Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 423-433

Epilogue to the two-part series: Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short forms

Jeanne A. Teresi[1,2,3] *& Bryce B. Reeve*[4,5]

Abstract

The articles in this two-part series of Psychological Test and Assessment Modeling describe the psychometric performance and measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form measures in ethnically, socio-demographically diverse groups of cancer patients. Measures in eight health-related quality of life domains were evaluated: fatigue, depression, anxiety, cognition, pain, sleep, and physical and social function. State-of-the-art latent variable methods, most based on item response theory, and described in two methods overview articles in this series were used to examine differential item functioning (DIF).

Findings were generally supportive of the performance of the PROMIS measures. Although use of powerful methods and large samples resulted in the identification of many items with DIF, practically none were identified with high magnitude. The aggregate level impact of DIF was small, and minimal individual impact was detected. Some methodological challenges were encountered involving positively and negatively worded items, but most were resolved through modest item removal. Sensitivity analyses showed minimal impact of model assumption violation on the results presented.

A cautionary note is the observance of a few instances of individual-level impact of DIF in the analyses of depression, anxiety, and pain, and one instance of aggregate level impact just below

[1] *Correspondence concerning this article should be addressed to:* Jeanne A. Teresi, Ed.D., Ph.D., Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York 10032-3702, USA; email: Teresimeas@aol.com

[2] Research Division, Hebrew Home at Riverdale; RiverSpring Health, 5901 Palisade Avenue, Riverdale, NY 10471

[3] Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College, 1300 York Avenue, New York, NY 10168

[4] Health Policy and Management, University of North Carolina at Chapel Hill

[5] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill

threshold in the analyses of physical function. Although this sample of over 5,000 individuals was diverse, ethnically, a limitation was the lack of ability to examine language groups other than Spanish and English and specific ethnic subgroups within Hispanic, Asian/Pacific Islander, and Black subsamples.

Extensive qualitative and quantitative analyses were performed in the development of PROMIS item banks. These sets of analyses, performed by several teams of psychometricians, statisticians, and qualitative experts, were the first to examine measurement equivalence of PROMIS short forms among ethnically diverse groups, and were also the first examination of PROMIS short forms among adults with cancer. Results presented in these articles provide strong evidence supporting the measurement equivalence of PROMIS short forms.

Key words: Patient Reported Outcomes Measurement Information System®, PROMIS®, short forms, measurement equivalence, differential item functioning, item response theory, reliability, validity, ethnic diversity, cancer

Summarized in this and an earlier issue of Psychological Test and Assessment Modeling (Reeve & Teresi, 2016) are the first studies of the measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short forms in a large, ethnically diverse sample of over 5,000 people (Jensen, Moinpour, et al., 2016). This was also the first study to examine PROMIS short forms in a population-based sample of adults with cancer, including those with co-morbidity. A goal of PROMIS is to provide end-users access to psychometrically-robust and standardized measures of health-related quality of life that may be used comparatively in clinical research, population surveillance, and healthcare delivery settings.

Differential item functioning (DIF) was examined in PROMIS short forms measuring eight domains: fatigue, depression, anxiety, cognitive concerns, pain interference, sleep, and physical and social function. This series also included overviews describing the methods used as well as methodological challenges (Kleinman & Teresi, 2016; Teresi & Jones, 2016). Generally, the short form measures performed well, all with high reliability. Because of the large sample sizes and power of the methods, significant DIF was observed for many items; however, the magnitude and impact of the DIF was negligible. Nonetheless, some items were singled out for further study.

The main models used in the analyses were based on item response theory (IRT). Challenges included local dependencies (LD) manifest for some models, often resulting in inflated estimates of the discrimination (slope parameter), as well as problems in obtaining sufficient numbers of DIF-free anchor items to set the metric of the scale.

In the area of sleep, four items were removed because of content overlap or because reverse-scored items produced methods effects. The use of both positively and negatively worded items resulted in residual correlations that required modeling (Jensen, King-Kallimanis, et al., 2016). The physical function item set also posed a challenge, in that many items evidenced high residual correlations with other items (Jones, Tommet, Ramirez, Jensen, & Teresi, 2016). There were a mixture of eight negatively worded limitations and eight positively worded ability items, and embedded in the latter item set were four “less difficult” items measuring basic activities of daily living, e.g., dressing, toileting. Positively correlated residuals among these items suggested the presence of a secondary factor. However, modeling the residuals did not affect model fit appreciably and the overwhelming evidence supported essential unidimensionality.

