Published in final edited form as:

J Hand Surg Am. 2019 April ; 44(4): 267–273. doi:10.1016/j.jhsa.2019.01.008.

Evaluation of Version 2.0 of the PROMIS Upper Extremity Computer Adaptive Test in Non-Shoulder Upper Extremity Patients

Andrew R. Tyser, MD¹, Man Hung, Ph.D.^{1,2}, Jerry Bounsanga, B.S.¹, Maren W Voss, Sc.D.¹, and Nikolas H. Kazmers, MD MSE^{1,*}

¹University of Utah, Department of Orthopaedics, 590 Wakara Way, Salt Lake City, UT 84108

²University of Utah, Division of Public Health, 375 Chipeta Way, Salt Lake City, UT 84108

Abstract

Purpose: The Patient-Reported Outcomes Measurement Information System (PROMIS) Upper Extremity Computer Adaptive Test (UE CAT) was recently updated to version 2.0 (v2.0). We hypothesize that the PROMIS UE CAT v2.0 will exhibit improved performance characteristics compared to the qDASH when administered to a non-shoulder upper extremity patient population.

Methods: The UE CAT v2.0, Physical Function (PF) CAT v2.0, and the QuickDASH were each prospectively administered via tablet computer to all patients presenting to a tertiary hand and upper extremity clinic between April 2017 and October 2017. Patient responses were analyzed, and the mean, range, floor and ceiling effect, and correlations between instruments were calculated.

Results: Among 825 patients, the mean UE CAT v2.0 score was 38.3 (SD 10.7) with a range of 15 – 61 and interquartile range of 15.4. The UE CAT v2.0 had a strong correlation with the qDASH ($r = -0.749$) and PF CAT 2.0 ($r = 0.719$). No patient scored between 56 and 60, indicating a gap in scoring in that range. The UE CAT 2.0 demonstrated a floor effect of 1%, ceiling effect of 6.9%, and a high internal consistency with a Cronbach's alpha of 0.99.

Conclusions: The PROMIS UE CAT v2.0 demonstrated improved ceiling effects, range, and a decreased gap in scoring compared to prior versions. Limitations of the PROMIS UE CAT v2.0 are still present, but updates have led to an incremental improvement over prior versions, demonstrating the ability to influence PROMIS instrument performance through upgrades.

Clinical Relevance: The updated PROMIS UE CAT v2.0 still demonstrates a ceiling effect and gap in scores at the upper end of the instrument, both of which may limit discrimination between different levels of upper extremity function for high-functioning patients.

*Corresponding author: nkazmers@gmail.com, Phone: 248-895-0568.

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.

Supplementary Online Material:

Appendix A - PROMIS UE CAT v2.0 questions.

Appendix B - PROMIS PF CAT v2.0 questions.

Keywords

Computerized adaptive test; Physical function; Patient-Reported Outcomes Measurement Information System (PROMIS); Quick DASH (qDASH); Upper extremity

Introduction:

Patient-reported outcome (PRO) metrics have increased in utilization and importance as part of a larger effort to more fully understand the value of health care delivered in the United States.^{1–4} Recently, the National Institutes of Health (NIH) has led an effort to develop a platform known as the Patient-Reported Outcomes Measurement Information System (PROMIS) in an attempt to standardize and simplify the administration, interpretation, and utility of PROs. PROMIS was designed to accurately assess health status in an extremely broad spectrum of domains, and utilizes Computer Adaptive Testing (CAT) in order to minimize responder burden.⁵ The resultant score of any PROMIS metric has been designed to be easily interpretable, with a mean score of 50 and a standard deviation (SD) of 10 for the reference population. Furthermore, the nature of the CAT design allows for refinement of the PROMIS metrics by updating and expanding the item banks that form the basis of each PRO.

