



Development and calibration data for the Medication Adherence Item Bank: a new computer adaptive test for persons with type 2 diabetes mellitus

Noelle E. Carlozzi⁵ · Michael A. Kallen¹ · Jonathan Troost² · Alexa Bragg³ · Jessica Martin-Howard⁴ · Ioana Moldovan⁴ · Jennifer A. Miner⁵ · Brian W. Jack^{3,4} · Suzanne Mitchell^{3,4}

Accepted: 7 October 2022 / Published online: 28 October 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract

Purpose To develop a new computer adaptive test that evaluates important aspects of medication adherence for persons with type 2 diabetes mellitus.

Methods Two hundred and twenty-five people with type 2 diabetes mellitus completed 41 items related to medication adherence.

Results Exploratory analysis supported the essential unidimensionality of the initial item pool. Five items were deleted due to low item-adjusted total score correlations (resulting in 36 items). Confirmatory factor analysis supported the retention of 27 items. A graded response model identified no items for exclusion, based on misfit. No items were flagged for meaningful differential item functioning (DIF). The final item bank is comprised of 27 items; an associated 6-item short form was constructed that balanced both psychometric factors (e.g., item information values) and clinical input. Initial analysis of the simulated CAT and static short form supported both the reliability (i.e., internal consistency and test–retest reliability) and validity (i.e., convergent, discriminant, and known groups) of both administration formats.

Conclusions The new medication adherence item bank provides a reliable and valid assessment of the ability to take medications accurately among people with T2DM; it will be available in early 2023 through healthmeasures.net.

Keywords Type 2 diabetes · Healthcare · Patient-reported outcomes

Introduction

The medical management of type 2 diabetes mellitus is complex and commonly includes both pharmacotherapy and lifestyle changes (e.g., diet and exercise). Given these complexities, medication adherence is often poor, as elucidated in several systematic reviews [1–3]. Poor medication adherence is associated with several negative effects, including acute and chronic complications, with increased health-care utilization and hospitalizations [1–3]; these effects are both costly and can have a deleterious impact on individual health-related quality of life (HRQOL). Medication adherence is also influenced by social determinants that present additional barriers, including socioeconomic factors, health literacy, comorbid mental health symptoms/conditions, and behavioral skills [4, 5].

Approaches to assessing medication adherence include both objective and subjective evaluation (see Clifford et al. [6]). With regard to objective assessments, the medication

✉ Noelle E. Carlozzi
carlozzi@med.umich.edu
https://coda.med.umich.edu/

¹ Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

² Michigan Institute for Clinical and Health Research, University of Michigan, Ann Arbor, MI, USA

³ Department of Family Medicine, Boston University School of Medicine, Boston, MA, USA

⁴ Department of Family Medicine, Boston Medical Center, Boston, MA, USA

⁵ Department of Physical Medicine & Rehabilitation, Center for Clinical Outcomes Development and Application (CODA), University of Michigan, North Campus Research Complex, 2800 Plymouth Road, Building NCRC B14, Room G216, Ann Arbor, MI 48109-2800, USA

possession ratio (MPR) [7], which reflects the number of days of medication supplied within the refill interval divided by the number of days in the refill interval, is a common objective approach to adherence assessment [6]. Although this is typically considered the gold standard approach for objective evaluation of medication adherence, it requires access to pharmacy claims databases or pharmacy/medical records to calculate, and it may not allow for the consideration of individuals that may fill their prescriptions across multiple pharmacies. In addition, this approach does not take into account whether or not the patient is taking the medication in the correct way, only that he/she is filling a prescription; as such, it commonly overestimates adherence: Patients who routinely refill prescriptions will have an inflated MPR, and those that refill in advance of the prescription running out will also have an inflated MPR. The use of electronic pill-caps that register the date and time of pill-bottle openings provides an alternative objective approach to medication adherence monitoring. Yet these devices are expensive and, while they provide more granular information about pill-taking behavior, they still do not confirm/ensure that an individual is actually taking the prescribed dose, only that he/she has opened the pill-bottle.

Given the disadvantages of these objective approaches, subjective ratings of medication adherence are commonly used [6]. Advantages of subjective approaches include the low cost and the low burden associated with the administration of this type of assessment. These advantages are mitigated by the fact that subjective reports, by design, rely on retrospective reporting, which can be biased and commonly overestimate actual adherence rates. One of the most common self-report measures of medication adherence is the Morisky Medication Adherence Scale [8–10]. The psychometric characteristics of this measure, albeit adequate (i.e., generally meeting minimal standards for internal consistency reliability, adequate test–retest reliability, yet suboptimal sensitivity/specificity [8]), are typically less robust than the psychometric reliability and validity that has been reported for newer-generation patient-reported outcome (PRO) measures.

The purpose of this report is to detail the development of a new patient-reported outcome (PRO) measure that was developed using classical test theory and item response theory methodologies [11–13].² This new item bank, the Re-Engineered Discharge for Diabetes Computer Adaptive Test (REDD-CAT) medication adherence item bank, was developed to evaluate medication adherence in people with T2DM. This item bank was developed according to the rigorous methodology [13] required for measure inclusion within the Patient-Reported Outcomes Measurement Information System (PROMIS) [14, 15] and Neuro-QoL measurement systems [16, 17], i.e., comprehensive measurement systems that capture important aspects of HRQOL. It

is expected that the REDD-CAT medication adherence Item Bank would be adopted within the PROMIS and Neuro-QoL measurement systems once it is fully developed. Below we describe the analytical process that informed the development of this new PRO, and we highlight the preliminary reliability (i.e., internal consistency reliability), and validity (i.e., concurrent validity, discriminant validity, known-groups validity) observed for this new tool.

Methods

Study participants

Two hundred and twenty-five people with type 2 diabetes were enrolled in this study. Participants had to be age 18+, have a diagnosis of T2DM, be able to converse in English, and be capable of providing informed consent in order to participate in the study. In addition, participants were required to correctly pronounce the first 10 words on the Wide Range Achievement Test 4th Edition (WRAT4) Reading Subtest [18] in order to complete study assessments independently; those with one or more errors on the first 10 words were assisted by a study coordinator in completing the study assessments. Participants were recruited through Boston Medical Center (BMC) by: (1) screening weekly lists of diabetes outpatients with upcoming appointments (using BMCs Clinical Data Warehouse to identify eligible participants), (2) using BMC internal census reports from the electronic health record to identify potential inpatient participants, and (3) by contacting individuals that had previously participated in research on T2DM and had agreed to be contacted for other studies. This study was conducted in accordance with institutional review board approval. Participants were required to provide informed consent prior to their participation in study-related activities.

