



Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 3742

Corresponding Measures:

Measure Title: ESRD Dialysis Patient Life Goals Survey (PaLS)

Measure Steward: Centers for Medicare & Medicaid Services

sp.02. Brief Description of Measure:

The PaLS is a patient self-report survey that includes eight items related to dialysis facility care team discussions about patient life goals. Six of the items are Likert-type items that are used to generate a "quality of facility care team discussion" score (described below). The remaining two items on the PaLS are checklist items: (1) a list of patient-reported life goals; and (2) a patient-reported list of dialysis care team members that the patient reports has talked with them about their life goals. These items are not scored. Instead, these items serve to provide contextual information for both the patient and the facility to guide care team discussions.

The PaLS is used to generate a patient-level *t*-score that reflects patient-reported satisfaction with how well his/her/their facility is doing in discussing life goals with the patient as part of the treatment planning process. For each individual patient at a given facility, the calculated *t*-score ($M=50$; $SD=10$) represents a patient's perceptions of their satisfaction with their dialysis care team discussions about life goals. A *t*-score greater than 40 and less than 60 reflects a score that is within normal limits of existing practices and should not warrant further action, assuming typical existing practices for patient and care team discussions are deemed adequate. Assuming a normal distribution, scores that are ≤ 40 would warrant follow-up by the facility. Specifically, scores ≤ 40 (i.e., ≥ 1 *SD* below the mean) suggest patient perceptions of care discussions are worse than 84% of their peers, whereas scores ≤ 30 (i.e., ≥ 2 *SDs* below the mean) suggest patient perceptions of care discussions are worse than 98% of their peers; scores that are > 40 would be within "normal limits" (Heaton et al., 2004). The *t*-score is based on the data collected for the instrument testing, as described in the section on scientific acceptability, but is currently not part of the process measure calculation.

The target population for the measure is patients on chronic dialysis who meet all of the following criteria:

- Are at least 18 years old
- Completed the PaLS survey at least once during the one-year reporting period

Reference:

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*. Lutz, FL: Psychological Assessment Resources, Inc.

1b.01. Developer Rationale:

sp.12. Numerator Statement: The numerator is the number of eligible patients from the denominator that completed at least one scorable item of the PaLS (i.e., at least one of the six Likert-type items).

sp.14. Denominator Statement:

All prevalent adult chronic dialysis patients (≥ 18 y/o) treated by the facility (both In-Center and Home Dialysis) for greater than 90 days during the reporting period, who read and understand English*.

*At present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

sp.16. Denominator Exclusions:

Exclusions are implicit based on eligibility criteria to complete the survey. These include:

- Persons under age 18
- Persons who are kidney transplant recipients with a functioning allograft
- Persons who had previously been on chronic dialysis but have recovered renal function, or are lost to follow up during the reporting period
- Persons with duplicate surveys – we used either the first or the more complete survey
- Persons that are unable to read and/or understand English (self-assessed and self-reported)*

*At present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

Measure Type: Process

sp.28. Data Source:

Claims

Instrument-Based Data

Registry Data

sp.07. Level of Analysis:

Other

IF Endorsement Maintenance – Original Endorsement Date:

Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?:

1. Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:

Current Submission:

Updated evidence information here.

Previous (Year) Submission:

Evidence from the previous submission here.

1a.01. Provide a logic model.

Briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

[Response Begins]

[Response Ends]

1a.02. Select the type of source for the systematic review of the body of evidence that supports the performance measure.

A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data.

[Response Begins]

[Response Ends]

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

Evidence - Systematic Reviews Table (Repeatable)

Group 1 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]

[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]

[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]

[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]

[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]

[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]

[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]

[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]

[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]

[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]

[Response Ends]

1a.13. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.

[Response Begins]

[Response Ends]

1a.14. Briefly synthesize the evidence that supports the measure.

[Response Begins]

[Response Ends]

1a.15. Detail the process used to identify the evidence.

[Response Begins]

[Response Ends]

1a.16. Provide the citation(s) for the evidence.

[Response Begins]

[Response Ends]

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]

[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.

[Response Begins]

[Response Ends]

1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., “topped out”, disparities data may demonstrate an

opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in above.

[Response Begins]

[Response Ends]

2. Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see [What Good Looks Like](#)).

[Response Begins]

ESRD Dialysis Patient Life Goals Survey (PaLS)

[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]

The PaLS is a patient self-report survey that includes eight items related to dialysis facility care team discussions about patient life goals. Six of the items are Likert-type items that are used to generate a “quality of facility care team discussion” score (described below). The remaining two items on the PaLS are checklist items: (1) a list of patient-reported life goals; and (2) a patient-reported list of dialysis care team members that the patient reports has talked with them about their life goals. These items are not scored. Instead, these items serve to provide contextual information for both the patient and the facility to guide care team discussions.

The PaLS is used to generate a patient-level *t*-score that reflects patient-reported satisfaction with how well his/her/their facility is doing in discussing life goals with the patient as part of the treatment planning process. For each individual patient at a given facility, the calculated *t*-score ($M=50$; $SD=10$) represents a patient’s perceptions of their satisfaction with their dialysis care team discussions about life goals. A *t*-score greater than 40 and less than 60 reflects a score that is within normal limits of existing practices and should not warrant further action, assuming typical existing practices for patient and care team discussions are deemed adequate. Assuming a normal distribution, scores that are ≤ 40 would warrant follow-up by the facility. Specifically, scores ≤ 40 (i.e., ≥ 1 *SD* below the mean) suggest patient perceptions of care discussions are worse than 84% of their peers, whereas scores ≤ 30 (i.e., ≥ 2 *SDs* below the mean) suggest patient perceptions of care discussions are worse than 98% of their peers; scores that are > 40 would be within “normal limits” (Heaton et al., 2004). The *t*-score is based on the data collected for the instrument testing, as described in the section on scientific acceptability, but is currently not part of the process measure calculation.

The target population for the measure is patients on chronic dialysis who meet all of the following criteria:

- Are at least 18 years old
- Completed the PaLS survey at least once during the one-year reporting period

Reference:

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*. Lutz, FL: Psychological Assessment Resources, Inc.

[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Surgery: General*

[Response Begins]

Renal

Renal: End Stage Renal Disease (ESRD)

[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]

Person-and Family-Centered Care: Person-and Family-Centered Care

[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]

Adults (Age >= 18)

Elderly (Age >= 65)

Populations at Risk: Dual eligible beneficiaries of Medicare and Medicaid

Populations at Risk: Individuals with multiple chronic conditions

[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]

Other

[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.

[Response Begins]

Outpatient Services

[Response Ends]

sp.09. Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials.

Do not enter a URL linking to a home page or to general information. If no URL is available, indicate "none available".

[Response Begins]

None available.

[Response Ends]

sp.12. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, [contact staff](#). Provide descriptors for any codes. Use one file with multiple worksheets, if needed.

[Response Begins]

No data dictionary/code table – all information provided in the submission form

[Response Ends]

sp.13. State the numerator.

Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome).

DO NOT include the rationale for the measure.

[Response Begins]

The numerator is the number of eligible patients from the denominator that completed at least one scorable item of the PaLS (i.e., at least one of the six Likert-type items).

[Response Ends]

sp.14. Provide details needed to calculate the numerator.

All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets.

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

We begin with the number of patients that took the PaLS survey and completed at least one of the six Likert-type scorable PaLS items that comprise the “quality of facility care team discussions” score. The response options for these six items are scored from 1 to 5. Higher scores indicate greater overall patient reported agreement that the care team is asking about and discussing life goals with the patient. IRT scores are initially estimated on the theta metric ($M=0$; $SD=1$). In order to enhance the clinical utility of our PaLS measure, we converted theta scores to standardized scores on the t -score metric ($M=50$; $SD=10$). The conversion from a theta score to a t -score can be made using the following linear transformation: $t\text{-score}=(\text{theta} \times 10)+50$. This patient-level t -score represents a patient’s perceptions about how well the facility is doing in discussing life goals as part of the treatment planning process.

Although missing PaLS responses are allowed, patients must answer at least one of the six Likert-type scorable PaLS items to receive a t -score.

The t -score is based on the data collected for the instrument testing, as described in the scientific acceptability, but is currently not part of the process measure calculation.

The numerator is comprised of the number of eligible patients from the denominator who completed at least one Likert-type scorable item of the PaLS.

