The Patient Reported Outcomes Measurement Information System (PROMIS®) Perspective on:

Universally-Relevant vs. Disease-Attributed Scales

By the PROMIS Statistical Center Working Group* on behalf of its Steering Committee

Background

The PROMIS Objective

A high priority of the U.S. National Institutes of Health is enhancement of the national clinical outcomes research enterprise by making available a dynamic system of psychometrically sound and efficient measures of patient-reported outcomes (PROs) for people with a wide range of chronic diseases and demographic characteristics. The PROMIS 1 RFA (http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-011.html), released on November 18, 2003, specified several research objectives, including:

> *To identify a core set of questions, derived primarily from existing, commonly-used instruments and supplemented by new and revised items, that will address the most common or salient dimensions of patient-relevant outcomes for the widest possible range of chronic disorders and diseases, and to collect responses to these items in a large, diverse sample. Domains of interest include, but are not limited to, self-reported: symptoms, physical functioning, participation in activities, social functioning; cognitive functioning, and emotional status. Of special interest are the assessment of pain severity, frequency, and impact and the assessment of fatigue as clinical outcomes of high importance to many people suffering from chronic diseases.*

There are compelling reasons to desire "universal" measures of self-reported health, including the ability to compare disease burden and treatment impact across various chronic conditions, and to be able to measure patient-relevant outcomes, using a common metric, with people who have more than one disease or medical condition. [1-3] The PROMIS Steering Committee endorses this vision of universally-relevant measures with a common metric and considers it a fundamental and defining characteristic of PROMIS. As expressed in the original charge to PROMIS investigators, measures in the PROMIS portfolio should be widely relevant to multiple chronic diseases (i.e., universal). As such, PROMIS measures typically do not include items with specific disease attributions in their stem.

Early in PROMIS 1, the decision was made to report scores on a T-score metric (mean=50, SD=10) and to center this metric based on a sample representative of the 2000 U.S. General Census with respect to important demographic variables (gender, age, race/ethnicity, education, marital status, and income).[4] Centering on a representative general population sample facilitates normative comparisons and provides an interpretative context for scores. This metric is consistent with the vision of universally-relevant measures whose scores can be compared across diseases and conditions. Another decision that was made was to use item response theory (IRT) models for calibrating items. IRT

*Karon Cook, Michael Kallen, David Cella, Paul Crane, Basil Eldadah, Ron Hays, Laura Lee Johnson, Carol Moinpour, Paul Pilkonis, Dennis Revicki, David Tulsky, James Witter

models estimate the mathematical relationship between how much individuals possess of a targeted domain and their responses to items.

Reissued in 2009, PROMIS 2 remained consistent with this vision. The ability to compare PROMIS scores across diseases and conditions and to the U.S. general population remained a priority in PROMIS 2. Among PROMIS 2 projects was the collection of responses to existing adult item banks from a sample of 3,000 individuals recruited through an online panel vendor. Data collection for this effort was completed in August 2013 and will be used to center the metric of PROMIS scores based on the 2010 U.S. General Census with respect to gender, race, ethnicity, and educational level.

PROMIS Expansion

In PROMIS 1, initial domains were selected to address a broad range of demographic and clinical groups, including those with multiple co-morbidities. With PROMIS 2 came an increased number of investigators, target outcomes, and clinical populations. This expansion introduced scientific questions about the role and the content validity of measures that were designed without reference to particular clinical populations. Questions included:

- Is a measure that does not include disease attribution likely to be as responsive to change as one that does include attribution?
- Can a measure that is developed based on input and responses from the general population and from an amalgam of chronic disease populations be effectively used in a group of persons with a single chronic disease, such as multiple sclerosis or fibromyalgia?
- Does measurement in some chronic disease groups improve with the addition of items to, or deletion of items from, existing PROMIS item banks?
- Do the PROMIS item banks have sufficient items targeting high and low levels of a domain? Can short forms or computer adaptive test (CAT) administrations be developed that are reliable, valid, and sensitive when used in chronic disease groups who experience very high or very low levels of the measured domain?
- Do different chronic disease groups require different IRT item calibrations?

In this document, we provide the PROMIS perspective on these questions and related topics.

Definition of Terms

Each of the above questions proceeds from a broader one—"What are the roles, functions, usefulness, and distinctions between disease-specific and universally-relevant scales?" Discussions of these issues often are impeded by a lack of precision or agreement regarding how terms are used and understood. Therefore, we begin with some definitions.