Measures

Medication Adherence

The Medication Adherence item bank (i.e., calibrated set of items) was developed to assess the knowledge and ability to take medications as prescribed, including getting prescriptions filled on time and taking medications at the correct time. Both qualitative and quantitative methodologies were used to develop this new item bank. This process involved the creation of an item pool (i.e., an uncalibrated set of items) that reflected important issues related to medication adherence for people with T2DM (see Cancino et al. 2014 [19] and <https://www.pcori.org/research-results/2012/pilot-project-figuring-out-which-patients-are-likely-return-hospital>). Response options for the items in this pool include two

different 5-point Likert scales: never (1), rarely (2), sometimes (3), usually (4), always (5); strongly disagree (1), disagree (2), neither agree nor disagree (3), agree (4), and strongly agree (5).

The initial item pool was refined using feedback from experts, an item reading-level assessment (to ensure no greater than a 5th-grade reading level for the final measure), translatability review (to allow for future translations of the measure into other languages), as well as feedback from patients with T2DM. The item pool was then field tested in a large sample of people with T2DM, and classical test theory and item response theory analytic approaches were used to inform the development of the final item bank. A brief summary of this process is provided below, with additional details reported in another manuscript within this issue [20]. The final Medication Adherence item bank provides a score that is on the T-score metric (mean = 50; $SD = 10$); higher scores indicate more problems with taking medications as prescribed (i.e., more difficulty with medication adherence). Note that these T-scores are normalized relative to the calibration sample which is comprised of other people with T2DM. Computer adaptive test (CAT) scores from the medication adherence item bank were simulated using Firestar Version 1.3.2 [21].

Neuro-QoL Anxiety [15–17, 22]

The Neuro-QoL Anxiety CAT was used to assess perceptions of worry, fear, hyperarousal, and anxiety using a 5-point Likert scale (1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always). This measure is scored on the T-score metric ($M = 50$, $SD = 10$), with higher scores representing more perceived anxiety relative to a calibration sample of individuals with neurological conditions. This measure was used to examine convergent validity of the new medication adherence measure. Reliability, validity, and responsiveness data for Neuro-QoL anxiety have been reported in the general population, as well as in people with chronic conditions including adults with epilepsy, multiple sclerosis, Parkinson's disease, and Huntington disease [23–29]. Internal consistency is typically excellent (all Cronbach's $\alpha > 0.91$ for the different cohorts), 3-day test–retest reliability is excellent ($ICC = 0.96$ in Huntington disease), 7-day test–retest reliability ranges from adequate to very good ($ICCs$ range from 0.68 to 0.81 across the other clinical cohorts), convergent validity is supported by moderate to strong correlations with other measures of mental health or emotional well-being (all $r's > 0.63$), discriminant validity is supported by less robust correlations with measures of social health or motor functioning ($r's < 0.50$ for measures of social health and $r's < 0.36$ for measures of motor function), known-groups validity was supported by group differences among patients with more severe disease in the epilepsy,

multiple sclerosis, and Parkinson's disease cohorts, responsiveness was supported by worsening anxiety over a 6-month period in the epilepsy and multiple sclerosis cohorts, and sensitivity and specificity that support clinical decision making were reported in Huntington disease.

Neuro-QoL Depression [16, 17]

The Neuro-QoL Depression CAT was used to assess perceptions of sadness and depression using a 5-point Likert scale (1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = always). This measure is scored on the T-score metric ($M = 50$, $SD = 10$), with higher scores representing more perceived depression relative to a calibration sample of individuals with neurological conditions. This measure was used to examine the convergent validity of the new Medication Adherence measure. Reliability, validity, and responsiveness data for Neuro-QoL Depression have been reported for the general population and in people with chronic conditions including adults with epilepsy, multiple sclerosis, Parkinson's disease, and Huntington disease [23–28, 30]. Internal consistency is typically excellent (all Cronbach's $\alpha > 0.92$ for the different cohorts), 3-day test–retest reliability is excellent ($ICC = 0.95$ in Huntington disease), 7-day test–retest reliability ranges from adequate to good ($ICCs$ range from 0.72 to 0.82 across the other clinical cohorts), convergent validity is supported by moderate to strong correlations with other measures of mental health or emotional well-being (all $r's > 0.67$), discriminant validity is supported by less robust correlations with measures of social health or motor functioning ($r's < 0.38$ for measures of social health and $r's < 0.36$ for measures of motor function), known-groups validity was supported by group differences among patients with more severe disease in the epilepsy, multiple sclerosis, and Parkinson's disease cohorts, responsiveness was supported by worsening depression over a 6-month period in the epilepsy and multiple sclerosis cohorts, and sensitivity and specificity that support clinical decision making were reported in Huntington disease.

Healing encounters and attitudes lists (HEAL) Patient–Provider Connection [31]

The 7-item HEAL Patient–Provider Connection short form (SF) was used to assess perceptions of trust and satisfaction with one's healthcare provider using a 5-point Likert scale (1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much). This measure is scored on the T-score metric ($M = 50$, $SD = 10$) relative to a calibration sample of individuals from the general population, with higher scores representing better perceived patient–provider connection. This measure was used to examine the discriminant validity of the new medication adherence measure. The measurement

development publication for the HEAL Patient–Provider Connection SF includes data to support its reliability and validity in the general population [31]; Cronbach’s alpha was 0.96 and there was a moderate relationship with a measure of outpatient clinical care ($r=0.38$).

HEAL Healthcare Environment [31]

The 6-item HEAL Healthcare Environment SF was used to assess perceived satisfaction with the healthcare staff and facility using a 5-point Likert scale (1 = not at all, 2 = a little bit, 3 = somewhat, 4 = quite a bit, 5 = very much). This measure is scored on the *T*-score metric ($M=50$, $SD=10$) relative to a calibration sample of individuals from the general population; higher scores represent more healthcare environment satisfaction. This measure was used to examine the discriminant validity of the new medication adherence measure. The measurement development publication for the HEAL Healthcare Environment SF includes data to support its reliability and validity in the general population [31]; Cronbach’s alpha was 0.92 and there was a moderate relationship with a measure of outpatient clinical care ($r=0.39$).

Economic Quality of Life (Econ-QOL) [32]

This 8-item Econ-QOL SF was used to examine economic and financial security. This measure is scored on the *T*-score metric ($M=50$, $SD=10$); higher scores indicate poorer perceived economic quality of life. To establish known-groups validity, we examined those with “better” (scores ≤ 40) versus “worse” (scores ≥ 60) economic problems. Reliability (all Cronbach’s alpha > 0.91 supporting internal consistency reliability) and validity (convergent validity supported by r ’s > 0.46 with self-reported income and known-groups validity supported by being able to differentiate those above versus below the poverty line) for the Econ-QOL SF have been reported in people with disabilities (i.e., traumatic brain injury, spinal cord injury, and stroke) [33–35] and in caregivers of people with traumatic brain injury [36].