[Response Ends]

sp.15. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]

All prevalent adult chronic dialysis patients (≥ 18 y/o) treated by the facility (both In-Center and Home Dialysis) for greater than 90 days during the reporting period, who read and understand English*.

*At present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

[Response Ends]

sp.16. Provide details needed to calculate the denominator.

All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets.

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

To be in the denominator, chronic dialysis patients at the facility must be eligible to complete the PaLS; that is, they must be (a) at least 18 years of age; (b) receiving long-term dialysis in the United States or any U.S. Territory

for greater than 90 days during the reporting period; and (c) able to read and understand English (self-assessed and reported). Receiving long-term dialysis in the 90 day period was selected in order to allow time for the patient to stabilize after beginning chronic dialysis, and for the dialysis care team to initiate discussions about patient life goals as part of the treatment planning process. This 90 day period also reduces facility-related burden. At present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

To construct our denominator for testing, we used the following self-report data from survey participants: first name, last name, sex, birthdate, last four digits of their social security number (SSN), race, ethnicity, and level of education completed. The first four of these data elements were required; the last four elements participants could elect to not report. Using self-reported first name, last name, last four digits of SSN (if provided), and birthdate, participants were then matched to our ESRD database, which contains treatment history data on all U.S. ESRD patients. We used CMS administrative data to confirm dialysis modality for participants linked to the UM-KECC ESRD database (in-center hemodialysis, home hemodialysis, peritoneal dialysis, or kidney transplant). In some cases, we could not match participants to their data in the UM-KECC ESRD database (i.e., if self-reported first or last name, birthdate, sex, or last four SSN digits were either missing, illegible or incomplete). In these cases, participants were not included in the analysis using dialysis modality.

We implemented two different field-testing data collection efforts as part of our measurement development process, which we refer to hereafter as: 1) the calibration sample; and 2) the validation testing sample. For the calibration sample, 10.4% of participants were not able to be matched to the ESRD database. For the validation testing sample, 20.2% of participants were not able to be matched to the ESRD database.

[Response Ends]

sp.17. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]

Exclusions are implicit based on eligibility criteria to complete the survey. These include:

- Persons under age 18
- Persons who are kidney transplant recipients with a functioning allograft
- Persons who had previously been on chronic dialysis but have recovered renal function, or are lost to follow up during the reporting period
- Persons with duplicate surveys – we used either the first or the more complete survey
- Persons that are unable to read and/or understand English (self-assessed and self-reported)*

*At present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

[Response Ends]

sp.18. Provide details needed to calculate the denominator exclusions.

All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

To be in the denominator, chronic dialysis patients at the facility must be eligible to complete the PaLS; that is, they must be (a) at least 18 years of age; (b) receiving long-term dialysis in the United States or any U.S. Territory for greater than 90 days during the reporting period; (c) able to read and understand English (self-assessed and reported). Receiving long-term dialysis in the 90 day period was selected in order to allow time for the patient to stabilize after beginning chronic dialysis, and for the dialysis care team to initiate discussions about patient life goals as part of the treatment planning process. This 90 day period also reduces facility-related burden.

Again, at present, this instrument is available to patients who read and understand English. Generalizing the survey to other languages will require additional development work.

For our testing (see sp.15, above) we used CMS administrative data to confirm patients were ESRD and on a chronic dialysis modality.

Exclusions are implicit based on eligibility criteria to complete the survey. These include age less than 18; patient has a kidney transplant; patient with recovered renal function, or lost to follow up; and unable to read and/or understand English (whether self-assessed or self-reported). In our testing we also excluded duplicate patient surveys.

[Response Ends]

sp.19. Provide all information required to stratify the measure results, if necessary.

Include the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate. Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format in the Data Dictionary field.

[Response Begins]

N/A

[Response Ends]

sp.20. Is this measure adjusted for socioeconomic status (SES)?

[Response Begins]

No

[Response Ends]

sp.21. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.

[Response Begins]

No risk adjustment or risk stratification

[Response Ends]

sp.22. Select the most relevant type of score.

Attachment: If available, please provide a sample report.

[Response Begins]

Rate/proportion

[Response Ends]

sp.23. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score

[Response Begins]

Better quality = Higher score

[Response Ends]

sp.24. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]

The response options for the six Likert-type scorable PaLS items range from “strongly agree” (5) to “strongly disagree” (1) for three items, and from “always” (5) to “never” (1) for the other three items. Response pattern scoring was applied to item responses, using the measure’s established item parameters.

See attached flowchart.

[Response Ends]

sp.25. Attach a copy of the instrument (e.g. survey, tool, questionnaire, scale) used as a data source for your measure, if available.

[Response Begins]

Copy of instrument is attached.

[Response Ends]

Attachment: 3742_Patient Life Goals Survey_508.pdf

sp.26. Indicate the responder for your instrument.

[Response Begins]

Patient

[Response Ends]

sp.27. If measure testing is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

Examples of samples used for testing:

- *Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.*
- *The sample should represent the variety of entities whose performance will be measured. The [2010 Measure Testing Task Force](#) recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.*

- *The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.*
- *When possible, units of measurement and patients within units should be randomly selected.*

[Response Begins]

Snowball sampling was used to obtain the calibration and validation testing samples. Participants received materials via web link or paper survey packets. Sample size requirements were based on IRT-related analyses (i.e., graded response model [GRM] analyses and differential item functioning [DIF] analyses). Sample size requirements for use of GRM analyses have been estimated to be between 200 and 1000, with larger sample sizes producing more stable parameter estimates (Muraki, 1990; Samejima, 1969; Samejima et al., 1996). For DIF analyses using lordif, a sample size should be at least n=200 participants per DIF factor sub-group of interest (Clauser & Hambleton, 1994).

References:

Clauser, B.E., & Hambleton, R.K. (1994). Review of differential item functioning. *Journal of Educational Measurement*, 31(1), 88-92.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14(1), 59-71.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society.

Samejima, F., van der Liden, W.J., & Hambleton, R. (1996). The graded response model. In: W.J. van der Liden (Ed.), *Handbook of modern item response theory* (pp. 85-100). Springer.

[Response Ends]

sp.28. Identify whether and how proxy responses are allowed.

[Response Begins]

Proxy responses are not allowed.

[Response Ends]

sp.29. Survey/Patient-reported data.

Provide instructions for data collection and guidance on minimum response rate. Specify calculation of response rates to be reported with performance measure results.

[Response Begins]

ESRD patients needed to answer at least one of the six Likert-type scorable PaLS items to receive a patient-level score. We were not able to calculate a facility response rate, given that data collection and testing were performed at the patient-level. Prior to possible implementation at the dialysis facility level, the response rate will need to be calculated at the facility level. The facility-level response rate should be calculated by dividing the number of patients who answer at least one scorable PaLS item (i.e., one of the six Likert-type items from the PaLS) by the number of patients who are eligible for the survey and complete its required demographic items.

[Response Ends]

sp.30. Select only the data sources for which the measure is specified.

[Response Begins]

Claims

Instrument-Based Data

Registry Data

[Response Ends]

sp.31. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]

To identify ESRD patients, we used the following self-report data from survey participants: first name, last name, sex, birthdate, last four digits of their social security number (SSN), race, ethnicity, and level of education completed. The first four data elements were required; the last four elements participants could elect to not report. Using self-reported first name, last name, last four digits of SSN (if provided), and birthdate, participants were then matched to our own ESRD database, which contains treatment history data on all U.S. ESRD patients. We used CMS administrative data to confirm dialysis modality for participants linked to our database (in-center hemodialysis, home hemodialysis, peritoneal dialysis, or kidney transplant). In some cases, we could not match participants to their data in the ESRD database (if self-reported first or last name, birthdate, sex, or last four SSN digits were missing, illegible or incomplete). In these cases, participants were not included in the analysis using dialysis modality.

ESRD data were used to confirm self-reported ESRD status and treatment information. These data were derived from the UM-KECC database, a national ESRD patient database that includes information from the Renal Management Information System (REMIS), CROWNWeb facility-reported clinical and administrative data (including CMS-2728 Medical Evidence Form, CMS-2746 Death Notification Form, and CMS-2744 Annual Facility Survey Form and patient tracking data), the Medicare Enrollment Database (EDB), and Medicare dialysis claims data (primarily outpatient). In addition, the UM-KECC database includes transplant data from the Scientific Registry of Transplant Recipients (SRTR), data from the Nursing Home Minimum Dataset, data from the Quality Improvement Evaluation System (QIES) Business Intelligence Center (QBIC; which includes Provider and Survey and Certification data from Automated Survey Processing Environment [ASPEN]), and data from the Dialysis Facility Care Compare (DFCC).

