

The draft FDA guidance document on PROMs is a major step forward; the team is to be congratulated for producing this excellent draft. Our comments for improving the guidance document are as follows:

Introductory remarks

The guidance document is likely to assume an even wider remit than the authors may have intended, which has implications that need to be kept in mind by those who will revise the draft to produce the final version. That is, as a new and long-awaited document that offers a comprehensive set of methodological guidelines re: the development and validation of PROMs, the guidance will undoubtedly be widely adopted by a range of stakeholders including those in industry, academia/research, clinical practice and policy making. In this sense, the document and the guidance it offers are likely to become reified in a way the authors may not have expected, i.e. statements in the guidance will quickly become “carved in stone”, assume the status of “gospel”, and be taken as the last word on the subject. One can easily imagine that much of the methodological guidance will eventually find its way into textbooks, academic publications, and clinical output in the general area of measurement and PROMs.

It is for this reason that it is essential that: i) the evidence base underlying the guidance is airtight and can stand up to rigorous scientific scrutiny, and ii) that the terms, concepts and language used throughout the document are clear and unambiguous. With respect to being on a firm evidence base, the document in large part achieves this. However, our concerns focus primarily on aspects of the guidance where the evidence base is either unclear, not unequivocal, or does not exist. With respect to clarity of terms, concepts and language, we offer suggestions for improvement. Specific comments include:

1. A distinction needs to be clearly made in the document between guidance that is evidence-based versus guidance based on consensus/expert opinion. For example, the basic principles and methods of reliability are well-established and have a strong evidence base. However, comments such as “Test-retest reliability is the most important type of reliability for PRO instruments used in clinical trials” (lines 491-492) are based strictly on opinion not evidence, and would clearly be disputed. Similarly, much of the guidance on the very important issue re: defining a MID (lines 537-577) is not based on a clear evidence base, nor is the guidance about weighting items (lines 416-430).
2. Related to this is the need for the document to make clear those aspects of the guidance for which there is currently no clear evidence base, and for which evidence needs to be gathered in order to advance methodological developments in this area. What would be most helpful is if the guidance included a section at the end on future research, unanswered questions, etc. which summarises specific methodological questions for which we currently do not have clear evidence and consequently must rely solely on consensus/expert opinion. Not only would this highlight and make

clear to readers those aspects of the guidance on which “the jury is still out”, but a clear indication of where further methodological research is needed would help set and define a much needed research agenda in the area of methodological aspects of PROMs. With so many unanswered methodological questions in this area, and with funding for such needed methodological work so notoriously difficult to obtain, a clear statement in this important FDA document about future research needs in this area would undoubtedly make funding bodies take note and more likely to consider funding such work.

3. Table 4 provides a good summary of measurement properties, but would be strengthened considerably by providing more specific, explicit, evidence-based guidance indicating the criteria for acceptability for each measurement property and test. The same comment applies to Table 3. An illustration of the type of more explicit guidance that would be useful to readers/users can be found in our published work in this area (available on request; also see Table 1 following these comments for an example of this approach):

Lamping, D.L., Schroter, S., Marquis, P., Marrel, A., Duprat-Lomon, I., & Sagnier, P.-P. (2002). The Community-Acquired Pneumonia Symptom questionnaire: A new, patient-based outcome measure to evaluate symptoms in patients with community-acquired pneumonia. *Chest*, 122, 920-929.

Lamping, D.L., Schroter, S., Kurz, X., Kahn, S.R., & Abenhaim, L. (2003). Evaluation of outcomes in chronic venous disorders of the leg: Development of a scientifically rigorous, patient-reported measure of symptoms and quality of life. *Journal of Vascular Surgery*, 37, 410-419.

Hilari, K., Byng, S., Lamping, D.L., & Smith, S.C. (2003). Stroke and Aphasia Quality of Life scale-39 (SAQOL-39): Evaluation of acceptability, reliability and validity. *Stroke*, 34, 1944-1950.

Schroter, S., & Lamping, D.L. (2004). Coronary Revascularization Outcome Questionnaire (CROQ): Development and validation of a new, patient-based measure of outcome in coronary bypass surgery and angioplasty. *Heart*, 90, 1460-1466.

Smith, S.C., Lamping, D.L., Banerjee, S., Harwood, R., Foley, B., Smith, P., Cook, J.C, Murray, J., Prince, M., Levin, E., Mann, A., & Knapp, M. (2005). Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technology Assessment*, 9 (10).