A reverse-scored item (the only positively worded item) was also removed from the fatigue short form item set analyses because of poor discrimination: having enough energy to exercise strenuously (Reeve et al., 2016). Similarly, an item was removed from the pain short form due to local dependency associated with item content overlap (Teresi, Ocepek-Welikson, Cook, et al., 2016). For example, the highest local dependency for the total sample was for the item pair: pain interferes with the things you usually do for fun and pain interferes with enjoyment of social activities. As a remedy, the analysis was repeated excluding the latter item from the set, resulting in overall improved item fit statistics, and reduced LDs and discrimination parameter estimates. An additional problem was that many items showed DIF in the initial analysis, resulting in the elimination of most of the items from the anchor sets, reducing the anchor sets to as low as two anchor items.

Generally, DIF analyses were robust to violations of model assumptions and anchor item selection. Sensitivity analyses increasing the number of anchor items and omitting items with high local dependency statistics resulted in similar DIF results. Assumption violations are likely to impact parameter estimates in a fashion leading to false DIF detection; additionally, the use of powerful methods and large sample sizes (as in this study) will also result in greater detection of significant DIF, even if trivial. In general, across these analyses, despite the large amount of DIF observed, the magnitude and impact of DIF were negligible. Finally, for one domain (pain) other models in addition to the graded response IRT model, e.g., nominal response models, were also attempted; however, the model fit for the other models was generally poor, and not as good as the original graded response model.

Reviewed below briefly are the findings related to DIF.

Fatigue

Because powerful methods for DIF detection based on IRT were used, small deviations in the item parameters resulted in all 13 fatigue items being identified with DIF in at least one of the several DIF comparisons across groups (i.e., gender, age, education, race/ethnicity, language translation; Reeve et al., 2016). Findings were considered in concert with the hypotheses generated. Although there were numerous hypotheses with respect to age in the direction of more fatigue in older patients, little consistent DIF was observed for age. The exception was for the item related to energy. One item was singled out for further review: able to think clearly. Of note was that the items that were on a frequency scale (*never* to *always*) evidenced somewhat more instances of DIF and lower discrimination parameters than those with amount (*not at all* to *very much*) response options. Follow-up magnitude and impact analyses showed little impact on the aggregated fatigue scale score. Although, more items were observed to evidence DIF in comparisons involving Asians/Pacific Islanders with other groups; differences in scores between the comparison groups with and without DIF adjustment were negligible.

Depression

As observed with the fatigue items, many short form depression items tested positive for DIF, particularly among Asians/Pacific Islanders as contrasted with the White reference group (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016a). No items evidenced high magnitude DIF for gender. One item showed slightly higher magnitude for age: nothing to look forward to; this item was also hypothesized to show age DIF, and might be targeted for further study. Conditional on depression, this item was more likely to be endorsed in the depressed direction by both older groups in contrast with the cohort aged 21 - 49. Only one item showed DIF of higher magnitude (just above threshold) for Whites vs. non-Hispanic Asians/Pacific Islanders in the direction of higher likelihood of endorsement for Asians/Pacific Islanders: felt like a failure. Because this item was also hypothesized to show DIF for minority groups, it might be singled out for further study. The magnitude of DIF was small, and the impact negligible, as shown by overlapping item characteristic curves; however, individual impact was observed for a small proportion (<1 %) of individuals.

Conditional on depression, the item, felt worthless was more likely to be endorsed in the depressed direction by the patients with less than high school education vs. the patients with a graduate degree; however, the magnitude of DIF was below threshold. Because this item was also hypothesized to show DIF in the direction of more feelings of worthlessness by groups with lower education, it might be targeted for further review. In summary, there was a correspondence of the DIF hypotheses to the findings of DIF in a number of instances; however, the magnitude and impact of DIF were negligible and reliability estimates were high across all studied groups, regardless of estimation method.

Anxiety

Less significant DIF was observed for anxiety items than for depression items (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016b). Contrary to the hypotheses, the findings were of very little DIF by gender group. Consistent with the hypotheses, conditional on anxiety, older respondents (aged 65 - 84) were less likely to express feelings of fearfulness and anxiety. As hypothesized, conditional on anxiety, Hispanics evidenced a significantly higher probability of responding in the anxious direction to the item, felt worried. Asians/Pacific Islanders (as contrasted with non-Hispanic Whites) evidenced a higher conditional probability of responding in the anxious direction to the items felt fearful, that situations made them worry, and that worries overwhelmed them, but were significantly less likely to report needing help for anxiety. Conditional on anxiety, the youngest age group in contrast to the oldest was more likely to express feelings of worry, and this item evidenced slightly higher magnitude of DIF; however, this item was not hypothesized to evidence DIF. Only the item, “Many situations made me worry”, showed DIF of higher magnitude (just above threshold) for non-Hispanic Asians/Pacific Islanders vs. Whites. As with depression, very little DIF of high magnitude was evidenced in the PROMIS short form items, and none of high aggregate impact. However, individual level impact for race/ethnic or education group comparisons was observed for a small proportion (< 3 %) of the sample.