[Response Ends]

sp.32. Provide the data collection instrument.

[Response Begins]

Available in attached appendix in Question 1 of the Additional Section

[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

- Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
- All required sections must be completed.
- For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
- An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
- Contact NQF staff with any questions. Check for resources at the [Submitting Standards webpage](#).
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the [2021 Measure Evaluation Criteria and Guidance](#).

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For instrument-based measures (including PRO-PMs) and composite performance measures, reliability should be demonstrated for the computed performance score.

2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure;

AND

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 14,15 and has demonstrated adequate discrimination and calibration
- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 differences in performance;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions

Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measure scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

Risk factors that influence outcomes should not be specified as exclusions.

With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of \$25 in cost for an episode of care (e.g., \$5,000 v. \$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:

Current Submission:

Updated testing information here.

Previous (Year) Submission:

Testing from the previous submission here.

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]

Claims

Instrument-Based Data

Registry Data

[Response Ends]

2a.02. If an existing dataset was used, identify the specific dataset.

The dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

[Response Begins]

ESRD data were used to confirm self-reported ESRD status and treatment information. These data were derived from the UM-KECC database, a national ESRD patient database that includes information from the Renal Management Information System (REMIS), CROWNWeb facility-reported clinical and administrative data (including CMS-2728 Medical Evidence Form, CMS-2746 Death Notification Form, and CMS-2744 Annual Facility Survey Form and patient tracking data), the Medicare Enrollment Database (EDB), and Medicare dialysis claims data (primarily outpatient). In addition, the UM-KECC database includes transplant data from the Scientific Registry of Transplant Recipients (SRTR), data from the Nursing Home Minimum Dataset, data from the Quality Improvement Evaluation System (QIES) Business Intelligence Center (QBIC; which includes Provider and Survey and Certification data from Automated Survey Processing Environment [ASPEN]), and data from the Dialysis Facility Care Compare (DFCC).

[Response Ends]

2a.03. Provide the dates of the data used in testing.

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]

Calibration sample:

The calibration sample was collected: 06-03-2020 – 12-29-2020. This is the primary development sample for the PaLS measure and was the sample that was used to inform a patient-level *t*-score (based on responses to the six Likert-type scorable PaLS items).

Validation testing sample:

The validation sample was derived from data collected: 04-07-2021 – 04-24-2022. These data were used to confirm the reliability and validity of a patient-level *t*-score (based on responses to the six Likert-type scorable PaLS items).

[Response Ends]

2a.04. Select the levels of analysis for which the measure is tested.

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]

Other (specify)

[Other (specify) Please Explain]

Patient level

[Response Ends]

2a.05. List the measured entities included in the testing and analysis (by level of analysis and data source).

Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample.

[Response Begins]

Calibration sample:

Five-hundred and seventeen (N=517) participants were included in the national field-testing data collection that established the calibration sample. This was a cross-sectional sample and did not include longitudinal follow-up. Participants were recruited using a snowball sampling approach in which recruitment materials were sent via email to several dialysis organizations, nephrology professional organizations, and kidney patient advocacy groups. We asked these stakeholders to disseminate recruitment materials to ESRD patients affiliated with their clinics, patient organizations, or via their own professional or advocacy networks. We only included surveys that met participation eligibility criteria and where participant consent was obtained. Upon request, we provided paper surveys to participants.

Validation testing sample:

Data collection was longitudinal and included baseline, 3-month, and 6-month follow-up study assessments. Four-hundred and twenty (N=420) participants were included in the national data collection that established the validation testing sample at baseline. One hundred and eighty-three (n=183) participants completed the 3-month follow up and one hundred and sixty-seven (n=167) participants completed the 6-month follow up. At each time point, participants completed the PaLS as well as additional surveys that were used to examine different aspects of psychometric validity (e.g., see section 2b.03: Known-groups validity for PROMIS measures). Participants were recruited using a snowball sampling approach in which recruitment materials were sent via email to several dialysis organizations, nephrology professional organizations, and kidney patient advocacy groups. We asked these stakeholders to disseminate recruitment materials to ESRD patients affiliated with their clinics, patient organizations, or via their own professional or advocacy networks. We only included surveys that met participant eligibility criteria and where participant consent was obtained. Upon request, we provided paper surveys to participants.

[Response Ends]

2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]

Calibration sample:

Only participants that met eligibility criteria and provided consent were included in this sample. There were 517 participants in the calibration sample. The average age of participants was 61.6 years old; 47.4% of participants

were female; 70.4% were White, 18.1% were Black, 6.8% were Other/Multi-racial/Unknown/Missing and 4.6% did not wish to report race; 9.3% were Hispanic; and 42.4% reported 4 years or more of college.

Validation testing sample:

Only participants that met eligibility criteria and provided consent were included in this sample. There were 420 participants in the validation testing sample at baseline. The average age of participants was 59.6 years old; 56.7% of participants were female; 69.8% were White, 21.7% were Black, 6.0% were Other/Multi-racial/Unknown/Missing and 2.6% did not wish to report race; 7.6% were Hispanic; and 38.3% reported 4 years or more of college.

At 3-month follow up there were 183 participants. The average age of participants was 60.8 years old; 51.4% of participants were female; 73.2% were White, 20.2% were Black, 3.8% were Other/Multi-racial/Unknown/Missing and 2.7% did not wish to report race; 6.6% were Hispanic; and 43.2% reported 4 years or more of college.

At 6-month follow up there were 167 participants. The average age of participants was 61.9 years old; 55.1% of participants were female; 74.9% were White, 18.9% were Black, 3.6% were Other/Multi-racial/Unknown/Missing and 3.0% did not wish to report race; 6.0% were Hispanic; and 41.9% reported 4 years or more of college.

[Response Ends]

2a.07. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing.

[Response Begins]

These two samples were obtained at different time points as noted above.

Calibration sample:

The calibration sample was derived from the field-testing data collected: 06-03-2020 – 12-29-2020. This is the primary development sample for the PaLS measure, which used classical test theory and item response theory analytical approaches. This was the sample that was used to estimate item parameters for calculating a patient-level *t*-score using responses to the six Likert-type scorable PaLS items. This sample was also used to generate preliminary validity data for the PaLS measure.

Validation testing sample:

The validation testing sample was derived from data collected: 04-07-2021 – 04-24-2022. These data were used to validate the patient-level *t*-scores (derived from responses on the six Likert-type scorable PaLS items). Baseline and follow-up data were collected for this sample, along with administration of additional surveys to examine different aspects of psychometric reliability and validity of the patient-level PaLS *t*-score (e.g., see section 2b.03: Known-groups validity for PROMIS measures).

[Response Ends]

2a.08. List the social risk factors that were available and analyzed.

For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

[Response Begins]

Calibration sample and Validation testing sample:

Age, Education, Sex, and Race were examined for item bias (using DIF analyses), and dual eligibility status was used to examine score disparities. Social risk factors were analyzed in both the calibration and validation testing samples. Social risk factors were assessed for potential sources of disparities in section 1b.04.

[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.09 check patient or encounter-level data; in 2a.010 enter “see validity testing section of data elements”; and enter “N/A” for 2a.11 and 2a.12.

2a.09. Select the level of reliability testing conducted.

Choose one or both levels.

[Response Begins]

Patient or Encounter-Level (e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)

[Response Ends]

2a.10. For each level of reliability testing checked above, describe the method of reliability testing and what it tests.

Describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used.

[Response Begins]

Calibration sample

The PaLS was developed according to established PRO development methodology (*PROMIS® Instrument Development and Psychometric Evaluation Scientific Standards*, 2019). This process relies on classical test theory (CTT) and item response theory (IRT) analyses. Specifically, we used exploratory and confirmatory factor analyses (EFA, CFA) to ensure that the six Likert-type scorable PaLS items were unidimensional (a condition for generating an IRT-based *t*-score; Cook et al., 2009; McDonald, 1999; Reise et al., 2007). For EFA, we considered the item set to have unidimensional characteristics if the ratio of eigenvalue 1 to eigenvalue 2 was ≥ 4 and the proportion of variance accounted for by eigenvalue 1 was ≥ 0.40 (Hu & Bentler, 1999; Kline, 2005; Lai et al., 2006, 2011).

