Published in final edited form as:

J Shoulder Elbow Surg. 2018 March ; 27(3): 515–522. doi:10.1016/j.jse.2017.10.015.

Psychometrics of the PROMIS PF CAT and DASH in orthopedic elbow patient population

Man Hung, PhD¹, Maren W. Voss, MS¹, Jerry Bounsanga, BS¹, Yushan Gu, BS¹, Erin K. Granger, MPH¹, and Robert Z. Tashjian, MD¹

¹University of Utah Department of Orthopaedic Surgery Operations, Salt Lake City, UT, USA

Abstract

Introduction—The purpose of this study was to evaluate the psychometric properties of the Patient-Reported Outcomes Measurement Information System's (PROMIS) Physical Function (PF) instrument administered through computerized adaptive testing (CAT) compared to the traditional full-length Disabilities of the Arm, Shoulder and Hand (DASH).

Methods—The PROMIS PF CAT and the DASH were administered to 1,759 patients seeking care for elbow conditions. This study used RASCH Partial Credit Modeling to analyze the instruments with item fit, internal reliability, response category thresholds, dimensionality, local independence, gender differential item functioning, and floor and ceiling effects.

Results—The PROMIS PF CAT and DASH had satisfactory item fit for all but one item on both measures. Internal reliabilities were high for both measures. Two items on the DASH and four items on the PF CAT show non-ordered category thresholds. Unidimensionality was adequate and local independence was supported for both instruments. Four items on the PF CAT showed gender bias while 12 items on the DASH were gender biased. Both measures had adequate instrument targeting and satisfactory floor and ceiling effects.

Conclusion—Both the PROMIS PF CAT and the DASH showed sufficient unidimensionality, good item fit, and good local independence with the exception of high levels of gender item bias, particularly for the DASH. Further scale evaluation should address item bias and item response categories for these instruments. Overall the PROMIS PF CAT is an effective outcome instrument to measure function in patients with elbow disorders requiring significantly fewer questions than the DASH.

Level of evidence—Basic Science Study, Validation of Outcome Instruments

Keywords

PROMIS; DASH; physical function; psychometric; orthopedics; elbow

Correspondence: Man Hung, PhD, Department of Orthopaedics, University of Utah School of Medicine, 590 Wakara Way, Salt Lake City, UT 84108, USA, Man.Hung@hsc.utah.edu.

Disclaimer: No financial biases exist for any author. All authors state "none". Institutional Review Board approval was obtained for the current study at the University of Utah, under IRB #00053404.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Value-based propositions in health care have been the focus of health care reform²⁶ and are evidenced by increases in bundling payment models¹¹ and patient safety based reimbursement penalties.²⁹ This increased emphasis on value over volume requires a metric or measuring system to accurately assess and compare treatment effects. Precise patient-reported outcome (PRO) measurement is critical to the process of improving care while controlling costs.³² Orthopaedics as a profession needs to take steps to understand the psychometric properties of outcome measures in order to collect useful data that can show measurable improvements to justify treatment costs.²

The widely used Disabilities of the Arm, Shoulder, and Hand (DASH) has been shown to be a reliable, responsive, and valid measure of upper extremity disability.^{12,17,28} However, depending on the population being assessed, it is prone to floor and ceiling effects.¹⁶ The DASH is a 30-item questionnaire, which can take as much as 5 minutes for patients to complete, adding to testing burden particularly if it is administered with other instruments.³⁰ Newly developed measures utilizing computerized adaptive testing (CAT) can minimize test burden and improve the patient experience.³³

The National Institutes of Health funded the Patient-Reported Outcomes Measurement Information System (PROMIS) to develop and improve upon existing PRO instruments.⁷ Included is the PROMIS Physical Function (PF) CAT v. 1.2, a 124-item instrument which assesses upper extremity, lower extremity, central function as well as functioning in daily living activities.¹⁴ The CAT administration uses computer algorithms to select targeted questions that assess functioning ability without redundancy. This is accomplished as individual items are selected from the 124-item bank that incorporate or are altered based on a patient's prior answers. The PROMIS PF CAT has shown high correlation with its full-length test version, while significantly reducing testing time and test burden.^{6,20,21,23,36} Reduced questionnaire length has been associated with higher completion rates,²⁷ an important factor in obtaining high quality data for value-driven healthcare.