Data collection

Data were captured using REDCap, a HIPAA-compliant secure online data capture system. Participants completed the assessments in-person ($n=200$) or at home via an online survey link ($n=3$) or over the phone with study staff ($n=1$); $n=21$ participants completed the assessments using more than one survey approach. This includes $n=15$ participants who began the survey in-person, but finished with the survey via telephone with assistance from a study team member (participants include, but are not limited to, those who received reading assistance from staff due to failing the WRAT), $n=3$ participants who began in person and

completed the assessments at home using a study-specific url, and $n=3$ participants who began in person, continued via telephone with the assistance of study staff, and finished using the emailed study link.

Statistical analyses

Item bank development: qualitative analyses

As recommended by current measurement development standards, both qualitative and quantitative methods, incorporating classical test theory and item response theory analyses, were used in the development of the new medication adherence item bank [13]. The details for the development of the medication adherence item pool, which went through several iterations of expert review (with T2DM and/or psychometrics specialists), cognitive interviews with patients with T2DM, reading-level assessment (to ensure no more than a 5th-grade reading level), and Spanish translatability review (to facilitate future translations), are provided in the Supplemental Appendix.

Item bank development: quantitative analyses

Following our review of the literature and our analyses of qualitative interview data, we utilized classical test theory (CTT) and item response theory (IRT) analytic methods to develop the calibrated item bank. A detailed summary of our analytical approaches can be found in the Supplemental Appendix of the Healthcare Access manuscript which can be found in this same issue of *Quality of Life Research* [20]. In brief, our quantitative analyses were as follows. Full-sample exploratory and confirmatory factor analyses (EFA, CFA), in conjunction with clinical input, were used to refine the item pool into a unidimensional set of items [37–39]. Unidimensionality would be supported by EFA where: (1) the ratio of eigenvalue 1 to eigenvalue 2 is ≥ 4 ; and (2) eigenvalue 1 accounts for $\geq 40\%$ of item set variance. Candidate items for exclusion included: (1) items with sparse cells–response categories with $n < 5$ responses; (2) items with low (i.e., < 0.40) item-adjusted total score correlations; and (3) items that were non-monotonic (we investigated monotonicity using non-parametric IRT models of item-rest plots and expected score by latent trait plots) [Testgraf Software; 40]. With regard to CFA, essential unidimensionality would be supported if the: (1) comparative fit index (CFI) ≥ 0.90 ; (2) Tucker-Lewis index (TLI) ≥ 0.90 ; and (3) root mean square error of approximation (RMSEA) < 0.10 [38, 41–46]. For comparative fit purposes, we also obtained the chi-square value for model fit and its associated *p* value. Here, candidate items for exclusion included: (1) items with low factor loadings ($1x < 0.50$); (2) items with residual correlations > 0.20 ; and (3) items with correlated error

modification index values ≥ 100 [37–39, 47–51]. EFA and CFA analyses were conducted in Mplus (Version 7.4) [52]. When CFA overall model fit criteria were not fully met, we conducted confirmatory bi-factor analyses (CBFA) [39, 53] to obtain comparators to traditional fit analyses. CBFA can be used to assess whether data are “unidimensional enough” to fit with a unidimensional measurement model [54]. CBFA provides a set of indices to assess factor strength, including omega, omega-Hierarchical (omega-H), and explained common variance (ECV). For our interests, omega-H provides a dimensionality index: A threshold general factor omega-H value > 0.80 has been recommended for establishing a measure’s essential unidimensionality [55].

EFA and CFA analyses were followed by fitting a common-slope IRT model (i.e., a constrained graded response model; GRM) [56] to the data. The following item types were candidates for exclusion: (1) items with significant misfit ($S-X^2 / df$ effect size > 3) [57–60] and (2) items exhibiting impactful differential item functioning (DIF). We defined impactful DIF as (1) $p < 0.01$ for a group-specific item parameter difference, along with a weighted area beneath the curve (wABC) effect size > 0.30 for any DIF candidate item tested [61], and (2) $> 2\%$ of DIF-corrected vs. uncorrected score differences exceeding individual case uncorrected score standard errors. We examined DIF for age (< 60 vs. ≥ 60 years), sex (male vs. female), education (\leq high school vs. $>$ high school), and socioeconomic status (“have enough income to pay rent/mortgage” and “can afford to pay bills on time,” both categorized as *never/rarely/sometimes* vs. *usually/always*). GRM analyses and DIF analyses were conducted in IRTPRO (version 3.1.2) [62]; DIF analyses used iterative Wald-2 testing, an analytical approach that tests candidate items for DIF by establishing a DIF-free set of items [63, 64]. In Wald-2 testing Step 1, we identified a DIF-free set of anchor items, while in Step 2 we tested any identified candidate items for DIF. Subgroup-specific parameters were estimated for each candidate DIF item, using the constrained GRM. The parameters were then compared across subgroups—total DIF, slope-related DIF, and threshold-related DIF—to identify statistically significant parameter differences with non-trivial effect sizes. A final CFA was then conducted to confirm the unidimensionality of the final items in the bank (using the aforementioned CFA item and model fit criteria).

Slope and threshold information from our GRM analyses (i.e., the estimated item parameter values) were used to program computer adaptive test administration of the final item bank [65–71]. For a more clinically relevant CAT performance assessment, we simulated the item responses from $N = 2000$ cases drawn from a clinical population (i.e., having a mean one SD in the direction of worse health status). We optimized CAT administration parameters (e.g., number of items to administer, targeted score reliability level)

to balance response burden and score precision. Firestar software [72] was used to simulate CAT administration and obtain CAT scores for the new item bank. A 6-item short form (SF) was selected that balanced clinician input with psychometric considerations, emphasizing item content coverage and achievable score-level reliabilities.

Preliminary reliability and validity analyses

Normality was assessed to ensure that parametric analyses were appropriate for the examination of obtained scores and supporting data. Internal consistency reliability analyses included Cronbach’s alpha estimation for the full bank and SF scores and an IRT-based estimate [73] for the simulated CAT scores (a priori criterion ≥ 0.70) [74]. Floor and ceiling effects, which are revealed by investigating the percentages of participants with the lowest possible and highest possible scores, were examined for both administration formats, i.e., the full bank and the newly developed SF (a priori criteria for floor and ceiling effects were $\leq 20\%$) [75, 76]. We divided the raw CAT item response score by the number of items administered in order to examine floor and ceiling effects for the CAT (i.e., a quotient score of “1” was considered a “floor effect” and a quotient score of “5” a “ceiling effect”).

An examination of convergent and discriminant validity included Pearson correlational analysis between the medication adherence full bank scores and other self-report measures. We expected moderate to high correlations (r ’s ≥ 0.36 – $0.67 =$ “moderate” and r ’s between 0.68 and $0.89 =$ “high”) between medication adherence and mental HRQOL (i.e., Neuro-QOL depression and Neuro-QOL anxiety scores, supporting convergent validity [74]. Conversely, we expected low correlations (r ’s $\leq 0.35 =$ “low”) between the medication adherence full bank score and two of the HEAL measures (patient–provider connection and health-care environment) to support discriminant validity [74].

