

### Docket Management Comment Form

Docket: 2006D-0044 - Draft Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims Availability

Temporary Comment Number: T74595

**Submitter:**

Dr. James Fries

Date: 03/31/06

**Organization:** Stanford University

**Category:** Academia

**Issue Areas/Comments**

**General**

See Attachment

**Attachments**

2006D-0044-T74595-Attach-1.doc

Print

Comment on Another Docket

Exit

**Print** - Print the comment

**Exit** - Leave the application

2006D-0044

C7

**Comments on the Draft Guidance Document  
Docket 2006D-0044 Submitted FDA**

**James F. Fries  
John E. Ware, Jr  
Jakob Bjorner  
Bonnie Bruce  
Matthias Rose  
Helen Hubert**

**Stanford University and QualityMetric**

**31 March 2006**

**The Draft Guidance document for Industry use of Patient-Reported Outcome Measures for Medical Product Development and to Support Labeling Claims dated February 2006** is a thoughtful document, reflects much discussion, and will promote and encourage use of PRO endpoints for clinical trials. We write from the perspective of long experience in developing, validating, and disseminating major PRO instruments for use in clinical trials by the NIH and others, being the developers of the SF-36 and the Health Assessment Questionnaire (HAQ). The HAQ Disability Index, for example, has been used in essentially all new drug development clinical studies in rheumatoid arthritis for over ten years. We believe that these instruments have served the FDA and Industry well, but that they may be further refined and improved, making them more reliable and sensitive to change, and requiring smaller sample sizes to achieve the same power in trials, using item banking, item improvement, item response theory, and computerized adaptive testing. We and colleagues have been deeply involved in such projects. FDA and industry adoption of improved PRO instruments will help ensure broad dissemination.

Certainly our goals and those described in the draft document are the same, to have valid, sensitive, proven PROs, broadly available, which may meaningfully assist the new drug development process; The draft document has aided our deliberations in a number of areas and we are appreciative. We would add that a major secondary goal is to reduce sample size requirements by reducing error terms associated with the measures.

The draft document contains non-binding recommendations, which are not intended to be construed as requirements but usually do acquire substantial weight and sometimes have acted like requirements, causing industry to be wary of deviating from them. There is only one small part of the document which we believe requires reconsideration and change, since as originally described it will substantially increase the measurement error and unintentionally greatly reduce validity. This involves the lines 302-308 on page 10 of the document. Here we note this problem area, comment on the issues it raises, and

suggest new wording to remedy the problems while retaining and even strengthening the intent.

**(From the Guidance Document) B. Creation of the PRO Instrument**

**1. Generation of items...(302-308).**

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

**Comments and Rationale:**

Clinical trials do not usually assess the concept ‘performance of daily activities’ but rather latent traits such as ‘physical function or disability’ where it is the ‘ability to do’ rather than the ‘performance’ that is the latent trait usually sought. Current instruments, such as the HAQ-DI and the SF-36 PF10, have been very helpful to the FDA in approving new drugs and new indications, and have been widely used for these purposes for many years. They seek to estimate the latent trait ‘physical function/disability’ by using items which are ‘ability’ or ‘limitation’ items. They have been validated in hundreds of studies and performance validated (e.g. observer vs questionnaire) in many studies. In trials they usually move in the same direction as biological outcomes, physician-assessed outcomes, and radiographs but are often more sensitive measures. The FDA arthritis group recognizes an ‘indication’ for ‘improvement in physical function’ which essentially requires a two year statistically significant improvement in HAQ-DI scores compared to comparator arms.

There are a number of reasons that a ‘performance’ requirement is problematic as a clinical trial endpoint. A central problem is that the ‘ability’ is always more sensitive than the ‘performance’ because of the difference between ‘could have’ and ‘did’. We all could do many things which we do not do, hence the mammoth problem of the ‘false negatives’ with ‘performance’ items. It is not unusual for 80 % or more of ‘performance’ item negatives to be false negatives! Clearly, this ascribes an erroneous value to the latent trait, and clearly, the sample size requirements increase dramatically.

