples would be helpful e.g. <i>line 1047</i> concept, <i>line 1050</i> conceptual framework and <i>line 1055</i> domain, as well as positioning it at the front of the document</p> <p><i>Line 1019:</i> Review document to see if other recommendations can be made for example in the section Interpretation of study results</p>
Line 1034, Page 30	second line from bottom: Change "definition" to "change".
	<p>Line 1040 (page 31): ...label or in any advertisement [was: advertising] of prescription drugs.</p> <p>Line 1055 (page 31): ...within a multi-domain [was: multidomain] concept.</p> <p>Line 1058 (page 31): A multi-domain [was: multidomain] concept that...</p>
	Throughout the document it is stated what the FDA plans to review, though very often words as "usually and generally" are used which introduces uncertainty as to what is important in a specific case
	In general it was felt that the guidance provides clarity as to what can be expected from the FDA in relation to label claims. Throughout the document it is stated what the FDA plans to review, though very often words as "usually and generally" are used which introduces uncertainty as to what is important in a specific case
	It should be noted that there are a number of generic PRO instruments, such as the EuroQOL, Quality of Wellbeing, Health Utilities Index, and SF-36 (and related instruments) that have been subject to years of validation and verification in myriad populations. The FDA should encourage standardization and discourage a proliferation of instruments. Many of these generic instruments also have disease specific components. The use of standard measures assists in building a large database of values associated with specific health states and lends to comparability of assessments across studies. The Agency should view instruments developed ad hoc for clinical trials with extreme suspicion and so state.
	In summary, the reporting of outcomes directly from patients is very valuable. The draft guidance on patient-reported outcomes acknowledges and describes the benefits well.
	This report is a technically excellent report about PRO instruments. PRO instruments are used to support the effectiveness of drugs and medical products because they can incorporate the patients' perspective alongside traditional physiological and functional assessments, assessing both observable and un-observable health benefits.
	As a participant at the meeting this past weekend I am heartened that the FDA may actually heed the concerns voiced. HOWEVER, it is important to note that the draft document is written like a text book on how to create the perfect PRO instrument and how to conduct the perfect PRO analysis. Other measures, like physical measures (BP or weight) or toxicity measures (NCI Common Toxicity Criteria), are not held to the SAME rigor as this document suggests PROs should be held. It was the consensus of the 400 people in the room (as assessed by a show of hands) that this document is out of balance with standards or guidance for the measure of ANY OTHER outcome instrument. At the bottom of this email is a small sample of abstracts

	documenting the low to moderate reliability of physical measures that are generally accepted for outcomes measurement without being held to the rigor the draft document proposes for PROs. There is no known PRO measure, in oncology or in any other field, that has been developed using every method listed in the guidance. Not one of the drugs that have to date gained approval based on PRO claims would have received approval if this document was in place and used as a “checklist” for approval. Further, although the document specifically states “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.” it is clear by the attendance of so many NCI folks, including almost all of DCP, that this document will inform how NCI (including DCP and CTEP) review PRO study aims.
	This draft guidance by the FDA "... describes how the FDA evaluates patient-reported outcome (PRO) instruments used as effectiveness endpoints in clinical trials" and "... describes [the FDAs] current thinking on how sponsors can develop and use study results measured by PRO instruments to support claims in approved product labeling", but it does not at all give any statement of the FDA about the need, support or request to sponsors to add PRO instruments into CTs.
	It would be helpful to have a statement early on in the guidance that the document applies only to regulatory application of PRO research, and to acknowledge that there are broader issues for use of PROs in clinical and research contexts that are not expressly addressed in this document.
	An important objective of this report is to address the comparability between studies and to achieve a “gold standard” in using the PRO measures. However, this has received lesser attention, especially, uniformity, reproducibility and transparency of methods and results need more emphasis. Also, it would be helpful to identify if PRO is measured as a primary or secondary endpoint? This identification has important implications for product development and analysis.
	References need to be provided for all definitions, diagrams, figures and tables. The concepts, definitions and processes presented are not novel and have originated from the field of psychometric/survey research and have been widely documented within measurement theory, sociology, psychology and education literature and it would be ethical to provide references of the various sources wherever appropriate.
	Why is this guidance limited to labeling claims? It would be worthwhile to include a section related to what is required (to demonstrate) to get appropriate wording in the clinical trial section of the product labeling.
	In general it was felt that the guidance provides clarity as to what can be expected from the FDA in relation to label claims. Throughout the document it is stated what the FDA plans to review, though very often words as "usually and generally" are used which introduces uncertainty as to what is important in a specific case