Independent sample t -tests were used to examine known-groups validity. We compared (a) those at high risk for readmission (i.e., ≥ 2 inpatient admissions in the past six months) vs. those at low risk for readmission (no or one inpatient admission in the past six months) and (b) those with “worse” economic quality of life (Econ-QOL scores ≥ 60) vs. those with “better” economic quality of life (Econ-QOL scores ≤ 40). Specifically, we expected those at high risk for readmission to have lower scores on medication adherence than those at low risk for readmission, and we expected those with “worse” economic quality of life to have lower scores on medication adherence than those with “better” economic quality of life. Finally, we expected rates for impaired medication adherence scores (i.e., percentage of participants whose scores were > 1 SD worse than the sample mean of 50) to be higher than the normative population

(i.e., to exceed 16%) for those at high risk for readmission and those with worse economic problems [77].

Sample size requirements

We established our sample size requirements based on ensuring stable parameter estimation for both the constrained GRM modeling and the Wald-2 DIF analyses conducted while developing our item bank. A constrained GRM model is appropriate for sample sizes that are less than $N=500$ [78]; existing recommendations suggest that a minimum $N=200$ sample size is needed for stable parameter estimation with the constrained GRM [79]. Established recommendations also indicate that iterative Wald-2-based DIF analyses are appropriate when a minimum sample size of ~ 100 participants for each DIF subgroup investigated is available [80].

Results

Study participants

A total of 225 people with T2DM were included in these analyses; data come from a study focused on the development of new patient-reported outcomes that capture important social determinants of health, including the development of the Healthcare Access item bank [20] and the Illness Burden item bank [81], which are also published in this issue of *Quality of Life Research*. Table 1 presents a summary of the sample's descriptive characteristics. Briefly, participants were 57.7 ($SD=11.0$) years of age, on average. There was an approximately equal number of males (48%) and females (52%), and the majority of the sample was black (75%); 80% had a high school education or more.

Item bank development

Findings from the item bank development process are outlined in Table 2. To summarize, the initial pool contained 41 items. EFA analyses supported the unidimensionality of the items in the bank (i.e., the ratio of eigenvalue 1 to eigenvalue 2 was 5.5, and eigenvalue 1 accounted for 43.0% of modeled variance). Five items were flagged for deletion due to low item-adjusted total scores. CFA flagged one item for having a low factor loading, five items for excessive residual correlations, and three items for high model-able correlated error values. There were no items with sparse cells, nor did any items evidence problems with monotonicity. Finally, GRM analyses identified no items as having significant misfit (see Supplemental Table 1 for item fit chi-square values, degrees of freedom, p values, and chi-square / degrees of freedom quotients).

No items were flagged for impactful DIF for any of the factors investigated. The final Medication Adherence item bank included 27 items. The final CFA (Table 3) supported the unidimensionality of the final identified item set. Because our CFA overall model fit criteria were fully met, it was not necessary to conduct CBFA.

Supplemental Table 2 includes the final calibration estimates for the 27 items in the Medication Adherence item bank. The common slope value was 1.66; thresholds ranged from -4.05 to $+0.51$ for the full bank. Figure 1, which displays the test information function and standard errors plotted by theta, indicates that information was excellent (i.e., reliability ≥ 0.90) for theta scores between approximately -4.0 and $+0.3$ (scaled scores between 10 and 53). Figure 2, which displays the number of items administered by examinee plotted as a function of theta, illustrates the minimum and maximum number of items administered by the CAT: With CAT settings of minimum number of items = 4, maximum number of items = 12, and targeted score-level reliability = 0.85, CAT administration tended to use the fewest items (i.e., 4) from the item bank at approximately ≤ -3.5 SD units and the maximum number of items (i.e., 12) at approximately $\geq +0.1$ SD units.

The six items selected for inclusion in the short form are indicated by bolding in Supplemental Table 2. A conversion or “look-up” table for converting SF raw summed scores to T-scores is presented in Supplemental Table 3.

Preliminary psychometric analyses

Table 6 reports the internal consistency reliabilities for the CAT and SF: For both assessments, internal consistency was acceptable (i.e., ≥ 0.70). Table 4 also indicates that both administration formats were devoid of floor and ceiling effects. Table 5 presents the findings for convergent and discriminant validity. As hypothesized, there were moderate correlations between medication adherence scores and scores on Neuro-QOL Depression and Neuro-QOL Anxiety (i.e., $r = -0.54$ and -0.50 , for Neuro-QOL Depression and Neuro-QOL Anxiety, respectively). Correlations between medication adherence and HEAL Patient–Provider Connection and Healthcare Environment were borderline low (the relationship with HEAL Healthcare Environment met our established criterion, but the relationship with Patient–Provider Connection fell short of our criterion by some two hundredths of a point, i.e., 0.377 vs. the criterion upper limit of 0.350). Finally, Table 6 indicates that Medication Adherence full bank scores were able to differentiate between those with “better” versus “worse” economic quality of life and that those with “worse” economic quality of life had elevated risk for poor medication adherence.

Table 1 Descriptive data for the study sample

Variable	T2DM (N = 225)
Age (Years)	
M (SD)	57.7 (11.0)
Sex (%)	
Female	52.4
Male	47.6
Ethnicity (%)	
Not Hispanic or Latino	83.3
Hispanic or Latino	16.7
Race (%)	
White	17.8
Black/African American	75.1
Other	7.1
Education (%)	
Less than High School	4.9
Some High School (9 through < 12)	13.8
High School Graduate or Equivalent	32.4
Attended vocational, trade, or Business School after High School	7.6
College: < 2 years	15.6
College: Associate's Degree	5.3
College: 2 or > years, No Degree	6.2
College Graduate (B.S. or B.A.)	8.4
Postgraduate/No Degree	0.4
Postgraduate Degree (M.S., M.A., PhD, M.D., etc.)	5.3
Marital Status (%)	
Single, without partner	37.8
Single, with partner	16.9
Married	15.1
Separated/divorced	21.3
Widowed	8.4
Missing	0.4
How confident are you in filling out medical forms by yourself? (%)	
Extremely confident	46.7
Quite a bit confident	20.4
Somewhat confident	20.0
A little confident	8.0
Not at all confident	4.4
Missing	0.5
Do you usually ask someone to help you read materials you receive from the hospital or doctor?	
Yes	24.0
No	75.1
Missing	0.9
HbA1c	
Mean (SD)	8.1 (2.2)

Entries in the table represent percentage of participants unless otherwise specified

Discussion

The results reported in this manuscript supported the development of a new item bank, REDD-CAT Medication