In the field of hand and upper extremity care, the PROMIS Upper Extremity (UE) CAT and Physical Function (PF) CAT have both been investigated for their utility and performance in upper extremity patient populations.^{6–13} Previous versions of both instruments have been found to reduce responder burden and to be highly correlated with each other and with legacy measures in upper extremity patients, yet concerns regarding performance characteristics were identified.^{6,7} Beckmann *et al.* reported on a ceiling effect with the UE CAT v1.0, whereby 10.8% of non-shoulder upper extremity patients reached the maximum score.¹⁴ Similarly, Beleckas *et al.* reported that the UE CAT v1.0 exhibited a ceiling effect in 7% of hand and upper extremity patients, and furthermore, identified a ceiling score of just 56.⁷ Both authors suggested further development of the UE CAT was likely required, and in particular suggested the addition of questions to the item bank to evaluate the upper end of function.

Recently, updated iterations of both the PROMIS UE and PF CAT were released for clinical use. Much like updated computer software, updated versions of the PROMIS CATs are identified by their version numbers (v1.0, v1.2, v2.0, etc). Updated versions of a PROMIS CAT should, in theory, be able to address perceived shortcomings in various performance characteristics of prior iterations of the metric when applied to target populations in a clinical setting.

The purpose of the study was to evaluate the performance characteristics of the PROMIS UE CAT v2.0, PF CAT v2.0, and qDASH in a tertiary, non-shoulder hand and upper extremity outpatient clinic setting. Our secondary aim was to compare the ceiling effects and maximum achievable score of the UE CAT v2.0 to reported values for the UE CAT v1.0.

Methods:

This study was approved by our Institutional Review Board (IRB). Between August and November 2017, the QuickDASH and Versions 2.0 of the PROMIS UE CAT (Appendix A) and PROMIS PF CAT (Appendix B) were each administered via tablet computer immediately prior to the clinic visit to all patients presenting to a tertiary non-shoulder hand and upper extremity clinic in a prospective manner. Patients under age 18 were excluded. Patients who completed all three metrics during a clinic visit were included in this retrospective study. For those patients who completed multiple surveys at different clinic visits during the study period, only data from their first encounter were utilized.

Surveys were not monitored for completeness during administration, patients could decline the survey, and the surveys did not force answers. qDASH scores were not calculated for surveys with more than one missing response, and the PROMIS surveys were considered scoreable based on the CAT protocols utilized. The instruments were randomized in the order of administration, and data tabulated online via wireless connection to a secure server.

Basic descriptive statistics were calculated to illustrate the mean, median, SD, and range of scores for the three questionnaires. The PROMIS UE CAT and PF CAT scores were evaluated for normality in their distribution using the Shapiro-Wilk test. Potential association between the PROMIS UE CAT v2.0, and PF CAT 2.0 and qDASH was assessed by calculating Pearson correlation coefficients. The magnitude of resulting r-values was interpreted as follows in terms of strength of the association: no association ($r < 0.2$), weak ($0.2 \leq r < 0.4$), moderate ($0.4 \leq r < 0.6$), strong ($0.6 \leq r < 0.8$), and very strong ($r \geq 0.8$).

Using a Rasch Partial Credit Model, ceiling and floor effects and reliability were evaluated for each metric. A ceiling effect was defined as the percentage of surveys that reached the maximum possible score, with the converse being the case for floor effects. Reliability was assessed using person and item reliability, as well as the more commonly reported Cronbach's alpha. Person reliability is a measure utilized in Rasch analyses that assesses the degree to which a predicted response is reflected by the actual response for a given respondent. Item reliability is a statistical measure of the internal consistency, as is Cronbach's alpha (where high internal consistency is typically desired).

An *a priori* sample size estimate was conducted. Assuming 10% measurement error, a minimum number of 119 respondents were needed to achieve a 2-tailed 99% confidence interval in a Rasch analysis (with an item calibration within 0.5 logits).