We excluded items with sparse cells (response categories with $n < 5$ participants), items with low item-adjusted total score correlations (< 0.40), and items that were non-monotonic (monotonicity was examined using non-parametric IRT models of item-rest plots and expected score by latent trait plots; Testgraf Software; Ramsay, 2000).

For CFA, we considered an item set to be unidimensional if: the comparative fit index (CFI) was ≥ 0.90 , the Tucker-Lewis index (TLI) was ≥ 0.90 , and the root mean square error of approximation (RMSEA) was < 0.10 (Cook et al., 2009; Hu & Bentler, 1999; Kline, 2005; Bentler, 1990; Lai et al., 2011, 2014). For comparative fit purposes, we also obtained the chi-square value for model fit and its associated *p* value. We considered items for removal if they had low factor loadings (< 0.50) or were locally dependent (i.e., residual correlation > 0.20 ; correlated error modification index ≥ 100 ; Cook et al., 2009; Hair et al., 2009; Kaplan, 1989; Luijben & Boomsma, 1988; McDonald, 1999; Reise et al., 2007; Saris et al., 1987, 2009; Whittaker, 2012).

Next, a graded response model (GRM) was used to estimate item parameters. We excluded items with significant misfit ($S-X^2 / df$ effect size > 3 ; Crisan et al., 2017; Drasgow et al., 2017; Stark et al., 2006; Zhao, 2017). We also excluded items with impactful differential item functioning (DIF). Items were evaluated for DIF using the lordif R package, version 0.3-3 (Choi et al., 2011). This statistical package iteratively applies a hybrid logistic ordinal

regression (LOR) and IRT approach. Items with McFadden pseudo-R² change ≥ 0.02 were flagged for DIF. DIF analyses were conducted for age (median split), education (4 year college degree or more versus less than a 4 year college degree), sex (male versus female), race (white versus other) and modality (in-center versus at home).

We investigated the practical impact of flagged DIF items by quantifying change in individual scores when adjusted for DIF. Two scores were calculated for each patient, one based on item parameters calibrated for the entire sample and another in which items flagged for DIF were calibrated separately (DIF-adjusted). For example, items flagged for DIF based on sex would have parameters calibrated separately in males and females. Scoring impact was evaluated based on: (a) Pearson correlation, (b) mean difference, (c) root mean squared difference (RMSD), and (d) percentage of score differences > their associated unadjusted score standard error (SE), i.e., >2% of DIF-corrected vs. uncorrected score differences exceeding individual case uncorrected score standard errors (Edelen et al., 2007).

These IRT-based analyses were followed by a final CFA analysis designed to confirm that the final item set remained essentially unidimensional (using the same item-level and overall model fit criteria outlined above).

Following this process, we converted the theta-based score to *t*-score. An IRT score is initially on a theta metric, with a mean of 0 and a SD of 1. We converted the theta score to standardized scores on the *t*-score metric ($M=50$; $SD=10$). The conversion from a theta score to a *t*-score can be obtained using the following linear transformation: $t\text{-score} = (\text{theta} \times 10) + 50$.

The *t*-score was then used to examine the preliminary reliability and validity of the PaLS *t*-score. Specific methods are described below for each level of testing.

Cronbach's alpha was used to establish the internal consistency of the six Likert-type scorable PaLS items.

Cronbach's alpha is defined below, where N is the number of items, c is the average inter-item covariance among items and v is the average variance (Cronbach, 1951).

$$\alpha = \frac{Nc}{v + (N - 1)c}$$

Marginal reliability is calculated as the ratio of the true score variance to the total variance, expressed with respect to the estimated latent abilities. Marginal reliability refers to the reliability with regard to the population as a whole. Marginal reliability is defined below with a density of g and variance of 1 (Andersson & Xin, 2018).

$$\rho_{\theta}(\alpha) = \int_{-\infty}^{\infty} \frac{I(\theta; \alpha)}{I(\theta; \alpha) + 1} g(\theta) d\theta$$

Response pattern reliability is the reliability based on the median *t*-score standard error and *t*-score standard deviation for *t*-scores ± 3 standard deviations from the mean. Thus, response pattern reliability was calculated using the median *t*-score standard error and *t*-score standard deviation where *t*-scores were between 20 and 80. Participants were not excluded if they did not answer all six Likert-type scorable PaLS items. Response pattern reliability is defined below (Pilkonis et al., 2014).

$$Reliability = 1 - \frac{SE^2}{SD^2}$$

Validation testing sample

Cronbach's alpha, marginal reliability and response pattern reliability were also examined for the patient-level *t*-score in the validation testing sample using the methods outlined in the calibration sample description above.

In addition, we calculated test-retest reliability using intraclass correlation coefficients (ICC). Minimum acceptable criteria for test-retest reliability was set at ≥ 0.70 for intraclass correlations (Cohen, 1969). ICC was calculated using the SAS macro intracc.sas.

We estimated two forms of the ICC, one including both systematic and random error in the estimation denominator and one including only random error. We used a two-way mixed effects ICC model, where people

effects were randomized and measure effects were fixed. We report our test-retest reliability estimates based on our ICC results from systematic plus random error-based estimations.

Minimal detectable change (DC_{95}) and standard error of measurement (SEM) were calculated for the PaLS scores. Minimal detectable change identifies the amount of change that can be detected with 95% confidence that it is not due to measurement error from baseline to the time point of interest, in this case 3 months and 6 months. The SEM percentage is calculated by dividing the SEM by the mean of all observations across time points and multiplying by 100. <10% indicates good (i.e. acceptable) measurement error.

Minimal detectable change (DC_{95}) is calculated using the equation:

$$DC_{95} = SEM * 1.96 * \sqrt{2}$$

Standard Error of Measurement (SEM) is calculated using the equation:

$$SEM = 1.96 * \sqrt{1 - ICC}$$

References:

- Andersson, B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational and psychological measurement*, 78(1), 32–45.
<https://doi.org/10.1177/0013164417713570>
- Bentler, P.M. (1990). *Comparative fit indexes in structural models*. *Psychological Bulletin*, 107(2), 238–246.
- Choi, S.W., Gibbons, L.E., & Crane, P.K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of statistical software*, 39(8), 1–30. <https://doi.org/10.18637/jss.v039.i08>
- Cohen J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.
- 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.
- Crısan, D.R., Tendeiro, J.N., & Meijer, R.R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, 41(6), 439–455.
<https://doi.org/10.1177/0146621617695522>
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B. & et al. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143–165.
[doi:10.1177/014662169501900203](https://doi.org/10.1177/014662169501900203)
- 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.
<https://doi.org/10.1007/s11136-007-9198-0>
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (Eds.). (2009). *Multivariate data analysis* (7th edition ed.). Upper Saddle River, NJ: Prentice Hall.
- Hatcher, L. (1994). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Cary, NC: SAS Institute, Inc.
- 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.
<https://doi.org/10.1080/10705519909540118>
- Kaplan D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate behavioral research*, 24(3), 285–305.
https://doi.org/10.1207/s15327906mbr2403_2
- Kline, R.B. (2005). *Principles and practice of structural equation modeling, Second Edition*. New York: Guilford Press.

Lai, J.S., Cella, D., Choi, S., Junghaenel, D.U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of physical medicine and rehabilitation*, 92(10 Suppl), S20–S27. <https://doi.org/10.1016/j.apmr.2010.08.033>

Lai, J.S., Crane, P.K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Qual Life Res*, 15(7), 1179–1190.

Lai, J.S., Zelko, F., Butt, Z., Cella, D., Kieran, M.W., Krull, K.R., Magasi, S., & Goldman, S. (2011). Parent-perceived child cognitive function: results from a sample drawn from the US general population. *Child's Nervous System*, 27(2), 285–293.

Lai, J.S., Zelko, F., Krull, K., Cella, D., Nowinski, C., Manley, P., & Goldman, S. (2014). Parent-reported cognition of children with cancer and its potential clinical usefulness. *Quality of Life Research*, 23(4), 1049–1058.