Adherence, which was designed to assess self-reported knowledge and ability to take medications as prescribed among persons with T2DM. In addition, the psychometric properties for the full item bank, 6-item SF and simulated

Table 2 Unidimensional modeling and analyses for the REDD-CAT Medication Adherence item pool

Domain	Unidimensional modeling										Initial item performance		IRT modeling	
	Item pool	EFA E1/E2 ratio (crit- rion > 4)	Percent of variance for E1 (crit- rion > 40)	1-factor CFA loading (crit- rion < .50)	1-factor CFA residual cor- relation (crit- rion > .20)	1-factor CFA modification index (crit- rion > 100)	Item-adjusted total score correlations (Crite- rion < .40)	Sparse cells (crite- rion < 10)	Problems with monoto- nicity	IRT item misfit	DIF	IRT modeling		
Medication adherence	41 items	5.5	43.0	1 item	5 items	3 items	5 items	0 items	0 items	0 items	0 items	27 items		

CFA Confirmatory factor analysis, EFA Exploratory factor analysis, IRT Item response theory

Table 3 Final overall model fit and reliability characteristics for the REDD-CAT Medication Adherence item bank

Domain	Item bank	Chi-square (degrees of freedom); p value	CFI (criterion > .90)	TLI (criterion > .90)	CFA-based RMSEA (criterion < .10)	Alpha reliability (criterion > .80)	IRT-based RMSEA (criterion < .10)	Response pattern/ person reliability (criterion > .80)
Medication adherence	27 items	751.2 (324); < .001	.93	.93	.08	.91	.00	.89

Chi-square chi-square test of model fit, CFI Comparative fit index, TLI Tucker-Lewis index, RMSEA Root mean square error of approximation

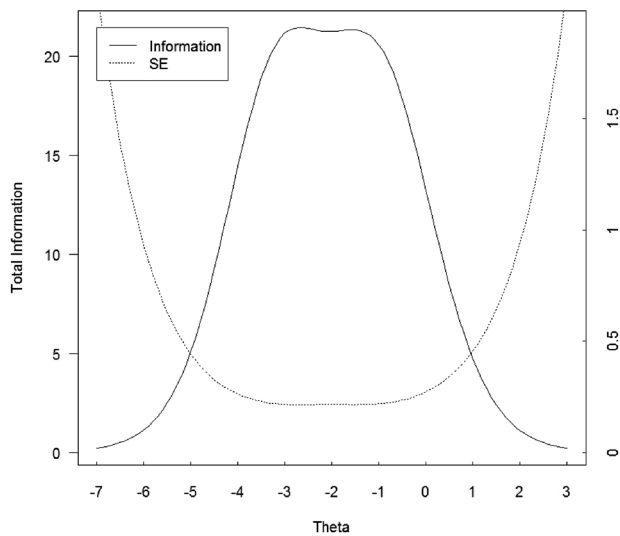


Fig. 1 Medication adherence item bank test information plot. In general, we would like total test information per score level to be ≥ 10.0 and the resultant standard error to be ≤ 0.32 (which provides a score-level reliability of ≥ 0.90). This figure shows excellent total test information (left y axis) and standard errors (right y axis) for medication adherence theta scores (x axis) between approximately -4.0 and $+0.3$ (scaled T-scores between approximately 10 and 53)

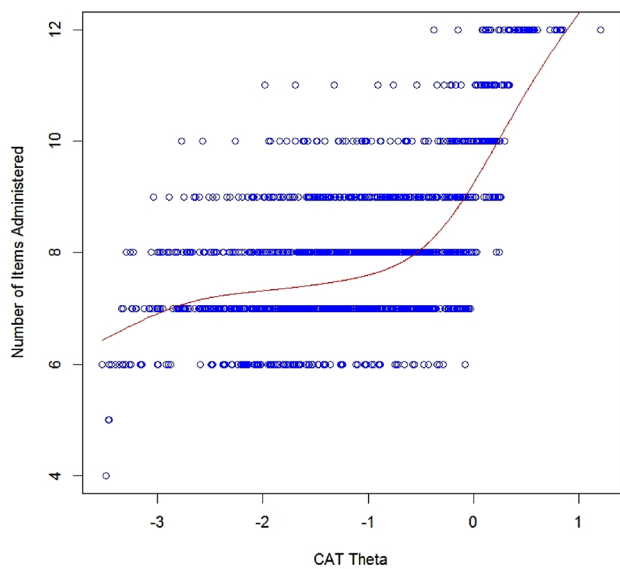


Fig. 2 Simulation data for the medication adherence number of CAT Items by CAT Theta. In this figure, the number of items administered by examinee (the individual blue circles) is plotted as a function (the red curvilinear line) of theta. The figure shows the number of CAT items used for different score levels in standard deviation units: At approximately ≤ -3.5 SD units, the CAT tended to use the minimum of four items from the item bank; at approximately $\geq +0.1$ SD units, the maximum of 12 items from the item bank was used by the CAT (Color figure online)

Table 4 Descriptive data for the different medication adherence administration formats

	<i>N</i>	Internal consistency reliability	Mean (SD)	% at floor	% at ceiling
Medication Adherence—CAT	225	0.91	49.9 (8.6)	0.0	0.0
Medication Adherence—SF	225	0.71	49.7 (8.3)	0.0	13.3

CAT Computer adaptive test, SF Short form

CAT administrations, were strong and support the reliability and validity of scores on this new measure. Specifically, this new measure is essentially unidimensional, its items are devoid of bias for age, gender, education, and socioeconomic status, internal consistency reliability was supported (with the CAT outperforming the SF in terms of magnitude of internal consistency reliability estimate), and scores for both formats were devoid of floor and ceiling effects, all of which serve these measures meeting established psychometric standards. This new measure will be publicly available through healthmeasures.net as a Neuro-QoL measure in early 2023, as well as through the PROMIS Application Programmable Interface (API).

There was also preliminary support for convergent validity (there were moderate relationships between Medication Adherence scores and the mental HRQOL measures). In addition, discriminant validity was supported by low correlations with HEAL Healthcare Environment; the relationship between medication adherence and HEAL Patient–Provider Connection fell just short of our criterion. Even with this one exception, 75% of our proposed convergent/discriminant analysis findings were in accordance with our a priori criteria, which meets established standards for establishing the construct validity of health status PROs [82]. Finally, known-groups validity was supported in that those individuals with better self-reported economic means had better medication adherence than those with worse economic means. Individuals with worse economic means were also more likely to have elevated risk for medication adherence problems, which is consistent with supporting a positive relationship between socioeconomic status and medication adherence [4, 5].