Results:

A total of 825 patients were included in the study. Baseline patient characteristics are described in Table 1. The study excluded 1478 patients due to a failure to complete all legacy and PROMIS surveys. The mean UE CAT v2.0 score was 38.3 (SD 10.7) with a range of 15 – 61, and interquartile range (IQR) of 15.4 (Table 2). Neither the UE CAT nor the PF CAT scores were distributed normally ($p < 0.05$). There was an observed gap in the UE CAT scores between 56 and 60, whereby no patient scored within this range (Figure 1).

The PROMIS UE CAT v2.0 demonstrated a strong correlation with the qDASH ($r = -0.749$, $p < 0.05$; Figure 2), and a strong correlation with the PF CAT 2.0 ($r = 0.719$, $p < 0.05$; Figure 3). The UE CAT v2.0 exhibited a floor effect of 1% and ceiling effect of 6.9%. New and return patients exhibited similar ceiling effects, which measured 7.1% and 6.6%, respectively. Rasch partial credit fit statistics are illustrated in Table 3. Rasch partial credit reliabilities (item and person), and Cronbach's alpha, are reported for the UE CAT v2.0 in Table 4. All three measures indicated high reliability.

Discussion:

The PROMIS UE CAT is a relatively new measure of patient-reported upper extremity function that has been shown to be reliable and responsive, exhibit low responder burden, and is correlated with other instruments such as the QuickDASH and PF CAT.^{7,9,14,15} As clinical testing of the performance of the UE CAT and PF CAT has progressed, targets for improvements in performance have been identified.^{6,7,11,16} With UE CAT v1.0, a ceiling effect has been noted by several authors that may limit the measure's ability to differentiate patients with higher upper extremity function. This is potentially problematic in the measurement of treatment outcome, as a metric with a large ceiling effect may not be able to capture accurately or fully the improvement in clinical outcome that a patient may truly experience. This could lead to the invalid conclusion in a clinical or research setting that a given treatment outcome was not clinically important to a patient.

The PROMIS platform differs from previously developed upper extremity PROs in several important ways, including its ability to undergo modifications that aim to improve its performance.^{5,16} Potential updates to PROMIS metrics include expanding the item bank and altering aspects of the software upon which the test relies.

The PROMIS UE CAT v2.0 item bank consists of 46 questions related to upper extremity function, the prior version's 16 items, and an additional 30 items.¹⁷ The computer adaptive test (CAT) algorithm selects questions from the item bank based on previously answered questions to increase the accuracy and efficiency of the metric. In theory, expansion of the item bank should allow for improved performance, and—if the added items assess the higher end of function—should serve to decrease ceiling effects.

The main finding of this study is that the UE CAT v2.0 shows some improved performance over prior versions of the UE CAT, yet still appears to demonstrate limitations at the upper end of the scale. Specifically, we noted a ceiling effect of 6.9% and a maximum score of 61. The reasons for this finding are speculative, but a persistent ceiling effect may be due to an item bank that continues to fail in assessing perceived limitations in patients with a high-level of baseline hand function. An additional finding of this study was that the scores of UE CAT and PF CAT scores were not normally distributed in this population, which could affect both the interpretation and analyses of PROMIS scores.

These findings are important for several reasons, and are worth illustrating. In the PROMIS framework, the stated goal of each metric is to represent the mean of a normally-distributed population with a score of 50, with a standard deviation of 10. In this case, one would

necessarily expect some percentage of patients to demonstrate baseline or “normal” hand function, which, for them, is higher (or lower) than the population mean. Certainly, various athletes, musicians, surgeons, and possibly a variety of other patients, might exhibit hand function that is potentially one or two standard deviations above the mean of the entire population, that such individuals would consider “normal” for themselves.

To illustrate this, assume such a patient’s hand function is 2-SD above the population mean at baseline, and that this is “normal” for them. In theory, this upper extremity function should be represented by a score of 70 on the PROMIS UE CAT—a score which currently does not appear to be achievable due to the functioning of the instrument. Instead, this patient would score a 61 on the metric, the current maximum possible score of the UE CAT 2.0.

