Comment on
Draft Guidance for Industry
Patient-Reported Outcome Measures:
Use in Medical Product Development to Support Labeling Claims
Docket No. 2006D-0044

Submitted by PRO Consulting™

PRO Consulting is a division of invivodata, inc.
2100 Wharton Street, Suite 505
Pittsburgh, PA 15203

Chad Gwaltney, Ph.D., Brown University
Alan Shields, Ph.D., East Tennessee State University
Brian Tiplady, Ph.D., invivodata, inc.

We applaud the FDA for the tremendous amount of commitment and energy that was clearly put into their Guidance for Industry document, *Patient Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* (hereafter referred to as the Draft Guidance). Use of PRO data in clinical trials is common but often complicated by the complex measurement theory that guides PRO measure development and use. Measurement specialists, behavioral scientists, and clinical trial developers have offered various recommendations to clarify the use of PRO instruments. The FDA's effort to consolidate these recommendations was a massive undertaking and we would like to acknowledge and publicly commend those efforts. Because of the centrality of PRO data and, ultimately, its impact and influence on public health and safety, we offer our support to the Agency for taking a science-based approach toward the larger domain of PRO assessment. We offer the following comments, suggestions, and feedback to the Agency in response to their open inquiry.

Table of Contents

Content	Page
Response to Introduction	3
Distinction between Symptomatic Assessments and Quality of Life Measures	3
Complexity and multi-dimensionality	3
Relation of PRO end-point to the disease and treatment	4
Communication of claims to prescribers and patients	4
Dynamic nature of symptom end-points	4
Summary	4
Response to evaluating PRO Instruments	5
PRO instrument validation evidence should be evaluated within the intended use in controlled clinical trials	5
The Draft Guidance is unnecessarily restrictive in response to instrument modifications	5
Draft Guidance unjustifiably requires psychometric revalidation studies in measurement situations where no or a lesser degree of evidence is sufficient	7
Reliability in PRO Assessment	10
Additional Points of Guidance Support	11
Hypothetical vs. actual behavior	11
Respondent burden and automated skip patterns	11
Missing data	11
Appropriate recall periods	11
References	12

List of Tables

List of Tables	Page
Table 1: Typical Characteristics of PROs Assessing Symptoms vs. HRQOL	3
Table 2: PRO Instrument Modification and Level of Validation Evidence	7

Response to the Introduction (Section I., page 1)

Distinction between Symptomatic Assessments and Quality of Life Measures: We commend the Agency for noting the distinction (Lines 33-37) between assessment of symptoms and assessment of more complex constructs related to health-related quality of life (HRQOL) within the text of the Draft Guidance. **However, we believe this distinction should be made clearer and stronger in the Draft Guidance, in order to clarify for sponsors the distinction between these domains to give sponsors appropriate guidance with regard to these different domains.** Although symptom assessment and HRQOL are both subject to similar considerations of reliability, validity, etc, as outlined in the Draft Guidance, assessments in these two domains are marked by substantial conceptual and practical differences that need to be recognized. Table 1 characterizes the typical properties of symptom and HRQOL assessment (we recognize that these are generalizations and are subject to exceptions, but believe that they characterize the typical approach in each domain).

Table 1

Typical Characteristics of PROs Assessing Symptoms vs. HRQOL

	Symptoms	**Health-Related Quality of Life**
Complexity of concepts	Simple	Complex
Behavioral “objective” measures	Often	Seldom
Dimensionality	Unidimensional	Multidimensional
Relation to disease	Direct	Indirect
Effect of treatment	Direct	Indirect
Primary target of treatment	Often	Seldom
Mirrors clinician and patient discourse	Often	Seldom
Dynamic time course	Often	Seldom