In summary, the item, worried might be singled out for more study, given that there was a confirmatory hypothesis regarding this item for Hispanics, who were hypothesized to express feeling worried, for reasons unrelated to anxiety, and DIF was observed for this item for this comparison as well as for age.

Cognition

DIF was examined in the PROMIS applied cognition (concerns) short form, containing eight items relating to subjective cognitive concerns. DIF was hypothesized for the item, brain not working as well as usual in the direction of higher self-reported impairment for Latinos and respondents with higher education. After correction for multiple comparisons, this item showed significant DIF in both primary and sensitivity analyses for both the race/ethnicity and education comparisons (Fieo et al., 2016), thus confirming the hypothesis for education. However in the race/ethnicity comparisons, Hispanic and Black respondents were less likely to endorse the item in the cognitive difficulties direction compared to non-Hispanic White respondents. In the language analysis, no DIF was found even though it was hypothesized that conditional on cognition, non-English and Spanish speakers would report higher impairment in having to work hard to pay attention.

Conditional on cognitive complaints, older respondents had a higher likelihood than younger respondents of endorsing two items in the cognitive complaints direction: “I have had to work really hard to pay attention or I would make a mistake” and “I have had trouble shifting back and forth between different activities that require thinking”. Overall the magnitude of DIF was below threshold, and the aggregate and individual impact was minimal. High reliability was observed across comparison groups.

Pain interference

DIF analyses were performed on the pain interference short form 10 item set (Teresi, Ocepek-Welikson, Cook, et al., 2016). No items were hypothesized to show DIF for race and ethnicity; however, five items showed DIF after adjustment for multiple comparisons in both primary and sensitivity analyses: ability to concentrate, enjoyment of recreational activities, tasks away from home, participation in social activities, and socializing with others. The items, ability to concentrate and enjoyment of recreational activities were also identified consistently with DIF for education and age. No item showed DIF above the magnitude threshold and the impact of DIF on the overall measure was minimal. Spanish speakers were hypothesized to experience less pain interference on one item, enjoyment of life. The DIF findings confirmed the hypothesis; however, the magnitude was small.

The reliability estimates were high across comparison groups. Despite the many DIF findings, the magnitude and aggregate impact of DIF was minimal. However, some individual-level impact was observed for about 4 % of the sample.

Sleep

Because of various methods effects, the 10 items did not fit a unidimensional model (Jensen, King-Kallimanis, et al., 2016). Standardized residuals, correlations, modification indices, and expected parameter changes were examined to arrive at a version that fit the data well. This six-item model was used for further analyses of race/ethnicity using multiple groups confirmatory factor analysis. First fit was an unrestricted model, followed by placement of equality constraints on both the factor loadings and intercepts. Only one item with DIF was identified. Non-Hispanic White participants were less likely to agree with the statement “I had difficulty falling asleep” compared to the other ethnic racial groups (Blacks, Hispanics, and Asians/Pacific Islanders). Thus, partial strong factorial invariance was established for the six item sleep disturbance scale for race/ethnicity. Although hypothesized for many items, no DIF by age or sex was identified, suggesting that the sleep disturbance scale was invariant with respect to age and sex.

Evidence also supported the reliability and convergent, discriminant, and known groups construct validity of the new six item short form sleep disturbance scale.

Physical functioning

Most of the five items which evidenced DIF for Black or African American respondents as contrasted with the White non-Hispanic reference group were also hypothesized to show DIF (Jones et al., 2016). For example, doing two hours of physical labor, vigorous activities, and walking more than a mile were posited to show DIF, and the latter two items were also found in other studies to evidence DIF for Black in contrast to White groups. However, the scale-level impact was negligible as evidenced by the estimated differences in physical function estimates for Blacks and Whites, adjusted for DIF.

Many items evidenced DIF in comparisons of Hispanics to Whites: vigorous activities, physical activities, physical labor, moderate work, chores, carry groceries, kneeling/bending/stooping, wash/dry body, walk a mile, and walk 15 minutes. The difference in physical function between Whites and Hispanics before and after DIF adjustment was substantial (a 38 % under-estimation of Hispanic group deficits in physical functioning); however, the scale level impact (absolute mean difference of 0.18 when controlling for DIF) approached, but did not meet an arbitrary effect size threshold established a priori. Similar to other domains (fatigue, depression, and anxiety), many items (15) evidenced DIF for the comparisons of Asians/Pacific Islanders to Whites; however, the scale level impact was small.