Condition or Disease-*Attributed* Measures

Condition or disease-attributed measures are measures containing items, responses, or other parts of the measurement that specify the etiology of the outcome being measured. For example, the item stem, "Because of my arthritis, I am too tired to do my usual activities," requires respondents to reference their fatigue to that which can be attributed to their arthritis. Disease-attributed measures require respondents to distinguish among potential causes for the symptom or outcome of interest and identify the part of that symptom or outcome that is due to the specified condition or disease. Responding to such items may be challenging for individuals with comorbidities, given the complexity of partitioning the impact of multiple clinical conditions on functioning and well-being. Furthermore, research has found that patients' causal attributions for symptoms are quite complex, influenced by psychological, physical, environmental, and cultural factors, as well as by age. (Note: A summary of influencing factors is discussed by Siegel and colleagues.[5])

Condition or Disease-*Relevant* Outcomes

A domain is condition or disease-relevant if individuals living with the condition or disease consider it (i.e., declare it) to be relevant. The perspective of clinicians caring for these patients is also important when defining the scope of relevant domains. In the case of condition-relevant outcomes, the domain need not be attributed to the condition or disease, but would be regarded as relevant to assess if people with that condition or disease have made it clear that it is important. This is usually determined through qualitative and observational research.

Universally (Widely) Relevant Outcomes

A common symptom like pain or fatigue can be relevant to more than one disease or condition and, therefore, measures of those domains can be relevant across many clinical populations. For such measures, we use the terms "universally relevant" or "widely relevant." This terminology acknowledges that there are outcomes and symptoms that individuals with different clinical conditions find relevant. Examples of such outcomes include pain, fatigue, physical function, depression, and social function.

Research Guide for Evaluating Relevance of the PROMIS Vision In Specific Clinical Populations

The PROMIS vision is one of universally-relevant measures whose scores can be compared across diseases and conditions. This vision has generated thoughtful conceptual and theoretical challenges. The questions bulleted above (see "PROMIS Expansion") proceed from such challenges. The scientific integrity of PROMIS requires the encouragement and accommodation of research that systematically evaluates the practicality and potential limits of the PROMIS vision. A basic and testable property of item response theory model calibrations is that they are sample independent. When this is achieved, scores are comparable across different populations. When empirical research finds this not to be the case, item bank modifications are warranted. Below we list several types of empirical results that would indicate the need for changes in the PROMIS measures. Some could perceive such empirical findings to be a challenge to the PROMIS vision. However, we recommend strategies for incorporating such findings without losing what is fundamental about PROMIS: universality of measured outcomes and wide comparability of obtained scores.

Potential Empirical Findings

Estimated Item Parameters for a Given Group Vary from the Published PROMIS Parameters

Comparisons of PROMIS scores across demographic and clinical groups require that item parameter estimates are invariant across the compared populations. Differential item functioning (DIF) detection methods are elegant approaches for evaluating, at the item level, whether this is the case. DIF occurs when the probability of responding within the individual response categories of an item differ by some characteristic other than a person's underlying level of the construct being measured. For example, in developing the PROMIS Depression item bank, in one item women were found to be more likely to report "crying" than men, even after controlling for level of depression. Because of the observed magnitude of its DIF, this item was excluded from the bank.[6] The PROMIS measures have been systematically evaluated for DIF with respect to several important demographic variables (e.g., age, gender, and education), and they have been found to be generally invariant. In the initial item bank development as part of the PROMIS I initiative, 643 items were evaluated for DIF; 14 items were excluded based on DIF magnitude.

Further research with diverse samples may identify items whose item parameters differ by subgroup membership, including membership in specific disease groups. One approach to such a finding would be wholesale recalibration of the item bank in a specific clinical subgroup. However, if all items of a PROMIS measure are re-calibrated based on responses from a single clinical condition or other disease-specific subsample, the resulting measure scores are no

longer on the PROMIS metric. They no longer are comparable across disease conditions, nor are the scores interpretable in the context of the general population. This disengagement from the established PROMIS metric would result in a loss of score meaning and interpretation. Thus, wholesale re-parameterization sacrifices the hallmark advantages of PROMIS scores, and we therefore strongly recommend against it.