As new instruments are developed, it is important to thoroughly evaluate their psychometric properties in defined patient populations, as well as to assess the relative performance of new measures compared to commonly used legacy measures. The purpose of this study is to evaluate the psychometric characteristics of the PROMIS PF CAT and the DASH in an upper-extremity orthopedic population with elbow conditions.

MATERIAL AND METHODS

Patients visiting a university orthopedic clinic for elbow conditions between February 2014 and March 2017 took the PF CAT and DASH on hand-held tablet computers (iPad; Apple, Inc., Cupertino, CA, USA) as part of their routine clinic care and prior to seeing the medical professionals. Data were collected using the mEVAL system via the institution's electronic medical record system through a secure wireless network and stored in the institution's enterprise data warehouse. Standard CAT algorithms from the PROMIS Assessment Center were applied. All patients aged 18 years and older were included, consecutively enrolled based on the inclusion criteria of upper-extremity elbow conditions. Non-English speakers

were included if patients felt comfortable with translators going through the questionnaire side by side. IRB approval was obtained prior to the start of the study.

The DASH is a 30-item questionnaire. Its scores range from 0 to 100, with higher scores representing higher levels of functioning. The PROMIS PF CAT draws from a 124-item test bank, with an average test length of 6–7 items in an orthopedic population.¹⁹ It has calibrated a T-score scale with mean of 50 and standard deviation of 10 in the general population.³⁴ T-scores can help in standardizing and better representing the general population in a consistent way.¹ Higher scores on the PROMIS PF reflect higher levels of function. The instruments were administered at the first clinic visit or within the seven days prior to the visit of a new patient with an upper extremity condition. Patient were emailed prior to their visit to complete the health assessment at home, and if not completed prior, complete the assessment on tablet at check-in.

Analyses

Psychometric evaluation of the PROMIS PF CAT and the DASH was completed using the Rasch Partial Credit Model (PCM). This model is based on logistic latent trait item-response theory (IRT).³¹ IRT poses many advantages for test development because it generates item rich information while linking to the underlying latent trait.⁸ Rasch modeling can determine the fit of individual items to the latent trait, and the PCM can model fit with polytomous items (items with more than 2 possible scores), as it does not require the response thresholds to be uniform across items. Rasch PCM has been used frequently in modern instrument development and evaluation.^{15,22} The analyses in this study included assessments of: 1) Item Fit, 2) Internal Reliability, 3) Response Category Threshold, 4) Dimensionality, 5) Local Independence, and 6) Differential Item Functioning.

Item Fit—The Rasch PCM is based on a form of Guttman scaling, where items and persons can be placed on the same linear continuum. In this analysis we first need to compare the response patterns in the observed data to the expected patterns determined by the model to see if Rasch PCM is suitable for analysis of the data. The items are considered to fit the model if the observed responses are not significantly different from the expected responses. The outfit Mean Square (MNSQ) statistic is used here as an indicator of model fit. Perfect fit is represented by an outfit MNSQ of 1. An outfit MNSQ less than 1.4 is a good fit and less than 2 is an acceptable fit.⁴

Internal Reliability—Internal reliability is evaluated by the Person Separation Index (PSI). The PSI assesses how well an item distinguishes between groups of patients. Interpretation of the PSI is similar to Cronbach's Alpha. The advantage of PSI over Cronbach's Alpha is that the PSI has no upper limit to its value. A PSI of 0.8 or above is a satisfactory indicator of internal reliability.