Complexity and multi-dimensionality*:* One important difference is that the concepts underlying symptomatic assessment are often unidimensional, whereas HRQOL assessments are often complex, abstract, and multidimensional. The Draft Guidance outlines the challenges that arise for the latter, which impose an especially high burden on sponsors to define the concept, assure that it is understood by patients, assure that the assessment tool taps the concept, and craft claims that directly and clearly communicate the concept in language clinicians and patients will understand. In contrast, assessment of symptoms typically involves constructs that are unidimensional and much less complex. One way to illustrate this is that symptomatic PRO assessment often involves reporting of “objective” behavioral end-points such as coughing, micturition, or defecation. Such simple symptomatic end-points, which often consist of frequency counts of events, do not pose the same conceptual and psychometric challenge posed by complex, abstract, and multidimensional end-points related to HRQOL. Even when PRO symptom assessment focuses on subjective and unobservable end-points, such as pain or anxiety, the concepts are considerably more straightforward, in part because they often closely mirror how patients and clinicians typically talk about the conditions and symptoms. Symptom measures are typically unidimensional, because symptoms are specific.

Symptom measures are sometimes aggregated into composite end-points, but this is usually based on thinking about the underlying condition or syndrome and the effect of the treatment, and not just on the psychometric or statistical properties of the measures.

Relation of PRO end-point to the disease and treatment*:* Another important distinction between symptoms and HRQOL is that symptoms have a direct relation to the underlying disease or condition being treated. In many cases, the symptoms define the disease or condition. Thus treating the symptoms is often the direct aim and target of treatment, and may be the primary basis of regulatory review and approval. In contrast, HRQOL is a benefit that accrues to the patient secondarily, often as a down-stream result of treating the disease and its symptoms. That is, HRQOL may improve precisely because symptoms are relieved, but not vice versa. Because symptoms are seen as the direct manifestation of the underlying disease or condition, and are often the direct and explicit target of treatment, symptom assessments (whether PRO or otherwise) can not only serve as the basis for label claims, but can and have also served as primary end-points for regulatory review and approval.

Communication of claims to prescribers and patients*:* Because PRO measures of symptoms often correspond closely to the terms that clinicians and patients naturally use to discuss the patient's condition, statements about symptom relief are typically easily understood by patients and clinicians. This is critical, as the purpose of labeling is ultimately to communicate to patients, clinicians, and others what the effects of the treatment are.

Dynamic nature of symptom end-points*:* Symptoms and HRQOL also may call for different approaches to assessment because symptoms may vary more dynamically than HRQOL. Whereas HRQOL is expected to change slowly and be relatively stable, and thus may be adequately captured by single pre- and post-treatment measures, symptoms can vary more quickly over time. Some symptoms may change rapidly in response to treatment, making it essential that assessments be scheduled frequently, and with appropriate timing to capture changes. As noted in the Draft Guidance (Lines 770-776), symptoms may also follow a particular natural history that should influence the scheduling of assessments. Symptoms may vary meaningfully even within a day (e.g., patients with gastroesophageal reflux disease tend to experience more symptoms in the evening hours), thus requiring a more dynamic assessment strategy that assesses symptoms repeatedly over time or target times of day when symptoms are more likely to be present or more severe. For this reason, diaries have often been used to gather data about symptoms that vary dynamically.

Summary*:* In summary, we agree with FDA that assessing complex, abstract, and multidimensional concepts places particular burdens on the developers and users of such measures. **We encourage the FDA to elaborate its distinction between measures of complex HRQOL concepts and measures of specific symptoms, which are often straightforward, concrete, readily interpretable, and unidimensional, and thus do not raise many of the same issues.**

Response to Evaluating PRO Instruments (Section IV., page 6)

The FDA is to be further commended for outlining the need for psychometric validation of PRO instruments. We also commend and support the Agency's taking a conceptual rather than highly prescriptive approach to the validation and revalidation of new and existing PRO instruments. **Consistent with the Agency's position, we believe that laying out highly specific, uniform procedures and benchmarks would be inappropriately restrictive and scientifically untenable.** The sheer numbers of a) existing PRO instruments, b) potential applications for new PRO tools, and c) unique measurement circumstances generates an infinite number of PRO measurement scenarios that simply cannot be accommodated by a single validation procedure. Indeed, as recognized in the draft, instrument validation is not an all-or-none phenomenon but rather an iterative process that aggregates data from a variety of sources in support (or not) of the use of a specific instrument or a more general measurement convention (e.g., self-administered questionnaires or visual analog scales). We have additional specific concerns about issues related to PRO instrument validation.