REDD-CAT Medication Adherence scores are on a T-score metric ($M = 50, SD = 10$), with lower scores indicating poor adherence; this increases the clinical utility of the measure, since obtained scores can immediately provide an

Table 5 Convergent and discriminant validity for the Medication Adherence item bank

Medication adherence	Convergent validity		Discriminant validity	
	Neuro-QOL depression	Neuro-QOL anxiety	Patient–provider connection	Healthcare environment
Full Bank Score	– 0.54	– 0.50	0.38	0.36
Mean (SD)	48.6 (8.22)	51.6 (9.72)	64.4 (9.6)	53.9 (9.8)

Table 6 Additional support for the validity of the REDD-CAT Medication Adherence item bank

Medication Adherence administration format	“Worse” economic quality of life			“Better” economic quality of life			<i>t</i>	<i>p</i>	<i>d</i>
	<i>n</i>	Mean (SD)	% Impaired ^a	<i>n</i>	Mean (SD)	% Impaired ^a			
Full bank score	63	45.8 (8.5)	25.4	16	59.3 (6.6)	0.00	– 5.92	<0.001	1.38

^aMedication Adherence score ≤ 40

estimation of an individual’s functioning relevant to a reference group (in this case, other people with T2DM). As such, persons with T2DM with Medication Adherence scores equal to or below 40 (i.e., ≤ 1 SD below the normative mean) have medication adherence problems that exceed 83.9% of persons with T2DM, whereas individuals with scores of less than or equal to 30 (i.e., ≤ 2 SDs below the normative mean) are exhibiting problems with medication adherence that are worse than 97.9% of their peers.

REDD-CAT Medication Adherence is the first self-reported PRO for medication adherence that can be administered as a CAT. CAT administration is such that only the most relevant items for a particular respondent are actually administered to that individual; in this type of “smart-test,” each item that is administered has been purposely selected based on the previous recorded item response. Given that the REDD-CAT measurement system is designed to provide a comprehensive assessment of important social determinants of behavior related to readmission risk, this type of administration assessment capitalizes on brevity while not sacrificing measurement sensitivity. In addition, for this new measure, the psychometric reliability of the CAT is superior to that of the SF administration, making it the preferred mode of administration. Regardless, the SF administration of this new measure has adequate psychometric properties and is still a viable option when practical considerations outweigh CAT administration (CAT administration requires the use of an electronic device, a reliable Wi-Fi connection, and access to a data collection platform along with access to the PROMIS API).

While this new self-report measure demonstrated adequate psychometric characteristics, future work is needed to mitigate several study limitations. First, this study did not assess MPR, nor did it include other self-reported medication adherence measures; thus, future work examining

the performance of this tool relative to other objective and subjective reports of medication adherence are needed to provide more robust support for convergence validity. Second, the study sample size, although it has been successfully applied in other PRO measurement development efforts [83, 84] and is appropriate for new measurement development using our study’s specified analyses (EFA, CFA, constrained GRM modeling, and DIF investigations), is relatively small. As such, future work with independent samples is needed to confirm these data; reliability and validity data should also be considered preliminary and warrant replication in an independent sample. Third, CAT analyses for this new measure utilized simulated scores and warrant replication in prospective data that are administered in a CAT format. In addition, this sample was recruited using an urban safety-net health system, which is likely to target individuals that have a higher probability of lacking health insurance and having a greater number of unmet social needs, and, as such, may not be generalizable to other types of hospital systems. Furthermore, future research is needed to examine responsiveness to change data, including establishing minimally important differences, to further contribute to the clinical utility of this new tool.

In sum, these preliminary data indicate that the new REDD-CAT Medication Adherence item bank is a psychometrically sound PRO that can be used to screen for problems with medication management in persons with T2DM; it has been accepted for inclusion within the Neuro-QoL measurement system and will be available through healthmeasures.net, as well as anywhere the PROMIS API is available, in early 2023. Furthermore, this PRO, as well as the other PROs that are included within the REDD-CAT measurement system, offer a comprehensive battery of measures that provide a sensitive assessment of the diverse social determinants of behavior related to readmission risk

in persons with T2DM. This new measure is the first CAT to assess medication adherence and is an excellent candidate for inclusion in studies that are focused on improving medication management of persons with T2DM. Finally, although this measure was developed specifically for use in persons with T2DM, it may also have clinical utility in other populations, especially those in which complex medication regimens are the norm (e.g., cardiovascular disease management, especially those individuals taking oral anticoagulant medications; or individuals on psychotropic medications for psychiatric conditions).

Conclusions

The REDD-CAT measurement system is the first comprehensive system designed to assess social determinants of health among persons with T2DM. The new REDD-CAT Medication Adherence measure captures an important social determinant of behavior, namely, patient self-reported medication adherence. This measure will be available for public use as a part of the healthmeasures.net platform. Additional research is needed to elucidate the mediation and moderation effects that self-reported medication adherence has on outcomes in these individuals. Understanding these relationships will help inform targeted interventions designed to minimize readmission risk and improve patient HRQOL.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11136-022-03275-z>.

Acknowledgements We thank the investigators and research associates/coordinators who worked on the study, the study participants, and organizations who supported recruitment efforts. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author contributions NEC: Principal Investigator; data coordination and analysis site; analysis design; initial draft of manuscript; incorporation of revisions. MAK: study Co-Investigator and Statistician; assistance with analysis design; primary statistician for measurement development portions of manuscript (e.g., factor analyses and IRT analyses); drafted statistical analysis section and a template for the results section. JPT: statistician and Data Analyst; primary statistician for reliability and validity analyses; assistance writing methods and results sections; review and feedback on manuscript drafts. AB: study Research Coordinator; responsible for data collection; review and feedback on manuscript drafts (critical review of the methods). JMH: study Grants Manager and Project Coordinator; review and feedback on manuscript drafts; assistance with study regulatory documents. IM: study Research Coordinator; responsible for data collection; review and feedback on manuscript drafts (critical review of the methods). JAM: study Data Manager; critical review of the methods; review and feedback on manuscript drafts. BWJ: Principal Investigator of the PRET study; review and feedback on manuscript drafts (critical review of the summary of the qualitative work that informed this study). SM: Principal Investigator; data collection site; study design; critical review of initial manuscript draft; review and feedback on manuscript drafts.

Funding Work on this manuscript was supported by Grant Number R21DK121092 (PIs Carlozzi; Mitchel [admin]) from the National Institutes of Health (NIH), National Institute of Diabetes and Digestive Kidney Diseases (NIDDK), and by the National Center for Advancing Translational Sciences (NCATS) for the Michigan Institute for Clinical and Health Research (UL1TR002240).

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. Dr. Mitchell has received compensation as a lecturer on Relationship Centered Care by Merck & co and holds equity in See Yourself Health Inc, a digital health company.

Ethical approval Approval was obtained from the ethics committee of Boston Medical Center. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate Informed consent was obtained from all participants prior to their participation in this study.