Let us further assume that this same high functioning patient perceives a loss of function due to some condition or injury, the magnitude of which should be represented by a loss of 10 points on the PROMIS UE CAT. In this scenario, with a theoretical “normal” score of 70, the UE CAT 2.0 should be able to measure the “injury state” of this patient with a score of 60.

In this situation, if the patient recovered back to their baseline of 70, the UE CAT 2.0 would again only be able to measure this patient’s “normal” state with a score of 61 (the current maximum score of the scale). The UE CAT 2.0 would demonstrate an improvement of 1 point, whereby the “actual” change in function was 10 points. In other words, the treatment effect would be incorrectly measured as 1, when in reality it was 10. This inability to accurately measure upper extremity function at the high end of the scale is clearly problematic in assessing changes in function in patients with higher levels hand function, including in response to treatments.

In comparison to the ceiling effect of 6.9% reported here, Beckmann *et al.* reported a ceiling effect of 10.8% in the UE CAT v1.0 in 379 hand and upper extremity (non-shoulder) patients.¹⁴ A separate study reported a ceiling effect of 7.2% in 5,202 adult patients (including shoulder) completing the UE CAT v1.0.⁷ The authors noted mean UE CAT v1.0 scores that were 4 points lower for patients presenting with shoulder or elbow conditions, compared to those presenting with hand conditions. In a smaller cohort, Doring *et al.* reported a ceiling effect of 0% in 84 patients using the UE CAT v1.0, although the method of calculating the ceiling effect differed from the current study.¹³

In addition, the current study noted a maximum score of 61 with the UE CAT v2.0, as well as a gap in scoring which occurred between 56 and 60, where no patient scored within that range. In comparison, Beleckas *et al.* reported a maximum score on the UE CAT v1.0 of 56, as well as a gap or secondary ceiling in scoring that occurred between 50 and 56. It is unclear to us why the maximum score achieved was only 61 in this study, as the PROMIS functional metrics have been designed to conform to a 0 – 100-point scale, with a score of 50 representing the mean of the population at large, 0 representing the minimum, and 100 representing the maximum function. It is possible that the calibration of the scoring system of the UE CAT requires further refinement, or that the addition of higher-level questions to the item bank may help to ameliorate this issue. Similarly, while speculative, we feel that the

observed scoring gap may be an artifact of the software used to score the UE CAT. As such, improvements in the UE CAT software may be required to improve this outcome measure.

Our correlative findings are congruent with prior literature utilizing prior versions of the PROMIS UE CAT. Kazmers *et al.* observed a strong correlation between the UE CAT v1.2 and PF CAT v1.2, and a strong correlation between the UE CAT v1.2 and qDASH.⁸ Doring *et al.* also reported a strong correlation between the UE CAT and qDASH.¹³ Although the UE CAT version was not reported, it was likely v1.0 or v1.2 given the publication date of 2014.

Our study has limitations that deserve mention. Our single center cross-sectional study design considered patients only at a single point in time, and we did not compare scores before and after treatment. Responsiveness^ the ability of an instrument to detect change when a meaningful change has occurred—could not be investigated based on this study design. The patient population treated within our practice is likely less diverse than the US population as a whole and may differ in other unrecognized ways from those in other parts of the US or the world. While we have no reason to suspect that unrecognized differences between our patient population and those found within other practices would influence the findings of this study, this is a possibility. Additional limitations of this study are that the PF CAT and UE CAT were compared with just 1 legacy instrument, the qDASH. This was chosen because the qDASH is a widely used instrument for non-shoulder upper extremity complaints,^{12,18,19}. The findings in this study should not be extrapolated to other instruments or patient populations. Additionally, our study excluded 1478 patients due to incomplete surveys. Low response rates can lead to non-response bias, which, in turn, can potentially limit the conclusions of clinical studies. It is possible that low response rates have effects on the evaluation of instrument performance, and we suggest that further study in this arena would be warranted given the increasing focus on patient-reported outcome and survey data. We are unaware, however, of any demonstrated effects of low response rates to surveys on the fundamentals of instrument performance such as floor and ceiling effects. The high Cronbach's alpha value observed in the current study suggests that question items have a high level of covariance (i.e. that the outcome measures are internally consistent), however exploratory factor analyses would be required to comment upon dimensionality of the instrument.