4. The language needs to be considerably tightened up, so that consistent terminology is used throughout. Use of related but different terms is likely to be confusing, particularly to novice readers/users, e.g.
 “measure” vs. “instrument” vs. “PRO” vs. “test” vs. “scale”
 “items” vs. “questions”
 “assessment” vs. “measurement”
 “concept” vs. “construct”
5. The document confuses conceptual models and measurement models. This is an important distinction which needs to be clarified in the final draft.
6. The guidance is not entirely clear about what qualifies as a PRO. The mention of both scales and single-item/single-rating measures suggests that either can be

considered a PRO. But given the focus on the psychometric approach to instrument development and validation, aren't we really using the term PRO to refer to multi-item scales? The mention of single item measures may confuse readers. The section on "Formal Assessment May be More Reliable than Informal Interview" (lines 120-137) may be confusing to readers, who may assume that the type of single-item question asked by clinicians "Do you cough at night" is a PRO.

7. Related to this, the guidance needs to be clearer about the relationship between PROs and traditional clinical outcome measures (including physician ratings), i.e. that they cannot substitute for each other but that PROs provide a complementary perspective on outcomes that clinician-based tools cannot reach. Making more explicit the distinction between manifest and latent variables, proximal vs. distal disease processes—as so clearly described in the seminal Wilson & Cleary paper—might help in clarifying this point.
8. It may be useful to readers to include a more explicit discussion about what to measure or which instrument to choose, i.e. the importance of matching the choice of outcome measure to explicit treatment objective(s).

Comments submitted jointly by:

Donna L Lamping, PhD
Reader in Psychology and Director, Research Degrees Programme
London School of Hygiene & Tropical Medicine
London, UK

Stefan Cano, PhD
Lecturer in Neurological Outcomes Measurement
Institute of Neurology, University College London
London, UK

Table 1 Psychometric Tests and Criteria

Psychometric Property	Definition/Test	Criteria for Acceptability
1. Item analysis/reduction	<p>identify items for possible elimination due to weak psychometric performance;^a assessed on the basis of:</p> <ul style="list-style-type: none"> unrotated principal component factor analysis (to determine whether all 18 items are measuring a single factor) item analyses for all 18 items 	<p><i>principal component factor analysis:</i></p> <ul style="list-style-type: none"> all items should load on the first unrotated factor >0.30 <p><i>applied to all 18 items:</i></p> <ul style="list-style-type: none"> missing data $<5\%$ no item redundancy (inter-item correlations <0.75) item-total correlations ≥ 0.25 evidence of item responsiveness as assessed by significant improvement between baseline and test of cure assessments maximum endorsement frequencies $<80\%$ (i.e. the proportion of respondents who endorse each response category), including floor/ceiling effects $<80\%$ (i.e. response categories with high endorsement rates at the bottom/top ends of the scale, respectively) aggregate adjacent endorsement frequencies $>10\%$
2. Acceptability	the quality of data; assessed by completeness of data and score distributions	<p><i>applied to items:</i></p> <ul style="list-style-type: none"> missing data $<5\%$ maximum endorsement frequencies $<80\%$ (see above), including floor/ceiling effects $<80\%$ (see above) <p><i>applied to summary scores:</i></p> <ul style="list-style-type: none"> missing data $<5\%$ floor/ceiling effects $<80\%$ skewness values between $+1$ to -1
3. Reliability		
3.1 Internal consistency	the extent to which items comprising a scale measure the same construct (e.g. homogeneity of the scale); assessed by Cronbach's alphas ⁴⁴ and item-total correlations	<ul style="list-style-type: none"> Cronbach's alphas for summary scores $> 0.70$⁴⁴ item-total correlations $\geq 0.25$²⁵
3.2 Test-retest reliability	the stability of a measuring instrument; assessed by administering the instrument to respondents on two different occasions and examining the correlation between test and retest scores ^b	<ul style="list-style-type: none"> intraclass correlation coefficients for summary scores $> 0.80$²⁵
4. Validity		