Similar small effects were found for gender and age, with very small differences in the pre- and post-DIF adjustment mean physical functioning levels. Many items were also shown to evidence DIF for age, sex, and education; however, the scale level impact was very small, and ignorable.

Although a great number of measurement model parameters were different across the various focal and reference groups, the overall scale-level impact of this DIF was trivial. However, the Hispanic subgroup comparison revealed an impact estimate just below an arbitrary threshold for small impact. Within the limitations of local dependency violations, it was concluded by the authors (Jones et al., 2016) that short form items derived from the PROMIS physical functioning item bank performed well.

Social function

Despite many hypothesized DIF relationships across groups, no significant DIF was identified for the PROMIS social function (ability to participate in social roles and activities) short form using primary cutoffs for DIF (Hahn et al., 2016). Sensitivity analyses identified only one item with DIF of trivial impact. Reliability was high and correlations with other PROMIS scales supported the criterion-related validity of the 10 item scale. Evidence was generally supportive of the concurrent criterion-related validity of the social function short form in terms of correlations with other PROMIS short forms. Thus the performance of the PROMIS Ability to Participate in Social Roles and Activities short form was viewed by the authors as evidence in support of its use among ethnically diverse groups of individuals with cancer and other chronic conditions.

Discussion

This is one of the first studies of PROMIS short forms among a large sample of ethnically diverse groups. The findings regarding item and scale performance were generally favorable across domains; however, some methodological challenges were encountered. These included the presence of methods effects, model assumption violations, and inadequate numbers of anchor items. These problems were handled by (a) removal of items, (b) modeling, and (c) sensitivity analyses to examine the effects of violations on DIF detection. Some PROMIS items, worded positively and mixed with negatively worded items were problematic, resulting in inconsistent responses. Local dependencies were observed for most of the short forms, often resulting in inflated discrimination parameters. In general, DIF findings were robust to most of the violations. It is recommended that physical function short form items be examined further because the local dependencies observed are suggestive of an additional unmodeled specific factor, possibly the result of the mixture of basic activities of daily living with more difficult functional items. Local dependency usually results in false DIF detection; in the case of physical function, although many items were indeed observed with DIF; the impact of DIF on the scale score was negligible.

Findings were of many items with significant DIF, given the powerful methods and large sample sizes. More DIF was observed for Asians/Pacific Islanders; albeit of low magnitude. Thus more research with this group is needed. In general, the magnitude and aggregate-level impact was minimal. However, some DIF of slightly higher aggregate impact was observed for the physical function short form for Hispanics, and in the analyses of depression, anxiety, and pain, individual-level impact was observed for some individuals (1 to 4 % across analyses of race/ethnicity and education), the most among those with lower education in the analyses of pain.

Summary: The findings are supportive of the positive performance of the PROMIS short form items. The measures performed well across most comparison groups and evidenced high reliability. Although many items were identified with DIF, given the powerful methods and large sample sizes, very few were of magnitude above threshold that would be concerning. Thus, analyses across comparison groups and domains were generally supportive of measurement equivalence. There were few instances of gender DIF. With some exceptions, the items performed equivalently across the life span from age 21 to 84 years, and across different education groups from less than high school to graduate level. Most items were invariant with respect to race/ethnicity for the global groups of Hispanics, Blacks, non-Hispanic Whites, and Asians/Pacific Islanders, although slightly more DIF was observed for some domains among the latter group. The items also performed well among Spanish and English speaking Hispanics. The DIF magnitude and aggregate impact was ignorable; although consistent with earlier findings, some impact just below threshold was observed for the physical function short form item set for Hispanics. A cautionary note is the observance of some individual-level impact of DIF, particularly for the pain interference domain among those with lower education. It is also noted that the findings of ignorable DIF are based on treating the short forms as continuous. DIF could have more of an impact if scales were used as screens with cutoffs for inclusion or exclusion. Additionally, the effect of DIF from an item could be more of an issue if a computerized adaptive test (CAT) were being used and an item with DIF was selected as one of the four or five items administered to the respondent. PROMIS permits different group calibrations based on DIF; however, more research is needed to make definitive recommendations regarding such calibrations. A limitation of these sets of analyses is the inability to study specific ethnic subgroups; this is an area that warrants further research.

Measurement equivalence is a basic requirement for valid assessment. Given the increasingly wide use of PROMIS measures in research and clinical practice, the analyses described in this set of papers is a major advance in providing evidence regarding their performance. Despite some methodological challenges, the PROMIS short forms examined in this set of papers generally performed well among this ethnically diverse sample, and the evidence supports their continued use.