In interpreting the findings of the current study in relation to prior work on the PROMIS UE CAT v1.0, the updated UE CAT v2.0 has served to slightly increase the maximum score achieved by upper extremity patients by 5 points, from 56 to 61, and to slightly decrease the apparent gap in scoring. In non-shoulder hand patients, the ceiling effect was only slightly decreased from 10.8% in the UE CAT v1.0 to 6.9% in v2.0, while maintaining a low floor effect (1%). Importantly, there does not appear to be a consensus as to what value constitutes an acceptable ceiling effect, given that it is difficult to entirely eradicate all ceiling effects of functional measurement instruments.

In order to address a persistent ceiling effect that appears to be present, it is our opinion that the UE CAT item bank should be expanded by adding questions that are better able to assess higher-level upper extremity function.

While our findings demonstrate the ability to improve PROMIS instrument performance through upgrades, we suggest that additional modification of the PROMIS UE CAT is indicated to further reduce the ceiling effect and scoring gap.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

This study was funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under award number U01AR067138.

References:

1. Smith MV, Calfee RP, Baumgarten KM, Brophy RH, Wright RW. Upper extremity-specific measures of disability and outcomes in orthopaedic surgery. *JBone Joint Surg Am.* 2012;94:277–285. [PubMed: 22298061]
2. Hoang-Kim A, Pegreffi F, Moroni A, Ladd A. Measuring wrist and hand function: Common scales and checklists. *Injury.* 2011;42:253–258. [PubMed: 21159335]
3. Alderman AK, Chung KC. Measuring outcomes in hand surgery. *Clin Plast Surg.* 2008;35:239–250. [PubMed: 18298996]
4. Chung KC, Burns PB, Sears ED. Outcomes research in hand surgery: Where have we been and where should we go? *J Hand Surg Am.* 2006;31:1373–1379. [PubMed: 17027802]
5. Cella D, Yount S, Rothrock N, et al. The patient-reported outcomes measurement information system (promis): Progress of an nih roadmap cooperative group during its first two years. *Med Care.* 2007;45:S3–S11.
6. St John MJ, Mitten D, Hammert WC. Efficacy of promis pain interference and likert pain scores to assess physical function. *J Hand Surg Am.* 2017;42:705–710. [PubMed: 28709796]
7. 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;42:867–874. [PubMed: 28709794]
8. Kazmers NH, Hung M, Rane AA, Bounsanga J, Weng C, Tyser AR. Association of physical function, anxiety, and pain interference in nonshoulder upper extremity patients using the promis platform. *J Hand Surg Am.* 2017;42:781–787. [PubMed: 28890332]
9. Hung M, Voss MW, Bounsanga J, Crum AB, Tyser AR. Examination of the promis upper extremity item bank. *J Hand Ther.* 2017;30:485–490. [PubMed: 27919565]
10. Anthony CA, Glass NA, Hancock K, Bollier M, Wolf BR, Hettrich CM. Performance of promis instruments in patients with shoulder instability. *The American Journal of Sports Medicine.* 2016;
11. Beckmann JT, Hung M, Bounsanga J, Wylie JD, Granger EK, Tashjian RZ. Psychometric evaluation of the promis physical function computerized adaptive test in comparison to the american shoulder and elbow surgeons score and simple shoulder test in patients with rotator cuff disease. *J Shoulder Elbow Surg.* 2015;24:1961–1967. [PubMed: 26321484]
12. Overbeek CL, Nota SP, Jayakumar P, Hageman MG, Ring D. The promis physical function correlates with the quickdash in patients with upper extremity illness. *Clin Orthop Relat Res.* 2015;473:311–317. [PubMed: 25099262]
13. Doring AC, Nota SP, Hageman MG, Ring DC. Measurement of upper extremity disability using the patient-reported outcomes measurement information system. *J Hand Surg Am.* 2014;39:1160–1165. [PubMed: 24799143]
14. Beckmann JT, Hung M, Voss MW, Crum AB, Bounsanga J, Tyser AR. Evaluation of the patient-reported outcomes measurement information system upper extremity computer adaptive test. *Journal of Hand Surgery-American Volume.* 2016;41:739–U146.