Luijben, T.C., & Boomsma, A. (1988). Statistical guidance for model modification in covariance structure analysis. *Compstat*, 335–340.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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. <https://doi.org/10.1016/j.jpsychires.2014.05.010>

PROMIS® Instrument Development and Psychometric Evaluation Scientific Standards, http://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf. (Vol. 2019).

Ramsay, J.O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. Montreal, Quebec: McGill University. Retrieved from www.psych.mcgill.ca/misc/fda/downloads/testgraf/TestGraf98.doc

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. <https://doi.org/10.1007/s11136-007-9183-7>

Saris, W.E., Satorra, A., & Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. In C.C. Clogg (Ed.), *Sociological methodology* (pp. 105–129). San Francisco, CA: Jossey-Bass.

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. <https://doi.org/10.1080/10705510903203433>

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? *The Journal of Applied Psychology*, 91(1), 25–39. <https://doi.org/10.1037/0021-9010.91.1.25>

Whittaker, T.A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education*, 80(1), 2644. <https://doi.org/10.1080/00220973.2010.531299>

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. <https://doi.org/10.1007/s11136-016-1467-3>

[Response Ends]

2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers).

If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, [NQF Measure Evaluation Criteria](#)).

[Response Begins]

Calibration sample

Establishing Unidimensionality.

Table 7: EFA Correlation matrix of the six Likert-type scorable items included in the PaLS

CORRELATION MATRIX	*	*	*	*	*	*
N = 510	*	*	*	*	*	*
*	2A	2B	2C	3A	3B	3C
2a. At least one member of my care team knows about my life goals	*	*	*	*	*	*
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.52	*	*	*	*	*
2c. My treatment plan is consistent with my life goals	0.56	0.36	*	*	*	*
3a. At least one member of my care team talks with me about my life goals	0.63	0.37	0.56	*	*	*
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.53	0.40	0.47	0.72	*	*
3c. At least one member of my care team helps me meet my life goals	0.61	0.34	0.60	0.84	0.70	*

Criterion: $r \geq 0.90$ indicates potentially redundant item content

* Cells intentionally left blank

Table 8: Eigenvalues from EFA analysis

EIGENVALUES FOR SAMPLE CORRELATION MATRIX	*	*	*	*
N = 510	*	*	*	*
*	1	2	Ratio of eigenvalue 1 to 2	% Variance
Eigenvalue	3.8	0.80	4.7	63.0

Criterion: Item set to have unidimensional characteristics if the ratio of eigenvalue 1 to eigenvalue 2 was ≥ 4 and the proportion of variance accounted for by eigenvalue 1 was ≥ 0.40 .

* Cells intentionally left blank

Table 9: Geomin rotated loadings from EFA analysis

GEOMIN ROTATED LOADINGS	*	*	*	*
N = 510	*	*	*	*
*	Factor 1	*	Factor 1	Factor 2

GEOMIN ROTATED LOADINGS	*	*	*	*
2a. At least one member of my care team knows about my life goals	0.73‡	*	0.85‡	0.02
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.51‡	*	0.71‡	-0.13
2c. My treatment plan is consistent with my life goals	0.66‡	*	0.38‡	0.33‡
3a. At least one member of my care team talks with me about my life goals	0.91‡	*	0.01	0.91‡
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.77‡	*	0.09	0.71‡
3c. At least one member of my care team helps me meet my life goals	0.90‡	*	-0.03	0.94‡

Criterion: Factor loadings >0.4 indicate items appear to have at least minimal construct validity.

‡ indicates significance at 5%

* Cells intentionally left blank

Table 10: Geomin factor correlation from EFA analysis

GEOMIN FACTOR CORRELATION	*	*	*	*	*
*	F1	*	*	F1	F2
F1	1	*	F1	1	*
*	*	*	F2	0.78‡	1

‡ indicates significance at 5%

* Cells intentionally left blank

Table 11: Fit Statistics from CFA analysis

1-factor model with modinidices at 100	*	*	*	2-factor model with modinidices at 100	*	*
N = 510	*	*	*	N = 510	*	*
Test	Value	90% CI	Probability RMSEA ≤0.05	Value	90% CI	Probability RMSEA ≤0.05
RMSEA	0.14	(0.12, 0.17)	0	0.09	(0.07, 0.12)	0.004
CFI	0.98	*	*	0.99	*	*
TLI	0.97	*	*	0.99	*	*
SRMR	0.06	*	*	0.04	*	*

Criterion: RMSEA <0.10, CFI and TLI ≥0.95, SRMR <0.08

* Cells intentionally left blank

Table 12: Standardized model results from CFA analysis, 1-factor model

Standardized model results - 1-factor model	*	*	*	*
N = 510	*	*	*	*
FACTOR1 by	Estimate	S.E	Est./S.E	Two-tailed p-value
2a. At least one member of my care team knows about my life goals	0.73	0.02	32.9	0
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.51	0.03	16.6	0
2c. My treatment plan is consistent with my life goals	0.66	0.02	27.1	0
3a. At least one member of my care team talks with me about my life goals	0.91	0.01	83.5	0
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.77	0.02	40.5	0
3c. At least one member of my care team helps me meet my life goals	0.90	0.01	75.1	0

Criterion: Poor construct validity indicated by factor loading <0.50.

* Cells intentionally left blank

Table 13: Standardized model results from CFA analysis, 2-factor model

Standardized model results - 2-factor model	*	*	*	*
N = 510	*	*	*	*
FACTOR1 by	Estimate	S.E	Est./S.E	Two-tailed p-value
2a. At least one member of my care team knows about my life goals	0.81	0.02	33.3	0
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.54	0.03	17.5	0
2c. My treatment plan is consistent with my life goals	0.73	0.03	27.5	0
FACTOR2 by				
3a. At least one member of my care team talks with me about my life goals	0.92	0.01	84.4	0
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.78	0.02	40.7	0
3c. At least one member of my care team helps me meet my life goals	0.91	0.01	74.7	0

Criterion: Poor construct validity indicated by factor loading <0.50.

* Cells intentionally left blank

Table 14: Residual correlation matrix from the CFA analysis, 1 factor model

RESIDUAL CORRELATION MATRIX	*	*	*	*	*	*
N = 510	*	*	*	*	*	*
*	2A	2B	2C	3A	3B	3C
2a. At least one member of my care team knows about my life goals	*	*	*	*	*	*
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.15	*	*	*	*	*
2c. My treatment plan is consistent with my life goals	0.08	0.03	*	*	*	*
3a. At least one member of my care team talks with me about my life goals	-0.03	-0.09	-0.04	*	*	*
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	-0.03	0.01	-0.04	0.02	*	*
3c. At least one member of my care team helps me meet my life goals	-0.05	-0.11	0.00	0.02	0.01	*

Criterion: Local item dependency indicated by residual correlation >0.20.

* Cells intentionally left blank

Table 15: Item fit statistics from CFA Analysis, 1-factor model

Item	X ²	df	Probability	Item fit
2a. At least one member of my care team knows about my life goals	84.4	53	0.0039	1.6
2b. I believe it is important at least one member of my care team talks with me about my life goals	115.5	57	0.0001	2.0
2c. My treatment plan is consistent with my life goals	72.8	54	0.04	1.3
3a. At least one member of my care team talks with me about my life goals	57.1	39	0.03	1.5
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	87.6	49	0.0006	1.8
3c. At least one member of my care team helps me meet my life goals	66.8	41	0.0067	1.6

Criterion: Good item fit indicated by item fit ≤3

* Cells intentionally left blank

Table 16: Item Bias

Total (N=517)	Count (%)
Sex	*
Female	245 (47.4)
Male	272 (52.6)
Race	*
White	364 (70.4)

Total (N=517)	Count (%)
Other ⁱ	153 (29.6)
Education	*
4-year college degree or more	298 (57.6)
Less than 4-year college degree*	219 (42.4)
Age	*
Greater or equal to median age split	258 (49.9)
Less than median age split ⁱ	259 (50.1)
ⁱ Other includes: Black/African American, Native American or Alaska Native, Asian, Pacific Islander, Do not wish to report, and More than One	*
*Less than 4-year college degree includes those that chose not to report	*
Age < median age split includes missing age	*