Response Category Threshold—Item response categories should be adequate to capture the differences between responses. Otherwise, a different response format should be considered, such as collapsing the response categories. Checking whether the response

thresholds (i.e., the transition points between response categories such as “very good” and “good”) are logically ordered is necessary in assessing category thresholds.

Dimensionality—Dimensionality refers to whether a measure is assessing just one or multiple underlying ideas or factors. If only one factor is influencing scale scores, the instrument is considered unidimensional. If there is more than one factor, the instrument is considered multidimensional. By controlling for the first factor of the Principal Component Analysis of the residuals, if more than 5% unexplained variance remains, this indicates extra noise in the measurement model and an indicator of multidimensionality.³⁷

Local Independence—Local independence suggests that the response to one item is not dependent on the response to another. It can be assessed through the item residual correlations which compare observed and expected partial correlations. A residual correlation greater than 0.7 is possible evidence for trait violation of local independence.

Differential Item Functioning (DIF)—The presence of DIF is an indicator of item bias, meaning that the item is biased toward different groups, and that individuals will perform differently on the item based on their group characteristics. This study examined gender DIF to see if there is evidence of item score bias toward female or male. A two-sided t-test was used to identify DIF by gender group, with *p*-values less than 0.05 indicating potential bias.

Targeting—Targeting is an assessment of instrument coverage. It can be analyzed to determine whether items cover an acceptable range of patients’ function or abilities. A person-item map was constructed to directly compare item and person functioning scores. On a person-item map, item and person distributions in identical and standardized scales are separated from right to left. From top to bottom, higher function levels shift towards the top of the vertical plot whereas lower function levels would shift towards the bottom. Sufficient targeting can be indicated by items and persons having similar distributions. Ceiling and floor effects are calculated as empirical indicators of instrument targeting. The proportion of patients’ measures that exceed the maximum measure of the instrument is the ceiling effect. The proportion of patients’ measures that are below the minimum indicates a floor effect. Ceiling and floor effects of less than 10% are considered satisfactory.

RESULTS

There were 1,759 patients seeking care for elbow conditions who took either the PF CAT or the DASH and were included in this study. Of these patients, 985 (56.0%) were male and the average age was 55.75 (SD=16.16; Range 18–93). The sample was 88.8% (n=1,562) Caucasian and 92.8% (n=1,633) non-Hispanic. Race was reported as “other” by 4.8% (n=85) of the sample, with 1.5% (n=26) as Black or African American, 1.3% (n=23) as American Indian/Alaska Native, 0.9% (n=16) as Asian, and 0.4% (n=7) as Hawaiian/Pacific Islander. Race data were missing for 2.3% of the sample (n=40) and ethnicity data were missing for 3.0% of the sample (n=52). Table 1 contains the details of the patient demographic characteristics. On average, there were 6 PF CAT items administered (SD = 2; median = 5.00; mode = 4.00) and 29 DASH items administered (SD = 2.17; median = 30; mode = 30). The importance of evaluating Item Fit, Internal Reliability, Response Category

Threshold, Dimensionality, Local Independence, and Differential Item Functioning, along with potential interpretations of values, are presented above in the methods section. Results for these six parameters are presented below.

Item Fit—Twenty-six items from the PROMIS PF CAT were fit to the RASCH PCM with outfit MNSQ statistics ranging from 0.12 to 2.99 (see Table 2). Of those items, twenty-two items had good fit with outfit MNSQ values less than 1.4, three items (PFB12, PFA1, PFA55) had acceptable fit with outfit MNSQ values less than 2.0, and one item (PFC46) did not fit the RASCH PCM model well, with outfit MNSQ values over 2.0.

All thirty items from the DASH were fit to the RASCH PCM with outfit MNSQ ranging from 0.62 to 2.52 (see Table 2). Twenty-five items had good fit with outfit MNSQ values less than 1.4, and four items had acceptable fit (DASH2, DASH21, DASH22, DASH26) with outfit MNSQ values less than 2.0. One item (DASH30) did not fit the RASCH PCM model, with outfit MNSQ values over 2.0.