There are alternative strategies for addressing instances of identified measurement variance. DIF can be accommodated without wholesale recalibration. One such strategy when impactful DIF is identified is to estimate subgroup-specific item parameters *only for those items exhibiting DIF*. Above we described the decision to exclude an item from the PROMIS Depression item bank because of DIF by gender. An alternative would have been to use a scoring algorithm that applies gender-specific item parameters based on the sex of the respondent. This approach would retain the item and also maintain the comparability of PROMIS scores while accounting for the impact of DIF. This strategy was used in developing the Spanish-language version of the PROMIS Physical Function item bank. Based on co-calibration of PROMIS Wave 1 English-language data with a sample of 640 Spanish-language respondents, items were evaluated for DIF by language. Some items were found to have "substantial DIF" (defined in this study as a pseudo R^2 difference greater than or equal to 0.02). [7] For the Spanish item bank, English-language calibrations were used for all DIF-free items; Spanish calibrations were used for items with substantial DIF. By using this strategy, the scores of English-speaking patients can still be compared to those of Spanish-speaking patients, even though not every item in the Physical Function bank has the same item parameter estimates for English vs. Spanish respondents.

Finally, a few cautions with regard to DIF: There are several different techniques for identifying DIF, both in terms of its statistical significance and its magnitude. These different techniques may give different answers[6], and we suggest not relying on a single technique but rather cross-validating with multiple methods. Cross-validation may be especially important as sample sizes increase, with larger samples sometimes yielding DIF results that are statistically significant but clinically inconsequential.

The Meaning of an Item Changes over Time

An item bank is intended to be a living entity in which new items are developed, others are brought in, and still others may be dropped as the need arises. Sometimes the content of an item can take on a different meaning for respondents over time because of social, political, or linguistic changes. This phenomenon has been observed in other widely used measures, such as the Minnesota Multiphasic Personality Inventory-2 (MMPI-2).[8] When item meaning shifts occur, an item's parameters could be re-estimated in order to maintain it in the bank. An item whose meaning has more fundamentally changed might, instead, be dropped from the item bank.

Standard Approaches Prove Suboptimal for a Given Group

Creating and implementing measures using the item banking approach allows us to effectively manage variations in measurement context. For example, short forms can be administered as an alternative to CAT when computers are not available to respondents. CAT stopping rules can be varied to help balance concerns about precision versus response burden.

CAT starting rules also can be varied. When a PROMIS CAT is administered to samples whose trait distribution is known to differ from that of the general population, a more informative prior can be selected for the CAT scoring algorithm. Currently, the CAT algorithm in PROMIS Assessment Center begins a CAT assessment using a normal prior. This is reasonable since, if nothing is specifically known about a respondent, the best guess for an assessment's starting point is that the respondent's trait level is near the mean trait level of the overall population on the domain being measured. However, if the PROMIS Fatigue CAT is being administered to a sample of persons with, for example, fibromyalgia, the usual normal prior would not be the most informative. CAT assessment software often allows users to set different priors that take into account what is known or becomes known about the distribution of a trait in a given subgroup.

Item(s) Found to Be Inappropriate for a Given Clinical Population

Some items may be found to be inappropriate for particular diseases or conditions. For example, physical function questions that ask about walking and climbing stairs are not informative, and at times even inappropriate, for people who cannot walk or climb.[9] One method for handling such a situation is to identify, from an existing PROMIS item bank, the subset of items appropriate for a particular population. In addition, new items that target the population can be added and linked to the established PROMIS metric. This strategy was successfully used by PROMIS to develop a physical function item bank for users of mobility aids (Grant U01AR052171 Amtmann, PI). The advantage of this strategy is that the modified bank retains comparability with PROMIS scores while addressing the particular measurement needs of a specific subgroup, here, mobility aid users.

Existing Bank Items Fail to Target a Subgroup's Range of Concerns or Level of Outcome

A PROMIS item bank may be found to have insufficient items to distinguish among individuals in a particular subgroup. For example, though the PROMIS Physical Function item bank was shown to work well in general, it exhibited modest floor effects in individuals with very poor function (e.g., nursing home residents) and ceiling effects in healthier populations (e.g., elite athletes). In such situations, new items can be developed, calibrated, and linked to the original PROMIS metric. This approach extends the range of the PROMIS measure while maintaining the ability to compare scores across the continuum and across differing patient subpopulations. Currently, data have been collected on new physical function items, and analyses are planned to inform selection and incorporation of these into the existing item bank.

Recommendations for Planning and Conducting Research

Researchers with interest in particular clinical populations may wonder whether PROMIS item banks might be modified or recalibrated, with the hope of making them more responsive to the needs of people with a specific condition or disease. Potential steps these researchers might take could include adding disease-attributions or recalibrating entire banks in a particular clinical sample. However, before taking such steps, we recommend the following:

Before Choosing to Use or Create a Disease-Attributed Measure

- Approximately 26% of U.S. adults have multiple chronic conditions.[10] Consider if respondents will have difficulty identifying impact due to a particular disease, especially given existing comorbidities. For example, what is the basis for assuming that respondents can differentiate between their fibromyalgia pain and their rheumatoid arthritis pain or distinguish between difficulty with physical function due to their heart failure or due to their obesity?
- Consider the loss of ability to compare scores with those from other studies that would then be using different (i.e., PROMIS vs. PROMIS-like) instruments.
- Consider administering both PROMIS and disease-attributed measures in order to preserve comparisons with other diseases and the general population while gaining more specific information about the target disease. This will also enable testing of the often-expressed assumption that disease-attributed assessment is more responsive than the PROMIS disease-relevant approach.