1. **PRO instrument validation evidence should be evaluated within the intended use in controlled clinical trials.**

The Draft Guidance states, "*The adequacy of a PRO instrument as a measure to support medical product claims depends on its developmental history and demonstrated measurement properties*" (Lines 171-172). **While we agree that developmental history and measurement properties are vital components of instrument evaluation, we feel that it is imperative for the Agency to acknowledge that these issues should be considered relative to the intended use of the instrument.** With respect to the Draft Guidance, PRO instruments are used to make comparisons between and inferences about randomized treatment groups and not individuals. This is important as the standards for the former tend be easier met than standards for the latter. This difference in standards is a direct result of the precision in measurement required in clinical practice that is typically not necessary in well controlled clinical trials (c.f., Cicchetti, 2001). There are numerous examples of this in the psychometric literature, one of the most widely used rules-of-thumb (in regard to reliability estimates) is provided by Nunnally & Bernstein (1994), which suggests a minimum score reliability cut-off value of .70 for the early stage of measure development, .80 for research purposes, and .90 for one-on-one clinical decisions (note that this is only one "rule-of-thumb" and more liberal and conservative suggestions have been offered in the literature).

2. **The Draft Guidance is unnecessarily restrictive in response to instrument modifications.**

The Draft Guidance states "*When considering an instrument that has been modified from the original, the FDA generally plans to evaluate the modified instrument just as it would a new one*" (Lines 178-179). **We are concerned that this section of the Draft Guidance may be read to imply an absolute standard with regard to any and all changes in PRO instruments.** As almost any anticipated use of a PRO assessment will require at least some modification of an existing tool (e.g., it will almost never be the case that the same instrument will be administered within precisely the same populations or under precisely the same conditions), this standard seems to imply that all instruments must be revalidated de novo by additional psychometric revalidation studies. This reading of the Guidance would lead to standards that are not justified by science and would be unduly burdensome, impractical, and unrealistic and likely would have a strong disincentive effect for use of PROs in clinical development.

Consistent with the science-based and flexible tone elsewhere in the Draft Guidance, we encourage the FDA to revisit their draft language to clarify that minor modifications to PRO instruments do not require extensive revalidation. For example, plainly simple changes like changes in font (so long as the fonts are readable) or the color or size of paper would not require any empirical revalidation of any sort. At the other extreme, substantial changes in the nature or content of items or the content or number of response options within an assessment can influence the performance of that assessment (e.g., Menon & Yorkston, 2000; Schwarz, 1999) and would, therefore, require considerable empirical evidence to support their validity or equivalence.

Further, as acknowledged in the Draft Guidance, validation is a dynamic and iterative process that may use information from a variety of sources. In this way, it is important to recognize that validation does not necessarily equate with the running of large or complex measurement-focused studies designed primarily to look at the psychometric issues of a given PRO instrument. In other words, there are a variety of procedures that can justify, substantiate, or otherwise "validate" the decision to select and use a given PRO instrument in some future trial. We laud the FDA for acknowledging a range of validation procedures (e.g., cognitive debriefing, equivalence studies). **However, it seems important that the Draft Guidance explicate more clearly a spectrum or hierarchy of validation procedures.**

We offer one possible "validation hierarchy" (Table 2) that is not exhaustive but instead reflects broad validation procedure domains under which a variety of specific procedures could be subsumed. This hierarchy of required re-validation steps ranges from none to cognitive debriefing, equivalence testing, and psychometric validation and revalidation studies. We are using the term cognitive debriefing in the same way as defined in the Draft Guidance (Lines 1042-1045). By equivalence testing, we mean procedures that evaluate the extent of correlation between a modified instrument and its original by a statistic such as the intraclass correlation coefficient. By psychometric validation and revalidation studies, we mean more extensive and sophisticated studies designed specifically to assess the psychometric properties of scores generated by novel or substantially modified PRO instruments.