15. Hung M, Saltzman CL, Greene T, et al. The responsiveness of the promis instruments and the qdash in an upper extremity population. 2017;1–13.
16. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the promis physical function item bank in orthopaedic patients. Journal of Orthopaedic Research. 2011;29:947–953.
17. www.healthmeasures.net. Accessed 07/06/2018.
18. Beaton DE, Wright JG, Katz JN, Grp UEC. Development of the quickdash: Comparison of three item-reduction approaches. Journal of Bone and Joint Surgery-American Volume. 2005;87a:1038–1046.
19. Gummesson C, Ward MM, Atroshi I. The shortened disabilities of the arm, shoulder and hand questionnaire (quickdash): Validity and reliability based on responses within the full-length dash. BMC Musculoskeletal Disord. 2006;7:44.

Figure 1 -
Histogram representing distribution of PROMIS UE CAT v2.0 scores.

Figure 2 -
Scatterplot of PROMIS UE CAT v2.0 scores and qDASH.

Figure 3 -
Scatterplot of PROMIS UE CAT v2.0 scores and PROMIS PF CAT v2.0.

Table 1 -

Baseline Patient Characteristics.

Factor	Value (n = 825)
Age (years $\pm$ SD)	50.3 $\pm$ 16.8
Ethnicity	
Non-Hispanic	753 (91%)
Hispanic	58 (7%)
Unknown/Missing	14 (2%)
Insurance Status	
Commercial	525 (64%)
Medicare	181 (22%)
Medicaid	69 (8%)
Worker's Compensation	32 (4%)
Patient Pay	15 (2%)
Other Government Insurance ^a	4 (0.5%)
Race	
White or Caucasian	714 (87%)
Asian	20 (2%)
Black or African American	10 (1%)
Native American/Pacific Islander	8 (1%)
Other	60 (7%)
Unknown/Missing	13 (2%)
Smoking Status	
Never	580 (70%)
Quit	152 (18%)
Current	88 (11%)
Unknown	4 (0.5%)
Passive Exposure	2 (<0.5%)
Sex (female)	468 (57%)
Work Status	
Employed - Full Time	374 (45%)
Retired	158 (19%)
Unemployed	123 (15%)
Self Employed	51 (6%)
Employed - Part Time	41 (5%)
Disabled	34 (4%)
Student	29 (4%)
Unknown	15 (2%)

Continuous data presented as mean $\pm$ standard deviation. Categorical data presented as number of patients and (percentage).

^aIndian Health Service and US Department of Labor

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2 -
Descriptive statistics summarizing patient-reported outcomes for the study cohort.

Instrument	Mean (SD)	Median	Interquartile Range	Range
PROMIS UE CAT v2.0	38.2 (10.7)	36.8	15.4	15–61
PROMIS PF CAT v2.0	42.9 (10.4)	42.1	12.0	15–75
qDASH	32.4 (21.2)	29.5	34.1	0–95

Table 3 -

Rasch partial credit analysis results: PROMIS UE CAT v2.0 fit statistics.

Instrument	Mean	Maximum	Minimum	ZSTD
Outfit MNSQ	1.00	9.90	0.03	-0.03
Infit MNSQ	0.90	7.73	0.03	-0.04

Table 4 -
Rasch partial credit analysis results: PROMIS UE CAT v2.0 instrument descriptions.

Instrument	Mean
Item reliability	0.91
Person reliability	0.94
Cronbach's Alpha	0.99