Before Calibrating PROMIS Items in Disease-Specific Samples

- Recognize that scores from a measure using disease-specific calibrations may not be comparable to PROMIS scores.
    - Consider the loss of ability to compare scores with scores calculated using data from other diseases.
    - Consider, instead, employing a hybrid approach such as that used by Paz and colleagues.[7] This approach preserves comparability by having non-DIF items fixed to the general group item parameters and DIF-identified item parameters equated to the metric of those of the general group (see above section, "Estimated Item Parameters for a Given Group Vary from the Published PROMIS Parameters").

- As an alternative, consider an approach where disease-specific calibrations are obtained and utilized for administration and scoring, but then the derived scores are linked to the score one would obtain from PROMIS general population calibrations. If the scores obtained through general population versus disease-specific calibrations do not differ significantly, this would provide reassurance that the scores are comparable regardless of calibrations used.

- Conduct DIF analyses and evaluate findings with respect to DIF impact (i.e., not relying solely on DIF identification via statistical significance). For example, statistically significant DIF may result in trivial measurement bias, which can be investigated by comparing impact of DIF on individual and group scores. For more discussion, see the PROMIS Standards document.
- Evaluate the ability to retain PROMIS item parameters by changing the prior information available in the measurement model based upon responses from people with the specific disease or condition in question.

Suggestions for Future Research

Specific Research Questions

We recommend a research agenda that begins to address the following questions:

- Are the PROMIS measures, which were developed based on input and responses from individuals from the general population and with an amalgam of chronic diseases, universally relevant for persons identified by a single clinical condition (e.g., multiple sclerosis, fibromyalgia)?
- Do the PROMIS measures capture the important content embodied in each domain for persons with specific clinical conditions? If so, how would we know? These questions are usually addressed through qualitative research with people diagnosed with the specific condition.
- How could clinical studies that compare both PROMIS and disease-attributed measures in the same patients help address gaps in our understanding of the role for these measures to assess health outcomes relevant to researchers and regulators?

These questions can be divided into the following sub-questions.

- For a target domain of wide or universal relevance (e.g., fatigue, physical function, emotional well-being, social function, pain), are measures developed based on input from a single clinical group substantially more relevant to that group than measures developed based on more population-inclusive input?
- For a target domain of wide or universal relevance, do measures developed with input from many clinical groups "leave out" important domain content *specific to a single clinical group?*
- For a target domain of wide or universal relevance, is the construct it assesses (e.g., fatigue, physical function, emotional well-being, social function, pain), defined differently by different clinical populations?
- For a target domain of wide or universal relevance, is there evidence indicating that PROMIS measures display DIF that has non-trivial impact on scoring in studies involving different clinical populations?
- Are patients from a particular clinical population able to make disease-based or treatment-specific attributions for common symptoms such as pain or fatigue? If so, how do they make this determination, and how do they respond in the context of multiple comorbidities? If they are not able to make these distinctions, what implications does this have for regulatory science? This can be identified through careful cognitive debriefing of individuals, who read and respond aloud to answering questionnaire items, with interviewer follow-up.
- To what degree does disease severity affect the way patients with a specific condition respond to these items (e.g., COPD directly after an acute exacerbation)?

- Do disease-attributed or condition-specific (targeted) questionnaires offer any measurable advantages to PROMIS measures? Are they comparable in performance? In what settings are PROMIS measures more relevant or responsive? To evaluate these questions, administer both PROMIS and disease-attributed measures and compare their performance in the same cohorts, preferably with controlled longitudinal designs such as randomized clinical trials. This will enable testing hypotheses regarding relative validity (responsiveness) of targeted versus generic assessment, in the context of using PROMIS in specific disease groups.
- When both PROMIS and disease-attributed measures are co-administered, test the assumption that disease-attributed assessment is more responsive than the PROMIS domain-focused approach

Theoretical Considerations

Existing research in the field should be considered, as should prior literature regarding relevant theoretical considerations. Researchers may choose to use these considerations to select disease-relevant domains to evaluate. We anticipate that studies evaluating the relevance of PROMIS measures for persons identified by a single clinical diagnosis (e.g., MS; fibromyalgia) likely would require substantial qualitative as well as quantitative work.