References

- Sapkota, S., Brien, J. A., Greenfield, J. R., & Aslani, P. (2015). A systematic review of interventions addressing adherence to anti-diabetic medications in Patients with type 2 diabetes-components of interventions. *PLoS ONE*, *10*(6), e0128581.
- Rubin, R. R. (2005). Adherence to pharmacologic therapy in patients with type 2 diabetes mellitus. *American Journal of Medicine*, *118*(Suppl 5A), 27S-34S.
- Krass, I., Schieback, P., & Dhippayom, T. (2015). Adherence to diabetes medication: A systematic review. *Diabetic Medicine*, *32*(6), 725–737.
- Roblin, D. W., Platt, R., Goodman, M. J., Hsu, J., Nelson, W. W., Smith, D. H., et al. (2005). Effect of increased cost-sharing on oral hypoglycemic use in five managed care organizations: How much is too much? *Medical Care*, *43*(10), 951–959.
- Gonzalez, J. S., Tanenbaum, M. L., & Commissariat, P. V. (2016). Psychosocial factors in medication adherence and diabetes self-management: Implications for research and practice. *American Psychologist*, *71*(7), 539–551.
- Clifford, S., Perez-Nieves, M., Skalicky, M., Reaney, M., & Coyne, K. S. (2014). A systematic literature review of methodologies used to assess medication adherence in patients with diabetes. *Current Medical Research and Opinion*, *30*(6), 1071–1085.
- Peterson, A. M., Nau, D. P., Cramer, J. A., Benner, J., Gwadrý-Sridhar, F., & Nichol, M. (2007). A checklist for medication compliance and persistence studies using retrospective databases. *Value Health*, *10*(1), 3–12.
- Moon, S. J., Lee, W. Y., Hwang, J. S., Hong, Y. P., & Morisky, D. E. (2017). Accuracy of a screening tool for medication adherence: A systematic review and meta-analysis of the morisky medication adherence scale-8. *PLoS ONE*, *12*(11), e0187139.
- Morisky, D. E., Ang, A., Krousel-Wood, M., & Ward, H. J. (2008). Predictive validity of a medication adherence measure in an outpatient setting. *Journal of Clinical Hypertension (Greenwich, Conn.)*, *10*(5), 348–354.
- Morisky, D. E., Green, L. W., & Levine, D. M. (1986). Concurrent and predictive validity of a self-reported measure of medication adherence. *Medical Care*, *24*(1), 67–74.

11. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists. Multivariate applications book series* (pp. xi–371). Erlbaum Associates.
12. Wu, M., Jen, T.-H., & Tam, H.P. (2016). *Educational Measurement for Applied Researchers: Theory into Practice*. Imprint: p. 1 online resource (XIV, 306 pages 114 illustrations, 74 illustrations in color. Springer
13. PROMIS® Instrument Development and Psychometric Evaluation Scientific Standards, 2019 http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf
14. Cella, D., Riley, W., Stone, N., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested in its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*, 1179–1194.
15. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, *45*(5 Suppl 1), S3–S11.
16. Cella, D., Lai, J. S., Nowinski, C., Victorson, D., Peterman, A., Miller, D., et al. (2012). Neuro-QOL: Brief measures of health-related quality of life for clinical research in neurology. *Neurology*, *78*, 1860–1867.
17. Cella, D., Nowinski, C., Peterman, A., Vistorson, D., Miller, D., Lai, J.-S., et al. (2011). The neurology quality of life measurement (Neuro-QOL) initiative. *Archives of Physical Medicine and Rehabilitation, Supplement*, *92*(Suppl 1), S28–S36.
18. Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4: Wide range achievement test professional manual*. Psychological Assessment Resources Inc.
19. Cancino, R., Esang, M., Martin, J., Morton, S., Kazis, L., & Jack, B. (2014). *Development of a Patient-Centered Instrument to Evaluate 30-Day Hospital Readmission*, in *Academy Health*. San Diego
20. Carlozzi, N.E., Kallen, M.A., Troost, J.P., Bragg, A., Marin-Howard, J., DE La Cruz, B., et al. (In Press). *Development and calibration data for the Healthcare Access Item Bank: A new computer adaptive test for persons with type 2 diabetes mellitus*. *Quality of Life Research*.
21. Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, *33*(8), 644–645.
22. Cella, D., Riley, W., Stone, N., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194.
23. Carlozzi, N. E., Boileau, N. R., Roche, M. W., Ready, R. E., Perlmutter, J. S., Chou, K. L., et al. (2020). Responsiveness to change over time and test-retest reliability of the PROMIS and Neuro-QoL mental health measures in persons with Huntington disease (HD). *Quality of Life Research*, *29*(12), 3419–3439.
24. Carlozzi, N. E., Goodnight, S., Kratz, A. L., Stout, J. C., McCormack, M. K., Paulsen, J. S., et al. (2019). Validation of neuro-QoL and PROMIS mental health patient reported outcome measures in persons with huntington disease. *J Huntingtons Dis*, *8*(4), 467–482.
25. Gershon, R. C., Lai, J.-S., Bode, R., Choi, S., Moy, C., Bleck, T., et al. (2012). Neuro-QOL: Quality of life item banks for adults with neurological disorders: Item development and calibrations based upon clinical and general population testing. *Quality of Life Research*, *21*(3), 475–486.
26. Miller, D. M., Bethoux, F., Victorson, D., Nowinski, C. J., Buono, S., Lai, J.-S., et al. (2015). Validating Neuro-QoL short forms and targeted scales with people who have multiple sclerosis. *Multiple Sclerosis*, *22*(6), 830–841.
27. Nowinski, C. J., Siderowf, A., Simuni, T., Wortman, C., Moy, C., & Cella, D. (2016). Neuro-QoL health-related quality of life measurement system: Validation in Parkinson’s disease. *Movement Disorders*, *31*(5), 725–733.
28. Victorson, D., Cavazos, J. E., Holmes, G. L., Reder, A. T., Wojna, V., Nowinski, C., et al. (2014). Validity of the neurology quality-of-life (Neuro-QoL) measurement system in adult epilepsy. *Epilepsy Behavior*, *31*, 77–84.
29. Healy, B. C., Zurawski, J., Gonzalez, C. T., Chitnis, T., Weiner, H. L., & Glanz, B. I. (2019). Assessment of computer adaptive testing version of the Neuro-QOL for people with multiple sclerosis. *Multiple Sclerosis*, *25*(13), 1791–1799.
30. Tran, T. A., Williams, L. M., Bui, D., Anthonisen, C., Poltavskiy, E., & Szabo, R. M. (2018). Prospective Pilot Study Comparing pre- and postsurgical CTSAQ and Neuro-QoL questionnaire with median nerve high-resolution ultrasound cross-sectional areas. *J Hand Surg Am*, *43*(2), 1841–1849.
31. Greco, C. M., Yu, L., Johnston, K. L., Dodds, N. E., Morone, N. E., Glick, R. M., et al. (2015). Measuring nonspecific factors in treatment: Item banks that assess the healthcare experience and attitudes from the patient’s perspective. *Quality of Life Research*, *25*(7), 1625–1634.
32. Tulskey, D. S., Kisala, P. A., Lai, J.-S., Carlozzi, N. E., Hammel, J., & Heinemann, A. W. (2015). Developing an item bank to measure economic quality of life for individuals with disabilities. *Archives of Physical Medicine and Rehabilitation*, *96*(4), 604–613.
33. Heinemann, A. W., Magasi, S., Hammel, J., Carlozzi, N. E., Garcia, S. F., Hahn, E. A., et al. (2015). Environmental factors item development for persons with stroke, traumatic brain injury, and spinal cord injury. *Archives of Physical Medicine and Rehabilitation*, *96*, 589–595.
34. Tulskey, D. S., Kisala, P. A., Lai, J.-S., Carlozzi, N. E., Hammel, J., & Heinemann, A. W. (2015). Developing an item bank to measure economic quality of life for individuals with disabilities. *Archives of Physical Medicine and Rehabilitation*, *96*, 604–613.
35. Sherer, M., Nick, T. G., Sander, A. M., Melguizo, M., Hanks, R., Novack, T., et al. (2017). Groupings of persons with traumatic brain injury: A new approach to classifying traumatic brain injury in the post-acute period. *The Journal of Head Trauma Rehabilitation*, *32*(2), 125–133.
36. Carlozzi, N.E., Graves, C., Troost, J.P., Kallen, M.A., Tulskey, D.S., Hanks, R., et al. (In Press). *Initial evidence for reliable and valid use of scores on the 8-item Econ-QOL short form to measure economic quality of life in caregivers of persons with traumatic brain injury*. *Archives of Physical Medicine & Rehabilitation Psychology*.
37. McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates Inc.
38. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT’s unidimensionality assumption. *Quality of Life Research*, *18*(4), 447–460.
39. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(Suppl 1), 19–31.
40. Ramsay, J. (2000). *TestGraf a program for the graphical analysis of multiple choice test and questionnaire data*. McGill University.
41. Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (2nd ed.). Guilford Press.
42. Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.
43. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new