Internal Reliability—The PSI for the PF CAT was 1.13 and for the DASH the PSI was 1.42. Both are well above the 0.8 threshold for adequate reliability indicating the person measures are reproducible within the person item map and the instruments are able to distinguish high and low functioning persons.

Response Category Threshold—There were four items from the PFCAT that did not show ordered category thresholds (PFB12, PFC35, PFA55, PFC46). There were two items from the DASH that did not show ordered category thresholds (DASH2, DASH21).

Dimensionality—The unexplained variance in the 1st contrast for the PF CAT was 3.3% and for the DASH was 4.5%, providing evidence that both instruments were sufficiently unidimensional.

Local Independence—All residual correlations were less than 0.7 for both instruments correlation, indicating item local independence for the instruments.

Differential Item Functioning (DIF)—Four items on the PF CAT showed gender DIF (PFA1, PFA53, PFB5R1, PFC56)(see Table 3). The DASH had 12 items that showed significant differences by gender (DASH1, DASH5, DASH10-12, DASH19, DASH21-23, DASH27-28, DASH30).

Targeting—The person-item map separates item (on the right) and person (on the left) distributions. Higher function levels are represented at the top of the vertical graph with lower function towards the bottom. For the PF CAT there was a 2.84% floor effect observed and a 9.04% ceiling effect (see Figure 1). For the DASH there was a 3.07% floor effect observed and a 1.25% ceiling effect (Figure 2).

DISCUSSION

In assessing the psychometric properties of the PROMIS PF CAT and the DASH in an orthopaedic elbow population we found both instruments to be unidimensional, which means that they both are measuring just one dimension, likely the dimension of physical function (without comingling other factors such as pain or depression). There was only one item on each instrument that was considered a poor fit for the RASCH PCM. The items and categories reflect predicted patterns of responses and the model accounts for a sufficient level of variance between responses and categories. In general, the PROMIS PF CAT is a responsive and reliable tool that can be used as an alternate to the DASH to evaluate function in patients with elbow disorders with the benefit of requiring significantly fewer questions.

Both instruments showed high levels of internal reliability. The correlations between residuals were all less than 7.0, indicating that the individual items do not rely heavily on other items. Four of the 26 items in the PF CAT, or approximately 15% of the items, had category thresholds that were not ordered in this elbow population. When functioning properly, response categories will show an ascending or descending order, from lower to higher abilities or vice versa. The lack of ordered categories for 15% of the items means that responses of these items might not be well understood by the respondents. Instead of increasing or decreasing sequentially (12345 or 54321) the items showed non-monotonicity (for example, 12435). The 15% ratio of non-ordered category thresholds was also identified in initial studies of the PROMIS PF.¹³ The non-ordered categories suggests that for some items assessment might be more accurate if the response category options were collapsed from 5 choices to three or four. Only two items from the DASH (6.7%) did not show ordered category thresholds.

The DASH was well targeted to the orthopedic elbow patient population, suggesting the items have the correct levels of difficulty for elbow orthopedic patients. The DASH showed a 3.07% floor effect and a 1.25% ceiling effect, which is minimal. The items of the DASH were able to capture the full range of physical dysfunction and ability in this sample. The PROMIS PF CAT evidenced a slightly higher floor effect of 2.84% and a much higher ceiling effect of 9.04%, possessing less discrimination when assessing higher function in elbow orthopedics. This ceiling effect is below the 10% satisfactory ceiling cut-off, but is higher than previous evaluation of the PROMIS PF in other areas of orthopedics. Very minimal PROMIS PF ceiling effect was noted for an orthopedic trauma population,²⁴ a foot and ankle population,¹⁸ and an upper extremity population.³⁵ The ceiling effect of the PROMIS PF appears to be more pronounced in the current elbow population, possibly suggesting less efficient targeting of the PROMIS PF to this diagnostic group. There have been efforts to produce and validate additional items that will address floor and ceiling effects of the PROMIS PF.⁵ The recently developed PROMIS Upper Extremity (UE) may serve as a more targeted instrument in the hand, elbow, and shoulder population, though it has similarly evidenced ceiling effects as well.²⁵ The PROMIS UE shows a strong correlation with the PROMIS PF and a similar ceiling effect (7%) in an orthopedic sample, though patients with elbow conditions evidenced lower scores on the PROMIS UE than for patients with hand conditions.³ Until this work is completed on PROMIS refinements, the