Relevant Prior Research

An example of prior research in this area is an investigation by Cook, Amtmann and colleagues. [11] In this study, individuals with MS identified from the PROMIS Fatigue item bank the items they thought were most relevant to their experience of fatigue. An MS-specific short form was created based on their results. Scores based on this 8-item short form were then correlated with scores estimated using an existing 7-item PROMIS Fatigue short form and compared in terms of ability to distinguish known groups. While similar in length, only one item overlapped between the two short forms. Scores from the two short forms were very similar ($r = 0.92$; 95% of differences in individual scores were within 6 points on a T-score metric, or 0.60 SD, of one another). Neither short form demonstrated a superior incremental validity over the other. A follow-up question invited respondents to suggest items that were not covered in the list of PROMIS items. Though participants made a number of suggestions, almost all were related to some other domain or dimension (e.g., coping skills). The lone exception was an item that asked about the relationship of sleep to fatigue.

Another approach is to identify *a priori* what are the domain-related concerns of a sample of individuals who have a specific disease or condition using individual persons' qualitative interviews, either focus group or one-on-one interviews. The results could be tabulated using accepted qualitative methods, and then the identified themes could be cross-referenced against the content coverage of a given item bank. Researchers at the University of Washington are currently evaluating the results of such a study with several PROMIS health domains, in which individuals with HIV were interviewed. The qualitative protocol began with the question, "What would your doctor need to know regarding your experience of (target domain) to take good care of you?"

Summary

The vision of PROMIS and the PROMIS methodology support the development and use of universally-relevant measures sharing a common metric centered on the most recently available data from the U.S. general population. This "universal measures on a common metric" approach to measurement allows researchers to speak a mutually-understood language when studying and reporting on PROs, when conducting comparative effectiveness studies, and in clinical care itself. Equally important, it allows for use of a common measurement "yardstick" to more appropriately and more transparently evaluate findings across a wide variety of observational and interventional research. Investigation of measurement invariance is included in the PROMIS methodology as an essential part of the process of creating high quality measures. When there is empirical evidence of measurement variance, the PROMIS methodology offers means

and options for addressing it while retaining the hallmark advantage of PROMIS measurement: the wide availability of universally-relevant measures sharing a common metric.

References

1. Tinetti ME, McAvay GJ, Chang SS, et al. Contribution of multiple chronic conditions to universal health outcomes. *J Am Geriatr Soc.* Sep 2011;59(9):1686-1691.
2. Working Group on Health Outcomes for Older Persons with Multiple Chronic Conditions. Universal health outcome measures for older persons with multiple chronic conditions. *J Am Geriatr Soc.* Dec 2012;60(12):2333-2341.
3. Tinetti ME, McAvay G, Chang SS, et al. Effect of chronic disease-related symptoms and impairments on universal health outcomes in older adults. *J Am Geriatr Soc.* Sep 2011;59(9):1618-1627.
4. Liu H, Cella D, Gershon R, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clin Epidemiol.* 2010;63(11):1169-1178.
5. Siegel K, Lekas HM, Schrimshaw EW, Brown-Bradley CJ. Strategies adopted by late middle-age and older adults with HIV/AIDS to explain their physical symptoms. *Psychol Health.* May 2011;26 (Suppl 1):41-62.
6. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychol Sci Q.* 2009;51(2):148-180.
7. Paz SH, Spritzer KL, Morales LS, Hays RD. Evaluation of the Patient-Reported Outcomes Information System (PROMIS(R)) Spanish-language physical functioning items. *Qual Life Res.* Nov 3 2012.
8. Butcher JN, Dahlstrom WG, Graham JR, Tellegen A, Kaemmer B. *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring* Minneapolis, MN University of Minnesota Press; 1989.
9. Hays RD, Hahn H, Marshall G. Use of the SF-36 and other health-related quality of life measures to assess persons with disabilities. *Arch Phys Med Rehabil.* Dec 2002;83(12 Suppl 2):S4-9.
10. Ward BW, Schiller JS. Prevalence of Multiple Chronic Conditions Among US Adults: Estimates From the National Health Interview Survey, 2010. http://dx.doi.org/10.5888/pcd10.120203. *Prev Chronic Dis.* 2013;10:120203.
11. Cook KF, Bamer AM, Roddey TS, Kraft GH, Kim J, Amtmann D. A PROMIS fatigue short form for use by individuals who have multiple sclerosis. *Qual Life Res.* Aug 2012;21(6):1021-1030.