Acknowledgements

Partial funding for the preparation of one or more of these manuscripts was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: A. Potosky, C. Moinpour), by the National Institute on Aging (NIA) Claude D. Pepper Center, 1P30AG028741-01A2 (PI: A. Siu), the NIA Edward R. Roybal Center, P30AG022845 (PI: M.C. Reid, K. Pillemer, E. Wethington), NIA, K01AG045342 (PI: G. Kim), NIA Resource Centers for Minority Aging Research, 5P30AG031054 (PI: K. Burgio, M. Fouad) and from the National Cancer Institute, P30-CA051008-S1.

The authors thank Stephanie Silver, MPH for editorial assistance in the preparation of this manuscript.

The authors also thank over 30 internationally recognized experts in measurement statistics and/or in the substantive domains for their reviews of these manuscripts.

PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926) and thirteen Primary Research Sites which may include more than one institution (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD and Karon Cook, PhD, 1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR057956; Stanford University, PI: James F. Fries, MD, 2U01AR052158; Boston University, PIs: Stephen M. Haley, PhD and David Scott Tulsky, PhD (University of Michigan, Ann Arbor), 1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, PIs: Arnold L. Potosky, PhD and Carol. M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle),U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186).

NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce B. Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, Ellen Werner, PhD, and James Witter, MD, PhD.

The authors also thank members of the MY-Health Clinical Advisory Committee (Drs. Patricia Ganz, Julie Gralow, Maurie Markman, Jimmy Hwang, Anthony Back, and Bonnie Teschendorf), SEER study site collaborators (Rosemary Cress, Theresa H.M. Keegan, Xiao-Cheng Wu, Antoinette Stroup, Lisa Paddock, Laura Allen, Lauren S. Maniscalco, Lisa Moy, Natalia Herman, and Wendy Ringer), and the Georgetown MY-Health Research team (Caroline Moore, Charlene Kuo MPH, Tania Lobo MS, Marin Rieger MS, Deena Loeffler, MA, Aaron Roberts).

References

Fieo, R., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Cella, D., & Teresi, J. A. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Applied Cognition - General Concerns, short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 255-307.

Hahn, E. A., Kallen, M. A., Jensen, R. E., Potosky, A. L., Moinpour, C. M., Ramirez, M., ... & Teresi, J. A. (2016). Measuring social function in diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Ability to Participate in Social Roles and Activities short form. *Psychological Test and Assessment Modeling*, *58*(2), 403-421.

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour C. M., Potosky, A. L., … & Teresi, J. A. (2016). Measurement properties of PROMIS® sleep disturbance short forms in a large, ethnically diverse cancer cohort. *Psychological Test and Assessment Modeling*, *58*(2), 353-370.

Jensen, R. E., Moinpour, C. M., Keegan, T. H. M., Cress, R. D., Wu, X.- C., Paddock, L. A., … Potosky, A. L. (2016). The Measuring Your Health Study: Leveraging community-based cancer registry recruitment to establish a large, diverse cohort of cancer survivors for analyses of measurement equivalence and validity of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form items. *Psychological Test and Assessment Modeling, 58*(1), 99-117.

Jones, R. N., Tommet, D., Ramirez, M., Jensen, R., & Teresi, J. A. (2016). Differential item functioning in Patient Reported Outcomes Measurement Information System® (PROMIS®) physical functioning short forms: Analyses across ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 371-402.

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58*(1), 79-98.

Reeve, B. B., & Teresi, J. A. (2016). Overview to the two-part series: measurement equivalence of the Patient-Reported Outcomes Measurement Information System® (PROMIS®) short-forms. *Psychological Test and Assessment Modeling, 58*(1), 31-35.

Reeve, B. B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., … & Chen, W- H. (2016). Psychometric evaluation of the PROMIS® fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling, 58*(1), 119-139.

Teresi, J. A. & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in Patient Reported Outcomes Measures: Methods overview to the two-part series, "Measurement Equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Short Form Measures." *Psychological Test and Assessment Modeling*, *58*(1), 37-78.

Teresi, J. A., Ocepek-Welikson, K., Cook, K. F., Kleinman, M., Ramirez, M., Reid, M. C., & Siu, A. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) pain interference short form items: Applications to ethnically diverse cancer and palliative care populations. *Psychological Test and Assessment Modeling*, *58*(2), 309-352.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System® (PROMIS®) depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58*(1), 141-181.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, R., & Kim, G. (2016b). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58*(1), 183-219.