DASH may be selected as a better fit for assessing elbow orthopedic outcomes in high functioning patients.

In terms of item bias, or DIF, our analysis of gender differences in response patterns suggests that the PF CAT had 15% of items with gender bias, while the DASH had 40% of items that were gender biased. This is a distinct problem in interpreting the DASH in that scores on the DASH will not be comparable across gender groups. In past research in a different patient population the DASH only showed gender bias for two items,⁹ again suggesting there may be different in the characteristics of the orthopedic elbow population leading to different targeting and bias in both the DASH and the PROMIS PF. These concerns about item bias and ceiling effects might prompt test developers to perform additional testing which might refine the test items to better suit an elbow population, however their psychometric performance is still considered adequate and does not disqualify either measure from use in an orthopedic elbow population. Further testing should be conducted to examine the nature of the DIF of these two instruments in the elbow orthopedics.

One limitation in this study is the lack of racial diversity in the patient population which could influence the generalizability of these results. This study was cross-sectional and did not evaluate the responsiveness to change of the PROMIS PF CAT or the DASH, which are important aspects of instrument validity. Future research should address these longitudinal research questions on whether these measures adequately detect change or treatment response in this population.

A number of factors influence clinical decision making as well as the decisions about which tests will be administered to provide clinical information. As legacy instruments have been in use and have guided clinical decisions in the past, it will be necessary to link scores from older measures to newly developed tests like the PROMIS instruments. Future analysis should build crosswalks that link scores from newer tests developed with IRT to legacy measures such as the DASH which have been used regularly in clinical practice. This will enable future research and clinical interpretation to take into account prior assessments and make direct comparisons to outcomes assessed with these newer instruments, easing the transition to implementing IRT developed instruments across orthopedic practice.

CONCLUSION

Overall the DASH showed a higher level of item bias, meaning different groups perform differently on the DASH, while the PROMIS PF CAT showed a higher ceiling effect, meaning that the test does not contain questions that fully assess higher functioning elbow patients. The equivalent performance based on psychometric properties of the DASH and PROMIS PF CAT in this orthopedic sample allows either test to be recommended. Yet there are many potential advantages of the IRT development and CAT administration of the PROMIS. Overall, PROMIS instruments administered in computerized formats provide improved reliability, flexibility in item administration, greater precision at the high and low extremes of function, and the ability to tailor the test to individual patients in ways that ease test burden.^{10,23} The PROMIS PF CAT should be given strong consideration to be used to