Consider four simplified items very commonly used in physical function instruments as examples and comparing pure performance and pure capability responses. (numbers in parentheses are estimated percentage positives, illustrative of our ongoing studies of relatively healthy individuals, and will of course vary greatly across disease populations, but the differences are intuitively obvious)

1. [capability] Are you able to walk a block on level ground? (90 % yes)  
[performance] Over the past 7 days, did you walk a block on level ground? (88 % yes) Here, it seems to make little difference in healthy populations which item is used.
2. Are you able to jog or run two miles? (50 %) Over the past 7 days, did you jog or run two miles? (10%) Here, the performance item has a clearly unacceptable level of false negatives.
3. Are you able to climb several flights of stairs? (80%) Over the past 7 days did you climb several flights of stairs? (10%) Here, the performance item has too many false negatives.
4. Are you able to use a hammer to pound a nail? (90 %) Over the past 7 days did you use a hammer to pound a nail? (10%) Here again there are too many false negatives with the performance item. Of further interest, all of the performance "positives" in a recent focus group for this item were in male subjects.

Note that the capability questions are the latent trait suitable for a clinical trial endpoint of a treatment intended to improve physical functioning, such as an arthritis or cardiac or pulmonary drug, the performance type of question is not appropriate for this task but might appear suitable for a drug treatment designed to treat depression (although they would be problematic for the last three questions even in assessing depression). A survey research precept is that it is generally advantageous to estimate a latent variable by direct questioning.

In the first set of questions on walking a block on level ground, most people will answer positively to both versions or negatively to both, and the distinction of the wording is not particularly important. With the 'jog or run two miles' items it is not surprising that those who can do it may not do it very often, hence the major differences. With the 'climb several flights of stairs' items there is a surprisingly large majority of people who could do it but seldom do; most homes do not have several flights of stairs, ADA-qualified buildings have ramps and elevators, and using several flights of stairs in stores is unusual. With the 'hammer and nail' items most can do if they had to but seldom have to. These items are important because they involve the three most important 'mobility' items of walking, climbing stairs, and (at the floor) jogging or running, and the best 'strong grip' item, using a hammer. Differences would be even greater if a one day time frame had been used for the 'performance' items.

In our recent review of 1860 physical function items from 165 instruments (in press), 85 % were classified as 'capability' and only 6 % as 'performance', indicating that de facto

use is 'capability'. Of interest, these classifications were applied very differently by a number of outside reviewers. For example, many considered the SF-36 PF10 items (Because of your health, how much are you limited in... 'yes, limited a lot; yes, limited a little; no, not limited at all') as 'performance' items, whereas our primary review group required a past-tense "did" for an item to be classified as 'performance' and thus considered the PF-10 to consist of 'capability' items. So, much of the contention over issues of 'capability' versus 'performance' may be semantic and moot.

Use of the present rather than the past tense has other advantages, such as brevity, clarity, lower reading level, more reliable translation, easier cultural adaptation, no need to integrate over a time period, no recall bias. Patients surveyed (n=1200) indicated that the perceived clarity of 'performance' items was substantially and significantly less than that of corresponding 'capability' items.

'Performance' items are less frequently positive than 'capability items', and this has been suggested as an advantage for such items, since subjects may believe (and report) that they are able to do things that in fact they cannot do. To quantitate and minimize this effect we recommend 'performance testing' of items, where the performance comparison is between a questionnaire response and subsequent testing by a physical therapist or nurse who observes the subject attempting to perform the activity. Such testing with HAQ-DI items typically shows about a 6% trend toward poorer performance than reported. In a clinical trial situation this does not have much of an effect because the usual outcome measure is the 'change score' and the typical subject with a high estimate of abilities reports similarly high at both measurement times.

We note that validation of 'performance' items, which refer to past events, is often very difficult, and that 'performance' items also can be answered positively when the activity has not been performed as reported. This is an attribute of all PRO items and we believe indicates that items need performance validation which includes objective observation of the subject performing a task.

#### **Suggested re-wording for lines 302/308**

**"Items should be appropriately selected to accurately access the desired domain construct or 'latent trait', which could be 'physical function/disability', 'pain', 'fatigue', 'emotional distress', or others. Time frames, response categories, and context should be appropriate for the particular domain. Patient evaluations may be validated against external observation, e.g actual observed ability to perform a described activity. Such external validation provides a strong test of the measurement instrument and is recommended, when feasible. Use of items which are developed from very well-validated 'legacy' instruments or the instruments themselves is encouraged, especially when the FDA has a substantial experience with these instruments and items. Item banks being developed by current research activities are anticipated to provide documentation of reliability, validity, performance, information content, and other item response theory attributes of many items in many domains."**