* Cells intentionally left blank

Table 17: Differential Item Function (DIF), Nagelkerke pseudo-R²

*	Age	*	*	Education	*	*	Gender	*	*	Race	*	*
*	Model 1 vs. 2	Model 1 vs. 3	Model 1 vs. 3	Model 1 vs. 2	Model 1 vs. 3	Model 1 vs. 3	Model 1 vs. 2	Model 1 vs. 3	Model 1 vs. 3	Model 1 vs. 2	Model 1 vs. 3	Model 1 vs. 3
2a	0.0001	0.001	0.0009	0.0038	0.0038	0	0.0015	0.0029	0.0015	0.001	0.001	0.0001
2b	0.0042	0.0042	0	0.0148	0.0171	0.0023	0.0009	0.0016	0.0007	0.0012	0.0077	0.0066
2c	0.0034	0.0034	0	0.0009	0.0009	0	0	0.0004	0.0004	0.0008	0.002	0.0012
3a	0.002	0.002	0	0.0001	0.0005	0.0004	0.0013	0.0013	0.0001	0.0015	0.0019	0.0004
3b	0.0001	0.0007	0.0006	0.0003	0.0009	0.0006	0	0.0004	0.0004	0.003	0.0047	0.0017
3c	0.0023	0.0028	0.0005	0.0006	0.001	0.0004	0.0008	0.0008	0.0001	0.0001	0.0001	0

* Cells intentionally left blank

Reliability Analyses.

Table 18: Reliability values of the PaLS

Reliability test	Calibration Sample	Validation testing Sample
Response pattern reliability	0.91	0.91
Median t-score Standard Error	2.86	2.85
T-score standard deviation	9.47	9.69
Cronbach alpha reliability	0.84	0.85

Reliability test	Calibration Sample	Validation testing Sample
Marginal reliability	0.90	0.91

Table 19: Item-level Cronbach's alpha if an item is deleted

N = 510	*	*	*	*
Deleted Variables	Raw Variables	*	Standardized Variables	*
*	Correlation with Total	Alpha	Correlation with Total	Alpha
2a. At least one member of my care team knows about my life goals	0.65	0.82	0.66	0.81
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.40	0.86	0.40	0.86
2c. My treatment plan is consistent with my life goals	0.57	0.84	0.57	0.83
3a. At least one member of my care team talks with me about my life goals	0.77	0.80	0.75	0.80
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.66	0.82	0.65	0.81
3c. At least one member of my care team helps me meet my life goals	0.75	0.80	0.74	0.80

* Cells intentionally left blank

Table 20: Pearson correlation coefficients among the six Likert-type scorable PaLS items

*	2A	2B	2C	3A	3B	3C
2a. At least one member of my care team knows about my life goals	1	0.42 <.0001	0.49 <.0001	0.56 <.0001	0.47 <.0001	0.54 <.0001
2b. I believe it is important at least one member of my care team talks with me about my life goals	*	1	0.29 <.0001	0.30 <.0001	0.33 <.0001	0.28 <.0001
2c. My treatment plan is consistent with my life goals	*	*	1	0.48 <.0001	0.41 <.0001	0.52 <.0001
3a. At least one member of my care team talks with me about my life goals	*	*	*	1	0.64 <.0001	0.78 <.0001
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	*	*	*	*	1	0.62 <.0001
3c. At least one member of my care team helps me meet my life goals	*	*	*	*	*	1

* Cells intentionally left blank

Validation testing sample:

Reliability Analyses.

Table 21: Item level Cronbach's alpha if an item is deleted

N = 416	*	*	*	*
Deleted Variables	Raw Variables	*	Standardized Variables	*
*	Correlation with Total	Alpha	Correlation with Total	Alpha
2a. At least one member of my care team knows about my life goals	0.66	0.84	0.66	0.83
2b. I believe it is important at least one member of my care team talks with me about my life goals	0.40	0.88	0.40	0.87
2c. My treatment plan is consistent with my life goals	0.58	0.85	0.58	0.84
3a. At least one member of my care team talks with me about my life goals	0.79	0.81	0.78	0.80
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	0.71	0.83	0.69	0.82
3c. At least one member of my care team helps me meet my life goals	0.78	0.81	0.77	0.81

* Cells intentionally left blank

Table 22: Pearson correlation coefficients among six Likert-type scorable PaLS items

*	2A	2B	2C	3A	3B	3C
2a. At least one member of my care team knows about my life goals	1	0.39 <.0001	0.49 <.0001	0.59 <.0001	0.50 <.0001	0.57 <.0001
2b. I believe it is important at least one member of my care team talks with me about my life goals	*	1	0.28 <.0001	0.33 <.0001	0.31 <.0001	0.31 <.0001
2c. My treatment plan is consistent with my life goals	*	*	1	0.51 <.0001	0.44 <.0001	0.53 <.0001
3a. At least one member of my care team talks with me about my life goals	*	*	*	1	0.70 <.0001	0.79 <.0001
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	*	*	*	*	1	0.70 <.0001
3c. At least one member of my care team helps me meet my life goals	*	*	*	*	*	1

* Cells intentionally left blank

Table 23: Test-retest Reliability, Minimal Detectable Change, and Standard Error of Measurement of PaLS t-score, time points combined

*	ICC	SEM	SEM %	DC% ₉₅ (LDC, UDC)
PaLS t-score	0.80	4.2	8.4	11.7 (-10.7, 12.6)

Criterion: Minimum acceptable criteria for the intraclass correlation used to analyze test-retest reliability was set at ≥ 0.70 (Cohen, 1969). Standard Error of Measurement (SEM) percent less 10% indicates good (i.e., acceptable) measurement error.

* Cells intentionally left blank

Reference:

Cohen J. *Statistical power analysis for the behavioral sciences*. New York: Academic Press; 1969.

[Response Ends]

2a.12. Interpret the results, in terms of how they demonstrate reliability.

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]

The first set of analyses was focused on the identification of a unidimensional set of items and included EFA and CFA analysis. The findings from these analyses indicated that the six Likert-type scorable PaLS items were essentially unidimensional. An examination of item bias and DIF analysis did not identify any problem items for factors investigated, and a final CFA indicated good fit statistics supporting essential unidimensionality. Following these analyses we examined reliability at both the score and item-level, as well as measurement error. Specific results are discussed below.

Calibration sample:

Establishing a Unidimensional Set of Items.

CTT:

Item-adjusted total score correlations should be ≥ 0.40 . The six Likert-type scorable PaLS items all had item-adjusted total score correlations above 0.40.

EFA:

In EFA, correlations among items that ≥ 0.90 may indicate redundant item content, i.e., that items are measuring the same thing (Hu & Bentler, 1999; Kline, 2005; Lai et al., 2006, 2011). Therefore, we would like item correlations to be below 0.90. For EFA, we considered the item set to have unidimensional characteristics if the ratio of eigenvalue 1 to eigenvalue 2 was ≥ 4 and the proportion of variance accounted for by eigenvalue 1 was ≥ 0.40 . For geomin rotated factor loadings, a factor loading ≥ 0.40 supports item construct validity.

In the calibration sample, all six Likert-type scorable PaLS items had correlations below 0.90. The ratio of eigenvalue 1 to 2 for the calibration sample life goals survey was 4.7, indicating a unidimensional model appeared appropriate. For geomin rotated factor loadings, in the 1-factor model, all factor loadings were above 0.4, indicating that all items appeared to have construct validity.

CFA:

Good fit in CFA is indicated by RMSEA < 0.1 , CFI ≥ 0.95 , TLI ≥ 0.95 , and SRMR < 0.08 (Cook et al., 2009; Kline, 2005; Bentler, 1990; Hu & Bentler, 1999; Hatcher, 1994; Lai et al., 2011, 2014). Signs of a poor model and fit include factor loadings < 0.50 , residual correlations > 0.20 , and correlated error modification index values ≥ 100 (Cook et al., 2009; Kaplan, 1989; Luijben et al., 1988; McDonald, 1999; Reise et al., 2007; Saris et al., 1987, 2009; Whittaker, 2012). In the calibration sample, the RMSEA was 0.14, which was higher than the criterion of 0.1. A 2-factor model was explored; the 2-factor model was not deemed appropriate based on modeling results (presented above). CFI and TLI were 0.98 and 0.97, respectively which met the criteria for indicating good fit. Factor loadings (presented above) were 0.50 or higher. Residual correlations (presented above) were < 0.20 , which indicated no problems with

local item dependency. No items were flagged for MI-based correlated errors (therefore results not shown), which further supported there being no problems with local item dependency.