measure function in patients with elbow disorders based on its adequate psychometric performance combined with the ability to reduce testing time and ease patient burden.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of A Promis Item Bank to Measure Pain Interference. *Pain*. 2010; 150(1):173–182. DOI: 10.1016/j.pain.2010.04.025 [PubMed: 20554116]
2. Andrawis JP, Chenok KE, Bozic KJ. Health policy implications of outcomes measurement in orthopaedics. *Clinical orthopaedics and related research*. 2013; 471(11):3475–3481. DOI: 10.1007/s11999-013-3014-7 [PubMed: 23625577]
3. Beleckas CM, Padovano A, Guattery J, Chamberlain AM, Keener JD, Calfee RP. Performance of Patient-Reported Outcomes Measurement Information System (PROMIS) Upper Extremity (UE) Versus Physical Function (PF) Computer Adaptive Tests (CATs) in Upper Extremity Clinics. *J Hand Surg Am*. 2017; doi: 10.1016/j.jhsa.2017.06.012
4. Bond, TG., Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press; 2013.
5. Bruce B, Fries J, Lingala B, Hussain YN, Krishnan E. Development and assessment of floor and ceiling items for the PROMIS physical function item bank. *Arthritis research & therapy*. 2013; 15(5):R144.doi: 10.1186/ar4327 [PubMed: 24286166]
6. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health services research*. 2005; 40(5p2):1694–1711. DOI: 10.1111/j.1475-6773.2005.00446.x [PubMed: 16179003]
7. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of Item Candidates: The PROMIS Qualitative Item Review. *Medical care*. 2007; 45(5 Suppl 1):S12–S21. DOI: 10.1097/01.mlr.0000254567.79743.e2 [PubMed: 17443114]
8. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*. 2007; 16(1):5.doi: 10.1007/s11136-007-9198-0 [PubMed: 17375372]
9. Forget NJ, Jerosch-Herold C, Shepstone L, Higgins J. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. *BMC musculoskeletal disorders*. 2014; 15:361.doi: 10.1186/1471-2474-15-361 [PubMed: 25358527]
10. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical and experimental rheumatology*. 2005; 23(5 Suppl 39):S53–57.
11. Froemke CC, Wang L, DeHart ML, Williamson RK, Ko LM, Duwelius PJ. Standardizing Care and Improving Quality under a Bundled Payment Initiative for Total Joint Arthroplasty. *The Journal of arthroplasty*. 2015; 30(10):1676–1682. DOI: 10.1016/j.arth.2015.04.028 [PubMed: 26092251]
12. Gay RE, Amadio PC, Johnson JC. Comparative responsiveness of the disabilities of the arm, shoulder, and hand, the carpal tunnel questionnaire, and the SF-36 to clinical change after carpal tunnel release. *The Journal of Hand Surgery*. 2003; 28(2):250–254. DOI: 10.1053/jhsu.2003.50043 [PubMed: 12671856]
13. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. *Medical care*. 2007; 45(5):S32–S38. DOI: 10.1097/01.mlr.0000246649.43232.82 [PubMed: 17443117]
14. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. *Annals of the*