The FDA clearly recognizes that different degrees of modifications of PRO instruments require different degrees of revalidation. It is essential that the Draft Guidance reflect this approach throughout, and avoid the absolute requirement to revalidate an instrument or treat it as a new instrument whenever a modification is made, regardless of its magnitude. **Therefore, we further suggest that the guidance should align the level of evidence needed to justify use of a modified instrument with the level of the modification made to the instrument.** Table 2 offers a potentially useful heuristic that proposes a level of evidence required to support among spectrum of common instrument modifications.

Table 2 PRO Instrument Modification and Level of Validation Evidence			
Level of Modification	**Justification Instrument use**	**Examples**	**Level of Evidence (Validation Hierarchy)**
Small	The modification can be adequately justified on the basis of logic and/or existing literature. No change in content or meaning.	1) Changes in font or font size or color of paper or ink (so long as result is readable) 2) Embedding in a battery of instruments	None
Medium	The modification can be adequately justified on the basis of existing empirical evidence. No change in content or meaning.	1) Language / cultural translation 2) Minor changes in mode of administration (e.g., from paper to electronic)	Cognitive debriefing
Large	The modification cannot be justified as neutral or on the basis of existing empirical evidence. May change content or meaning.	1) Changes in item wording (e.g., slight editing to shorten item) or response options (e.g., vertical v. horizontal VAS). 2) Large change in mode of administration (e.g., from paper to IVRS).	Equivalence testing
Substantial	The modification cannot be justified as neutral or on the basis of existing empirical evidence. Changed content or meaning	1) Substantial changes in item response options 2) Substantial changes in item wording 3) Computer-adaptive testing	Psychometric revalidation study

3. The Draft Guidance unjustifiably requires psychometric revalidation studies in measurement situations where no or a lesser degree of evidence is sufficient.

The Draft Guidance offers a set of examples as to what represents an instrument "modification," including a very general "other changes" category (Lines 579-670). As noted just above, not all instrument modifications require intensive psychometric revalidation studies. Therefore, we encourage the Agency to specifically note within the Draft Guidance that the level of validation evidence required of a PRO instrument modification can generally be obtained from a review of the empirical literature relevant to *both* the specific PRO instrument at hand and the type of modification proposed. Specific examples of PRO instrument modifications that we cannot endorse as requiring a psychometric revalidation study are identified below. Additionally we offer reasoning as to why revalidation is generally not necessary in those instances:

a) **Paper-and-pencil self-administered PRO is modified to be administered by computer or other electronic device (Line 636)**: We have conducted an extensive literature review that unequivocally contradicts the FDA's direction for revalidating PRO instruments based simply on modifying them from paper to electronic administration (Gwaltney, Shields, and Shiffman, in preparation; more detailed results available upon request); representative studies from these analyses include Greenwood et al., (2006),

Kleinman et al. (2005); Kvien et al. (2005); Saleh et al. (2002). Briefly, among 197 direct comparisons (obtained from over 40 unique studies) for the equivalence of means, over 90% demonstrated no significant difference between means and the test-retest reliability between paper-and-computer administrations was equal to that of the paper-and-paper administrations. Importantly, the platform used in the electronic administration (PDA and PC) and the presentation of a single item on-screen in the computer administrations (where multiple items are visible at the same time) did not moderate this effect (i.e., all produced scores that were equivalent with paper and pencil measures). Because this extensive evidence indicates psychometric equivalence, only minimal additional revalidation (i.e., cognitive testing) is necessary when migrating PRO instruments from paper to electronic administrations.