IRT:

Using the IRT-based approach to calculate item fit, poor fit was identified if the misfit quotient value was >3 (Crisan et al., 2017; Drasgow et al., 1995; Stark et al., 2006; Zhao, 2017). Using the IRT-based item fit assessment, all fit values were below 3, indicating good item fit.

Item bias:

For item bias, we expected an equivalent estimation of item parameters across tested groups. We did not expect to see items flagged for differential item functioning (DIF; i.e., based on the methods described in section 2a.10). Overall there was no evidence for item bias for the factors of sex, education, age, or race, i.e., no items were flagged for DIF for any of the DIF factors investigated.

Calibration sample and Validation testing sample

Reliability of the PaLS t-scores

In the calibration and validation testing samples, response pattern reliability, Cronbach's alpha reliability, and marginal reliability all supported the internal consistency reliability of the set of six Likert-type scorable items included in the PaLS instrument (i.e., all values were ≥ 0.84 , indicating "very good" internal consistency; Cohen, 1969). Additionally, all measures of reliability were consistent across both testing samples.

Validation testing sample

T-score Level Reliability and Measurement Error

Minimum acceptable criteria for the intraclass correlation used to analyze test-retest reliability was set at ≥ 0.70 (Cohen, 1969). Standard Error of Measurement (SEM) percent $<10\%$ indicates good (i.e., acceptable) measurement error (Flansbjerg et al., 2005).

In the validation testing sample, the test-retest reliability was very good, with an ICC = 0.80. The SEM percent was 8.4%, indicating low-level measurement error. Minimal detectable change was 11.7%, with a 95% confidence of -10.7 to 12.6.

References:

- Bentler, P.M. (1990). *Comparative Fit Indexes in Structural Models*. *Psychological Bulletin*, 107(2), 238-246.
- Cohen J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- 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.
- Crişan, D.R., Tendeiro, J.N., & Meijer, R.R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, 41(6), 439-455.
<https://doi.org/10.1177/0146621617695522>
- Drasgow, F., Levine, M.V., Tsien, S., Williams, B. & et al. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143-165.
Doi:10.1177/014662169501900203
- Flansbjerg, U.B., Holmbäck, A.M., Downham, D., Patten, C., & Lexell, J. (2005). Reliability of gait performance tests in men and women with hemiparesis after stroke. *Journal of rehabilitation medicine*, 37(2), 75-82.
<https://doi.org/10.1080/16501970410017215>
- Hatcher, L. (1994). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Cary, NC: SAS Institute, Inc.
- 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.
<https://doi.org/10.1080/10705519909540118>

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate behavioral research*, 24(3), 285–305.

https://doi.org/10.1207/s15327906mbr2403_2

Kline, R.B. (2005). *Principles and practice of structural equation modeling, Second Edition*. New York: Guilford Press.

Lai, J.S., Cella, D., Choi, S., Junghaenel, D.U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, 92(10 Suppl), S20–S27.

<https://doi.org/10.1016/j.apmr.2010.08.033>

Lai, J.S., Crane, P.K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Qual Life Res*, 15(7), 1179–1190.

Lai, J.S., Zelko, F., Butt, Z., Cella, D., Kieran, M. W., Krull, K. R., Magasi, S., & Goldman, S. (2011). Parent-perceived child cognitive function: results from a sample drawn from the US general population. *Child's Nervous System*, 27(2), 285–293.

Lai, J. S., Zelko, F., Krull, K., Cella, D., Nowinski, C., Manley, P., & Goldman, S. (2014). Parent-reported cognition of children with cancer and its potential clinical usefulness. *Quality of Life Research*, 23(4), 1049–1058.

Luijben, T. C., & Boomsma, A. (1988). Statistical guidance for model modification in covariance structure analysis. *Compstat*, 335–340.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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. <https://doi.org/10.1007/s11136-007-9183-7>

Saris, W.E., Satorra, A., & Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. In C. C. Clogg (Ed.), *Sociological methodology* (pp. 105–129). San Francisco, CA: Jossey-Bass.

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. <https://doi.org/10.1080/10705510903203433>

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?. *The Journal of applied psychology*, 91(1), 25–39. <https://doi.org/10.1037/0021-9010.91.1.25>

Whittaker, T.A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education*, 80(1), 26–44.

<https://doi.org/10.1080/00220973.2010.531299>

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.

<https://doi.org/10.1007/s11136-016-1467-3>

[Response Ends]

2b.01. Select the level of validity testing that was conducted.

[Response Begins]

Patient or Encounter-Level (data element validity must address ALL critical data elements)

[Response Ends]

2b.02. For each level of testing checked above, describe the method of validity testing and what it tests.

Describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used.

[Response Begins]

We performed several series of tests to assess different aspects of validity for the PaLS *t*-scores in the calibration and validation testing samples. The methods and samples used for each set of validity tests are described below.

Calibration sample and Validation testing sample

Known-groups validity. Known-groups validity was used to analyze whether there were differences in PaLS *t*-score for groups expected to have different mean responses based on their demographic or clinical characteristics. In the calibration and validation testing samples, we looked for mean score differences based on dialysis modality. For modality, a one-sided upper-tail *t*-test was computed for the *t*-score, stratified by variable characteristic of interest. The *t*-test was limited to participants that answered all six Likert-type scorable PaLS items. Pooled *t* statistics were used for stratifications with equal variance based on the *F*-test.

In the validation sample we looked for PaLS mean *t*-score differences using "high" versus "low" health-related quality of life (HRQOL) for several PROMIS domain scores (i.e., Global Physical Health, Global Mental Health, Meaning and Purpose, Ability to Participate, Depression, and Self Efficacy). Mean *t*-tests were used to compare patients with low vs. high PROMIS measure-specific HRQOL status. The Folded *F*-test was used to determine whether variances were equal between the two groups; the pooled degrees of freedom method was used when variances were equal, and Satterthwaite degrees of freedom method was used when variances were unequal. We hypothesized that participants with poor mean HRQOL scores (i.e., >0.5 SD from the mean in the "worse health" direction) would report greater dissatisfaction with their life goals discussion (i.e., have poorer mean PaLS scores) than participants with good HRQOL scores (i.e., >0.5 SD from the mean in the "better health" direction).

Base Rates for Patients with Dissatisfaction with Life Goals discussions. Dissatisfaction with life goals discussions (i.e., participants whose PaLS *t*-score was ≤40) were evaluated to determine if participants that have poor HRQOL were at increased risk for dissatisfaction with life goals discussions compared to those with good HRQOL. We expected that, based on the mean, 16% of the US general population would have poor scores on the PaLS. We anticipated this rate would be higher for those with poor versus good HRQOL; rates >16% for those with poor HRQOL would indicate greater dissatisfaction than expected.

Effect Sizes. Cohen's *d* effect sizes for the patient-level PaLS *t*-score was computed for the high versus low HRQOL groups for each of the PROMIS domain measures using the standard equation:

$$Cohen's d = \frac{mean_{group1} - mean_{group2}}{standarddeviation_{pooled}}$$

Values of *d* between 0.20 and 0.49 were considered "small", values between 0.50 and 0.79 were considered "moderate," and values ≥0.80 were considered "large" (Cohen, 1988).

Floor and ceiling effects. Floor and ceiling effects assessments were used to determine if a high proportion of participants was responding "strongly disagree" or "strongly agree" to all six Likert-type scorable PaLS items. High floor and ceiling effects can be a threat to validity. The floor and ceiling effects were calculated by creating a measure score for each participant based on the numeric value of how the participant responded to the item. The number of participants with the lowest possible score of 6 was divided by the total number of participants to complete all six Likert-type scorable PaLS items for the floor effect. The number of participants with the highest possible score of 30 was divided by the total number of participants to complete all six Likert-type scorable PaLS items for the ceiling effect. *N*, mean, and standard deviation descriptive statistics were obtained for participants answering all six Likert-type scorable PaLS items. *A priori* criterion for acceptable floor and ceiling effects was specified as ≤20% (Andresen, 2000; Cramer & Howitt, 2001).