- alternatives. *Structural Equation Modeling-a Multidisciplinary Journal*, 6(1), 1–55.
44. Hatcher, L. (1994). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. SAS Institute Inc.
 45. Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christoudolou, C., Gershon, R., et al. (2011). How item banks and its applications can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine & Rehabilitation*, 92(Suppl 1), S20–S27.
 46. Lai, J.-S., Zelko, F., Krull, K., Cella, D., Nowinski, C., Manley, P., et al. (2014). Parent-reported cognition of children with cancer and its potential clinical usefulness. *Quality of Life Research*, 23(4), 1049–1058.
 47. Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, 80(1), 26–44.
 48. Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research*, 24(3), 285–305.
 49. Luijben, T.C. & Boomsma, A. (1988). *Statistical guidance for model modification in covariance structure analysis*. Compstat (pp. 335–340)
 50. Saris, W.E., Satorra, A., & Sorbom, D. (1987). *The detection and correction of specification errors in structural equation models*, in *Sociological methodology* In: C.C. Clogg, (Eds) Jossey-Bass, San Francisco, (pp. 105–129)
 51. Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models for detection of misspecifications. *Structural Equation Modeling*, 16, 561–582.
 52. Muthén, L. K., & Muthén, B. O. (2011). *Mplus User's Guide*. Muthén & Muthén.
 53. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552–566.
 54. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73, 5–26.
 55. Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, 98(3), 223–237.
 56. Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. *Psychometrika*, 34(4p2):p. 1-&
 57. Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91(1), 25–39.
 58. Crisan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, 41, 439–455.
 59. Drasgow, F., Levine, M. V., Tsein, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19(2), 143–165.
 60. Zhao, Y. (2017). Impact of IRT item misfit on score estimates and severity classifications: An examination of PROMIS depression and pain interference item banks. *Quality of Life Research*, 26(3), 555–564.
 61. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(Suppl 1), 5–18.
 62. Cai, L., Thissen, D., & du Toit, S. H. C. (2015). *IRTPRO for Windows [Computer software]*. Scientific Software International.
 63. Woods, C. M., Cai, L., & Wang, M. (2013). The longer-improved wald test for DIF testing with multiple groups: evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73(3), 532–547.
 64. Wang, M., & Woods, C. M. (2017). Anchor selection using the wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, 41(1), 17–29.
 65. Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research*, 16(Suppl 1), 95–108.
 66. Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16(Suppl 1), 133–141.
 67. van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: theory and practice* (pp. 12–323). Kluwer Academic.
 68. Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandel, B., Klapp, B. F., et al. (2012). Computerized adaptive testing—ready for ambulatory monitoring? *Psychosomatic Medicine*, 74(4), 338–348.
 69. Smits, N., Paap, M. C. S., & Bohnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*, 27(4), 1055–1063.
 70. Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C. H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, 16(Suppl 1), 109–119.
 71. Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: a primer* (2nd ed., pp. xxiii–335). Lawrence Erlbaum Associates.
 72. Choi, S. W., Podrabsky, T., & McKinney, N. (2012). Firestar-D: Computerized adaptive testing simulation program for dichotomous item response theory models. *Applied Psychological Measurement*, 36(1), 67–68.
 73. Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the patient-reported outcomes measurement information system (PROMIS) in a three-month observational study. *Journal of Psychiatric Research*, 56, 112–119.
 74. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
 75. Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine & Rehabilitation*, 81(12 Suppl 2), S15–20.
 76. Cramer, D., & Howitt, D. L. (2004). *The Sage dictionary of statistics*. Sage.
 77. Heaton, R. K., Miller, S. W., Taylor, J. T., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Psychological Assessment Resources Inc.
 78. Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Psychometric Society.
 79. Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.
 80. Clauser, B. E., & Hambleton, R. K. (1994). Review of differential item functioning. *Journal of Educational Measurement*, 31(1), 88–92.
 81. Mitchell, S., Kallen, M.A., Troost, J.P., Bragg, A., Marin-Howard, J., Moldovan, I., et al. (In Press). *Development and calibration data for the Illness Burden Item Bank: A new computer adaptive*

- test for persons with type 2 diabetes mellitus*. Quality of Life Research.
82. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42.
 83. Carlozzi, N. E., Lange, R. T., French, L., Kallen, M. A., Boileau, N. R., Hanks, R., et al. (2020). TBI-CareQOL military health care frustration in caregivers of service members/veterans with traumatic brain injury. *Rehabilitation Psychology*, *65*(4), 360–376.
 84. Ruo, B., Choi, S. W., Baker, D. W., Grady, K. L., & Cella, D. (2010). Development and validation of a computer adaptive test

for measuring dyspnea in heart failure. *Journal of Cardiac Failure*, *16*(8), 659–668.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.