a1) **It is erroneous and misleading to categorize the administration of PRO assessments via text-based computer platforms (e.g., PDA, PC, laptop) along with PRO assessments administered via other technologically advanced methods like, for example, interactive voice response systems (IVRS).** While categorizing computer screen adaptations of existing PRO assessments along with PRO tools modified to accommodate other administration systems like IVRS is common, it is nevertheless rationally and empirically unjustified to do so. Upon critical examination, each of these different technologies raises unique testing issues that require different levels of re-validation. When a paper instrument is administered on a computer, there is little change in the respondents' experience and task: the item is read and one of several serially presented response options, all visible to the respondent, is endorsed. Alternatively, when a paper instrument is administered via IVRS system, items and response options are presented aurally and respondents must hold this information in memory while determining their unique response to be endorsed on a non-serial keypad. Even this briefly presented contrast demonstrates the clearly different tasks and cognitive loads required of subjects responding to computer and IVRS assessments. Importantly, the near-perfect equivalence observed in the review discussed above (Gwaltney et al., in preparation), are relevant only to text-based computer administration of existing PRO instruments and are not generalizable to IVRS. **Because of this evidence, we encourage the FDA to reconsider their decision to group text-based computer administrations along with other technologically advanced administration platforms like IVRS (Lines 636-638). Instead, we believe the Draft Guidance should offer a clear distinction between the more simple scenario of modifying a paper PRO instrument for computer administration and the more complex and less well understood testing and assessment challenges raised by IVRS. The latter may require increased levels of re-validation (see Table 1).**

a2) It is erroneous and misleading to categorize the administration of PRO assessments via text-based computer platforms (e.g., PDA, PC, laptop) along with computer-adaptive testing (CAT). We extend the argument made in point a1. above to the categorization of text-based computer administrations with computer-adaptive testing (CAT) models. That is, it is not rational or empirically justifiable to submit text-based computer administrations of PRO instruments to the same level of re-validation testing

as those necessary to fully understand CAT models. Text-based computer modifications of PRO instruments are often administered exactly like the paper versions. That is, they are fixed-item instruments and all respondents answer all the test items. By definition, CAT models alter both the testing procedures (e.g., number of items) and the content of what is being assessed via a programmed mechanism by which subsequent PRO instrument items are selected based on scores obtained on previous items. Theoretically, it is possible that no two respondents complete the same assessment. These issues alone (altered PRO assessment item number and content), typically not encountered in paper instruments modified for text-based computer administration, should prompt psychometric re-validation procedures (see Table 2). **It is for these reasons that we cannot endorse the categorical grouping of text-based computer administrations along with CAT, particularly with respect to the level of validation evidence necessary to ensure their reliable and valid use in clinical trials. CAT models will require additional levels of validation evidence whereas the equivalence of paper and text-based computer administrations has been established in over 40 studies using a variety of tests and scales and across a wide range of sample and methodological characteristics.**

b) **The PRO instrument was not developed and validated for use in a clinical trial (Line 666).** Many valid PRO instruments have not been developed in clinical trials. Importantly, we are aware of no evidence suggesting that instrument validity is directly and inversely influenced by the fact it was administered in a clinical trial setting.

c) **A PRO instrument developed and previously used as a stand-alone assessment is included as a part of a battery of measures (Line 668).** Most instruments are developed in isolation or as stand-alone assessment products and instrument combinations, PRO or otherwise, differ for virtually every trial. Even in instances in which instruments are developed with a battery of other measures, only rarely are these other instruments listed in the primary reports or instrument administration manuals. For these reasons, it is unrealistic and unwarranted to require revalidation in every case particularly in light of the lack of evidence suggesting that it is even necessary.

Reliability in PRO Assessment

The Guidance proposes psychometric reliability standards for PROs (Lines 484-497), yet, in our opinion, does not fully explicate the conceptual differences in the assessment of complex, trait-like entities (e.g., HRQOL) versus highly variable symptoms or discuss the implications this has for reliability. The Draft Guidance puts strong emphasis on obtaining high levels of test-retest reliability stating, “Test-retest reliability is the most important type of reliability for PRO instruments used in clinical trials” (Lines 491-492). This is sensible when evaluating trait-like constructs; however, many symptoms fluctuate significantly from day-to-day and even from hour-to-hour, for instance, pain and fatigue. This means that a comparison of scores generated from two dependent symptom measures (i.e., the same measure) from the same individual would be likely to yield relatively low (by conventional standards) test-retest reliability estimates. The interpretation of low reliability coefficients with trait measures is that the measure has substantial unsystematic measurement error. However, this should not be the case with symptom measures, because the low reliability is largely attributed to true or actual variation in the construct of interest (e.g., a symptom). **Thus, we suggest that the statements in the Guidance concerning reliability be modified to reflect these considerations.**