4.1 Content validity	the extent to which the content of a scale is representative of the conceptual domain it is intended to cover; ^c assessed qualitatively during the questionnaire development stage through pre-testing with patients, expert opinion, and literature review	<ul style="list-style-type: none"> qualitative evidence from pre-testing with patients, expert opinion, and literature review that items in the scale are representative of CAP symptoms
4.2 Construct validity		
4.2.1 Within-scale analyses	evidence that a single entity (construct) is being measured and that items can be combined to form a summary score; assessed on the basis of evidence of good internal consistency, moderately high item-total correlations, and results from principal component factor analysis	<ul style="list-style-type: none"> internal consistency (Cronbach's alpha) > 0.70 item-total correlations ≥ 0.25 evidence from factor analysis that a single construct is being measured
4.2.2 Analyses against external criteria		
4.2.2.1 Known group differences/hypothesis testing	<p>the ability of a scale to differentiate known groups; assessed by comparing CAP-Sym scores of patients defined as clinically cured, according to the clinical variable "clinical evaluation of cure" between baseline and the day 7-10 (test of cure) assessments, with those of patients defined as clinical failures</p> <p><i>Note:</i> the comparative validity of the disease-specific CAP-Sym against the generic SF-36 was also evaluated by assessing the ability of the SF-36 Vitality scale to differentiate patients defined as clinical cure/failure</p>	<ul style="list-style-type: none"> CAP-Sym scores should be significantly higher (i.e. higher symptom bother) in patients in the clinical failure group than in patients in the clinically cured group SF-36 Vitality scores should be significantly lower (i.e. lower energy) in patients defined as clinically cured vs. clinical failures
4.2.2.2 Convergent validity	evidence that the scale is correlated with other measures of the same or similar constructs; assessed on the basis of correlations between CAP-Sym scores and other patient-based (SF-36) and clinical (temperature, Pneumonia Severity Index) outcome measures	<p>criteria for acceptability depend on the degree of conceptual similarity between the CAP-Sym scale and the other validation measures. Specific hypotheses:</p> <p><i>for patient-based outcome measures</i></p> <ul style="list-style-type: none"> moderate correlations between the CAP-Sym and SF-36 (because the two instruments are measuring constructs that are related but distinct--symptoms vs. quality of life) higher correlations between the CAP-Sym and SF-36 PCS/Vitality scores than between the CAP-Sym and SF-36 MCS scores because the CAP-Sym is more closely related to physical than mental health <p><i>for clinical measures</i></p> <ul style="list-style-type: none"> low correlations between CAP-Sym and temperature and

4.2.2.3 Discriminant validity	evidence that the scale is not correlated with other measures of different constructs; assessed on the basis of correlations with age and sex	<p>the PSI⁴⁰⁻⁴²</p> <ul style="list-style-type: none"> low correlations between CAP-Sym scores and age and sex
5. Responsiveness	<p>the ability of a scale to detect clinically significant change following a treatment of known efficacy;⁴⁵⁻⁴⁶ assessed by comparing mean scores for change in CAP-Sym scores at three assessment points (i.e. between baseline and day 3-5, day 7-10, and day 28-35) using two standard methods:</p> <ul style="list-style-type: none"> effect size, calculated for responsiveness at the three assessment points as the mean difference (change score) in symptom scores from baseline to follow-up divided by the standard deviation of the baseline score; effect sizes and standardized response means of 0.20 are considered small, 0.50 moderate and 0.80 or greater as large.⁴⁶ standardized response mean, calculated for responsiveness at the three assessment points as the mean difference (change score) in symptom scores from baseline to follow-up divided by the standard deviation of the change score <p><i>Note:</i> the comparative responsiveness of the disease-specific CAP-Sym against the generic SF-36 was assessed by comparing effect sizes</p>	<ul style="list-style-type: none"> effect sizes and standardized response means should increase in magnitude across time, i.e. CAP-Sym and SF-36 scores should improve over time larger effect sizes indicate better responsiveness

^aAn standard item reduction strategy was used to identify and eliminate items from the questionnaire that showed weak psychometric properties. To test the robustness of the item reduction strategy, cross validation analyses using the same tests and criteria were performed separately on two random split-half subsamples from the pooled dataset. Results of the item reduction analyses performed on the two randomly selected subsamples were then compared to results obtained in the pooled sample. ^bThe length of the test-retest interval must be short enough to ensure that clinical change in the symptom being measured is unlikely to occur, but sufficiently long to ensure that respondents do not recall their responses from the first assessment. In conditions such as CAP, where rapid changes in symptoms are expected to occur over a very brief time (i.e. within a few hours), a very short test-retest interval of 1-2 hours is necessary. This ensures that stability per se is being evaluated, rather than clinical change in symptoms during the test-retest interval, which will underestimate reliability. ^cA scale to measure CAP-related symptoms should include questions based on the wide range of symptoms that characterize the condition. If a CAP symptom questionnaire did not include an item about cough, content validity might be considered doubtful as an important dimension of the condition had been excluded.