**RESPONDENTS:****Pharmaceutical/Medical Device/Diagnostic/Biotech Industry**

- Rod Barnes MBA, Alcon Laboratories, Inc., Fort Worth, TX, USA
- Kati Copley-Merriman MS, MBA, Pfizer Global Pharmaceuticals, Ann Arbor, MI, USA
- Isabelle Côté PhD, Hoffmann-La Roche Ltd., Mississauga, ON, Canada
- Erwin De Vries MSc, NV Organon, Oss, , Netherlands
- Rahul Dhanda PhD, Bayer Healthcare Pharmaceuticals, West Haven, CT, USA
- Carol Fairchild PhD, Alcon Laboratories, Fort Worth, TX, USA
- Claire Gilbert MSc, BSc, Pfizer Ltd, Sandwich Kent, UK
- Shrividya Iyer PhD, Wyeth Pharmaceuticals, Norristown, PA, USA
- Ursula M. Kuehnel MS, Berna Biotech, Bern, Switzerland
- Oscar Leeuwenkamp PhD, PharmD, N.V. Organon, Oss, Netherlands
- Ann-Christin Mork PhD, Pfizer AB, Täby, Sweden
- Delia Schaffer MSc, BSc, Pfizer Australia, West Ryde, Australia
- Curtis Waycaster PhD, Alcon Laboratories Inc., Fort Worth, TX, USA
- Helmut Wenzel MS, Roche Diagnostics, Mannheim, Germany
- Kim Yoong BSc, Janssen-Ortho, Inc., Toronto, ON, Canada

**Academia**

- Ibrahim Alabbadi MBA, BSc, Queen's University Belfast, Belfast, UK
- Stephen Joel Coons PhD, Professor of Pharmacy and Public Health, The University of Arizona, AZ, USA
- David Feeny PhD, Institute of Health Economics, Edmonton AB, Canada
- Viliam Foltan ScD, RN, University of Comenius, Bratislava, Slovakia
- Amede Gogovor MSc, University of Montreal, Montreal, QC, Canada
- Ravishankar Jayadevappa, PhD, University of Pennsylvania, Philadelphia, PA, USA
- Meredith Kilgore PhD, MS, UAB SOPH Health Care Organization & Policy, Birmingham, AL, USA

- Ivan Kocic PhD, MD, Department Of Pharmacology, Academia Medica Gedanensis, Gdansk, Poland
- Paul Krabbe PhD, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands
- Rahul Shenolikar MS, Ohio State University, College of Pharmacy, Columbus, OH, USA
- Kojiro Shimozuma PhD, MD, University of Marketing and Distribution Sciences, Kobe, Japan
- Deborah Watkins Bruner PhD, RN, University of Pennsylvania, Philadelphia, PA, USA
- Inez Wu BS, MS, Harvard Medical School, Brigham & Women's Hospital, Brighton, MA, USA

**Research Organizations**

- Ljubica Besker-Ivasovic PhD, MD, Clinica S. Anna, 6500 Bellinzona, Switzerland
- Lori Frank PhD, Center for Health Outcomes Research United BioSource Corporation, Bethesda, MD, USA
- William Furlong MSc, Health Utilities Inc., Dundas, ON, Canada
- David Klingman PhD, MA, ValueMedics Research, LLC, Gainesville, VA, USA
- Steven Pashko PhD, ICON Clinical research, North Wales, PA, USA
- Pedro Plans-Rubió MD, MBA, Sant Pol, Spain
- George Torrance PhD, VP, Innovus Research Inc., Toronto, ON, Canada