Additional Points of Guidance Support

The remainder of our comments represents general support for issues raised in the Draft Guidance. It is our hope that each of these issues receives similar treatment in the FDA's revised Guidance documents.

Hypothetical vs. actual behavior: We also agree with the Draft Guidance's decision regarding the importance of measuring actual behaviors, rather than perceived ability to perform behaviors in hypothetical situations, in PRO assessment schedules (Lines 302-308). While patients' perceptions of capability may be of some theoretical interest, actual performance seems the most appropriate end-point for studies of treatment.

Respondent burden and automated skip patterns: The Draft Guidance notes that respondent burden is an important issue to consider when PRO assessments are administered (Lines 432-458); we concur. One way of decreasing respondent burden is to limit the complexity and frequency of item skip patterns. However, sometimes more complicated item branching is required. In these instances, automated skip patterns, which can be implemented on electronic platforms, can routinely and easily determine appropriate skip patterns with no extra effort required of the respondent, and this can minimize respondent burden even for complex assessments.

Missing data: The Agency appropriately notes the challenges posed by missing data and further delineates a number of strategies for managing these situations (Lines 956-1017). **Because, as also noted in the Draft Guidance, these strategies are imperfect solutions (Lines 1005-1007) we fully support the FDA's position on preventing missing data as preferable to managing missing data after the fact (Lines 753-768)**. Indeed, this is such an important protocol, data management, data analytic, and substantive issue that it is our hope that proven methods known to reduce missing data (e.g., electronic data capture) be explicitly discussed in future versions of the Guidance.

Appropriate recall periods: The Draft Guidance states, "…it is important to consider patients' ability to accurately recall the information requested as proposed" (Lines 329-330) and further notes, "PRO instruments that require patients to rely on memory, especially if they must recall over a period of time, or to average their response over a period of time may threaten the accuracy of the PRO data. It is usually better to construct items that ask patients to describe their current state than to ask them to compare their current state with an earlier period or to attempt to average their experiences over a period of time" (Lines 339-343). We agree with this scientifically justifiable position and while it is difficult to stipulate a standard recall period across symptoms, the Guidance recommendation that it be minimized is consistent with the empirical literature on biases in recall (e.g., Baddeley, 1990; Christiansen & Loftus, 1991; Loftus & Marburger, 1983).

References

Baddeley, A. (1990). *Human memory: Theory and Practice*. London: Erlbaum.

Christiansen, S.A. & Loftus, E.F. (1991). Remembering emotional events: The fate of detailed information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17*, 1-12.

Ciccehetti, D.V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23*, 695-700.

Gwaltney, C., Shields, A.L., & Shiffman, S. (in preparation). Equivalence of electronic and paper-and-pencil administration of Patient Reported Outcome measures: A review.

Greenwood et al. (2006). *Rheumatology, 45*, 66–71.

Hammersely, R. (1994). A digest of memory phenomenon for addiction research. *Addiction, 89*, 283-293.

Kleinman et al. (2005). *Medical Care, 39*,181–189.

Kvien et al. (2005). *Annals of the Rheumatic Diseases, 64*, 1480-1484.

Menon, G. & Yorkston, E.A. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgements. In Stone, A.A., Turkkan, J.S., Bachrach, C.A., Jobe, J.B., Kurtzman, H.S., & Cain, V.S. (Eds.), *The Science of Self-Report: Implications for Research and Practice*. Erlbaum: Mahwah, NJ.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.)*. New York: McGraw-Hill.

Saleh et al. (2002). *Journal of Orthopaedic Research, 20*,1146–1151.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.