- rheumatic diseases. 2015; 74(1):104–107. DOI: 10.1136/annrheumdis-2013-204053 [PubMed: 24095937]
15. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *The Lancet Neurology*. 2007; 6(12):1094–1105. DOI: 10.1016/S1474-4422(07)70290-9 [PubMed: 18031706]
 16. Hsu JE, Nacke E, Park MJ, Sennett BJ, Huffman GR. The Disabilities of the Arm, Shoulder, and Hand questionnaire in intercollegiate athletes: validity limited by ceiling effect. *Journal of shoulder and elbow surgery/American Shoulder and Elbow Surgeons [et al]*. 2010; 19(3):349–354. DOI: 10.1016/j.jse.2009.11.006
 17. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: The DASH (Disabilities of the Arm, Shoulder, and Head). *American journal of industrial medicine*. 1996; 29(6):602–608. [PubMed: 8773720]
 18. Hung M, Baumhauer JF, Latt LD, Saltzman CL, SooHoo NF, Hunt KJ. Validation of PROMIS (R) Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clinical orthopaedics and related research*. 2013; 471(11):3466–3474. DOI: 10.1007/s11999-013-3097-1 [PubMed: 23749433]
 19. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res*. 2011; 29(6):947–953. DOI: 10.1002/jor.21308 [PubMed: 21437962]
 20. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *Journal of Orthopaedic Research*. 2011; 29(6):947–953. DOI: 10.1002/jor.21308 [PubMed: 21437962]
 21. Hung M, Hon SD, Cheng C, Franklin JD, Aoki SK, Anderson MB, et al. Psychometric Evaluation of the Lower Extremity Computerized Adaptive Test, the Modified Harris Hip Score, and the Hip Outcome Score. *Orthopaedic Journal of Sports Medicine*. 2014; 2(12):2325967114562191.doi: 10.1177/2325967114562191 [PubMed: 26535291]
 22. Hung M, Hon SD, Franklin JD, Kendall RW, Lawrence BD, Neese A, et al. Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. *Spine (Phila Pa 1976)*. 2014; 39(2):158–163. DOI: 10.1097/BRS.0000000000000097 [PubMed: 24173018]
 23. Hung M, Stuart AR, Higgins TF, Saltzman CL, Kubiak EN. Computerized Adaptive Testing Using the PROMIS Physical Function Item Bank Reduces Test Burden With Less Ceiling Effects Compared With the Short Musculoskeletal Function Assessment in Orthopaedic Trauma Patients. *Journal of orthopaedic trauma*. 2014; 28(8):439–443. DOI: 10.1097/BOT.0000000000000059 [PubMed: 24378399]
 24. Hung M, Stuart AR, Higgins TF, Saltzman CL, Kubiak EN. Computerized Adaptive Testing Using the PROMIS Physical Function Item Bank Reduces Test Burden With Less Ceiling Effects Compared With the Short Musculoskeletal Function Assessment in Orthopaedic Trauma Patients. *J Orthop Trauma*. 2014; 28(8):439–443. DOI: 10.1097/bot.0000000000000059 [PubMed: 24378399]
 25. Hung M, Voss MW, Bounsanga J, Crum AB, Tyser AR. Examination of the PROMIS upper extremity item bank. *Journal of hand therapy : official journal of the American Society of Hand Therapists*. 2016; doi: 10.1016/j.jht.2016.10.008
 26. Institute of Medicine Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington (DC): National Academies Press (US); 2001.
 27. Jepson C, Asch DA, Hershey JC, Ubel PA. In a mailed physician survey, questionnaire length had a threshold effect on response rate. *J Clin Epidemiol*. 2005; 58(1):103–105. DOI: 10.1016/j.jclinepi.2004.06.004 [PubMed: 15649678]
 28. Kotsis SV, Chung KC. Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder and Hand questionnaire in carpal tunnel surgery. *The Journal of Hand Surgery*. 2005; 30(1):81–86. DOI: 10.1016/j.jhsa.2004.10.006 [PubMed: 15680560]
 29. Lansky D, Nwachukwu BU, Bozic KJ. Using financial incentives to improve value in orthopaedics. *Clinical orthopaedics and related research*. 2012; 470(4):1027–1037. DOI: 10.1007/s11999-011-2127-0 [PubMed: 22002826]

30. Morgan JH, Kallen MA, Okike K, Lee OC, Vrahas MS. PROMIS Physical Function Computer Adaptive Test Compared With Other Upper Extremity Outcome Measures in the Evaluation of Proximal Humerus Fractures in Patients Older Than 60 Years. *Journal of orthopaedic trauma*. 2015; 29(6):257–263. DOI: 10.1097/bot.0000000000000280 [PubMed: 26001348]
31. Rasch G. *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. 1960
32. Relman AS. Assessment and accountability: the third revolution in medical care. *The New England journal of medicine*. 1988; 319(18):1220–1222. DOI: 10.1056/nejm198811033191810 [PubMed: 3173460]
33. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol*. 2008; 61(1):17–33. DOI: 10.1016/j.jclinepi.2006.06.025 [PubMed: 18083459]
34. Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE Jr. The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol*. 2014; 67(5):516–526. DOI: 10.1016/j.jclinepi.2013.10.024 [PubMed: 24698295]
35. Tyser AR, Beckmann J, Franklin JD, Cheng C, Hon SD, Wang A, et al. Evaluation of the PROMIS Physical Function Computer Adaptive Test in the Upper Extremity. *The Journal of hand surgery*. 2014; 39(10):2047–2051. e2044. DOI: 10.1016/j.jhssa.2014.06.130 [PubMed: 25135249]
36. Wainer, H., Dorans, NJ., Flaugher, R., Green, BF., Mislevy, RJ. *Computerized adaptive testing: A primer*. Routledge; 2000.
37. Wright, BD., Masters, GN. *Rating Scale Analysis*. Rasch Measurement. ERIC; 1982.

Figure I.
Person-item map for the PROMIS PF CAT

Figure II.
Person-item map for the DASH

Table I

Patient Sample (N = 1,759).

Characteristics	Mean (SD)	n (%)	Range
Age (years)	55.75 (16.16)		18 – 93
Gender			
Male		985 (56.0%)	
Female		774 (44.0%)	
Race			
White or Caucasian		1,562 (88.8%)	
Asian		16 (0.9%)	
American Indian and Alaska Native		23 (1.3%)	
Native Hawaiian and Other Pacific Islander		7 (0.4%)	
Black or African American		26 (1.5%)	
Other		85 (4.8%)	
Missing		40 (2.3%)	
Ethnicity			
Not Hispanic/Latino		1633 (92.8%)	
Hispanic/Latino		74 (4.2%)	
Missing		52 (3.0%)	

Table II

Item Outfit MNSQ

	OUTFIT MNSQ	ZSTD
PF CAT		
Mean	1.06	0.1
Maximum	2.99	3.8
Minimum	0.12	-0.6
DASH		
Mean	1.04	-0.1
Maximum	2.52	9.4
Minimum	0.62	-2.9

* Note: DASH = Disabilities of the Arm, Shoulder and Hand; MNSQ = Mean Square; ZSTD = Z-Standardized; PF CAT = Physical Function Computerized Adaptive Testing

Table III

Items showing gender DIF

PF CAT Item	Item Description	DIF Measure (Female)	DIF Measure (Male)	p-value
PFA1	Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?	0.08	0.55	0.0036
PFA53	Are you able to run errands and shop?	-0.31	-0.06	0.0128
PFB5R1	Does your health now limit you in hiking a couple of miles (3km) on uneven surfaces, including hills?	-0.03	-0.24	0.0018
PFC56	Does your health now limit you in walking about the house?	-1.86	-1.52	0.0201

DASH Item	Item Description	DIF Measure (Female)	DIF Measure (Male)	p-value
<i>Please rate your ability to do the following activities in the last week.</i>				
DASH1	Open a tight or new jar	-0.73	-0.23	0.0053
DASH5	Push open a heavy door	-0.01	0.44	0.0103
DASH10	Carry a shopping bag or briefcase	-0.19	0.41	0.0006
DASH11	Carry a heavy object	-0.83	-0.36	0.0086
DASH12	Change a lightbulb overhead	-0.36	0.01	0.0233
DASH19	Recreational activities in which you move your arm freely (eg playing frisbee, badminton, etc)	-0.73	-1.14	0.0159
DASH21	Sexual activities	0.73	0.33	0.0152
DASH22	During the past week, <i>to what extent</i> has your arm, shoulder or hand problem interfered with your normal social activities with family, friends, neighbors, or groups?	0.51	-0.11	0.0007
DASH23		0.20	-0.35	0.0041
<i>Please rate the severity of the following symptoms in the last week</i>				
DASH27	Weakness in your arm, shoulder or hand	-0.49	-0.90	0.0249
DASH28	Stiffness in your arm, shoulder, or hand	0.28	-0.37	0.0005
DASH30	I feel less capable, less confident or less useful because of my arm, shoulder, or hand problem	-0.62	-1.53	<0.0001

* Note: DASH = Disabilities of the Arm, Shoulder and Hand; DIF = Differential Item Functioning; PF CAT = Physical Function Computerized Adaptive Testing