Validation testing sample

Convergent and discriminant validity. We investigated convergent and discriminant validity for the patient-level PaLS *t*-score using Pearson correlations. Convergent validity indicates whether the PaLS *t*-score is moderately or

highly correlated with similar measured concepts. Discriminant validity is used to test that two measures have low or no correlation and are measuring unrelated concepts.

The PaLS *t-score* was compared to the PROMIS domain measure scores administered in the validation testing sample. Convergent validity would be supported by observing “moderate” to “high” correlations between the PaLS *t-score* and PROMIS measure scores. Discriminant validity would be supported by “low” correlations between the PaLS *t-score* and PROMIS measure scores. “Low”, “moderate”, and “high” were defined as: “low” = $r \leq 0.35$, “moderate” = $r \geq 0.36 - 0.67$, “high” = r between 0.68 and 0.89 (Campbell & Fiske, 1959). We expected evidence of convergent validity for the PaLS *t-scores*’ association with PROMIS Meaning and Purpose.

Responsiveness. Responsiveness was tested using a modified version of the Life Events survey as well as the 3- and 6-month assessments of the PaLS. The Life Event survey (Holmes & Rahe, 1969) is a self-report survey where participants select from a list of personal, relational, health, financial and social related events experienced within the last 3 months. In order to tailor this to the ESRD chronic dialysis population for our testing, the existing Life Events survey was modified to remove events not applicable and add several specific ESRD and health-related events including switched to a different dialysis modality, switched vascular access used, were offered a kidney, switched dialysis facilities, switched to a different kidney doctor, had a change in care team, were hospitalized for any reason, had a health event requiring immediate care, death (for any reason) of someone you know that is on dialysis, COVID-19 infection or death in family member or close friend. These modifications were guided by clinical nephrologist input. Response options for all life events included “Have not experienced in the past 3 months”, extremely negative impact on life (-3) to extremely positive impact on life (+3). We anticipated a relationship between magnitude of life events experienced (agnostic to positive or negative) and changes in a participant’s PaLS *t-score*.

The absolute difference in PaLS *t-score* was calculated between baseline and 3 months as well as baseline and 6 months. Baseline to 3 months results are presented below; baseline to 6 months were explored but not found to be statistically significant (data not reported). The **number of life events** a participant experienced at each time point was summed to explore differences in *t-score* based on the number of life events a participant experienced. We hypothesized that the more life events a participant had at a given time point, the more likely there would also be a significant change in *t-score* for these associated time points.

We compared participants with between zero life events and four life events to participants with five or more life events. We also calculated the median split of life events, which was found to be three life events, comparing participants with ≤ 3 life events vs. participants with > 3 life events.

Finally, the **impact of life events** was explored using the rating each participant gave to selected life events. The absolute sum of the impact of life events was calculated and the median split of impact was found to be seven. Participants that had life events with an impact ≤ 7 were compared to participants with an impact > 7 . Participants that did not experience any life events at each time point were excluded from the impact analysis.

References:

- Andresen E.M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation*, 81(12 Suppl 2), S15–S20. <https://doi.org/10.1053/apmr.2000.20619>
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56, 81-105.
- Cramer, D., & Howitt, D.L. (2004). *The Sage dictionary of statistics*. Thousand Oaks, CA: Sage.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, New York: Lawrence Erlbaum Associates
- Holmes, T.H., & Rahe, R.H. (1967). The Social Readjustment Rating Scale. *Journal of Psychosomatic Research*, 11(2), 213–218. [https://doi.org/10.1016/0022-3999\(67\)90010-4](https://doi.org/10.1016/0022-3999(67)90010-4)

[Response Ends]

2b.03. Provide the statistical results from validity testing.*Examples may include correlations or t-test results.***[Response Begins]****Known-groups Validity**

Table 24: Known-groups validity for dialysis modality

*	PaLS Scores	*	*	PaLS Scores	*	*	*	*	*
*	HHD; PD	*	*	ICHD	*	*	*	*	*
Sample	N	Mean (SD) PaLS t-score	% of participants with poor PaLS t-score	N	Mean (SD) PaLS t-score	% of participants with poor PaLS t-score	t	p-value	Cohen's d
Calibration	125	53.3 (8.9)	8.8	333	48.8 (9.1)	18.3	4.7	<.0001	0.50
Validation	106	50.7 (9.9)	15.1	225	49.0 (9.3)	16.9	1.5	0.07	0.18

ICHD = in-center hemodialysis; HHD = home hemodialysis; PD = peritoneal dialysis.

* Cells intentionally left blank

Table 25: Known-groups validity for PROMIS Measures

*	PaLS Scores	*	*	PaLS Scores	*	*	*	*	*
*	Low HRQOL ^a	*	*	High HRQOL ^b	*	*	*	*	*
PROMIS Measure	N	Mean (SD) PaLS t-score	% of participants with poor PaLS t-score	N	Mean (SD) PaLS t-score	% of participants with poor PaLS t-score	t	p-value	Cohen's d
Global Physical Health	282	49.3 (9.3)	16.3	25	56.5 (11.5)	16.0	3.6	0.0002	0.76
Global Mental Health	217	48.2 (9.1)	19.4	58	54.6 (11.7)	17.2	3.9	0.0001	0.67
Meaning and Purpose	152	45.7 (7.4)	24.3	120	55.2 (10.4)	10.8	8.7	<0.0001	1.1
Ability to Participate	256	48.7 (9.2)	18.4	31	56.3 (11.6)	16.1	4.2	<0.0001	0.80
Depression†	197	47.7 (9.1)	20.3	97	52.9 (10.2)	11.3	-4.5	<0.0001	-0.56

*	PaLS Scores	*	*	PaLS Scores	*	*	*	*	*
Self efficacy	172	47.4 (8.6)	19.8	67	54.9 (11.8)	16.4	4.7	<0.0001	0.78

Criterion: ^a for each PROMIS domain, Low HRQOL reflects participants that are 0.5 SD away from the mean in the “worse health” direction; ^b for each PROMIS domain, High HRQOL reflects participants that are 0.5 SD away from the mean in the “better health” direction; ‡Depression negatively worded concept. All other PROMIS measures are positively worded concepts

* Cells intentionally left blank

Floor and Ceiling Effects

Calibration sample and Validation testing sample

Table 26: Floor and ceiling effects for measure score

Sample	N	Floor effect (%) (a priori <20%)	Ceiling effect (%) (a priori <20%)	Mean measure score	STD of measure score
Calibration	510	0.39	6.1	20.4	5.6
Validation	416	0.48	5.8	20.2	5.7

* Cells intentionally left blank

Convergent and discriminant validity

Validation testing sample

Table 27: Pearson correlation examining convergent validity of PaLS *t*-score

*	PaLS (a priori ≥0.36)
PROMIS Meaning and Purpose	0.46

Criterion: $r < 0.36$ discriminant validity, $r \geq 0.36$ convergent validity

* Cells intentionally left blank

Table 28: Pearson correlation examining discriminant validity of PaLS *t*-score

PROMIS measure	PaLS (a priori <0.36)
PROMIS Ability to participate	0.27
PROMIS Depression	-0.29

PROMIS measure	PaLS (a priori <0.36)
PROMIS Global Health Mental Health	0.30
PROMIS Global Health Physical Health	0.22
PROMIS Self Efficacy	0.34

Responsiveness**Validation testing sample**

Table 29: 3-month responsiveness of the absolute difference in PaLS t-score relative to different groupings of self-reported number of life events

*	N	Mean (std)	N	Mean (std)	t	p-value
*	BETWEEN ZERO AND FOUR LIFE EVENTS	*	FIVE OR MORE LIFE EVENTS	*	*	*
PaLS	111	4.8 (4.6)	69	6.6 (5.8)	-2.3	0.03
*	BETWEEN ZERO AND THREE LIFE EVENTS	*	≥ THREE LIFE EVENTS	*	*	*
PaLS	96	4.4 (4.2)	84	6.7 (5.8)	-2.9	0.004
*	≤ IMPACT OF SEVEN	*	>IMPACT OF SEVEN	*	*	*
PaLS	85	4.4 (4.8)	73	6.9 (5.7)	-3.1	0.003

* Cells intentionally left blank

[Response Ends]

2b.04. Provide your interpretation of the results in terms of demonstrating validity. (i.e., what do the results mean and what are the norms for the test conducted?)

[Response Begins]

Results from validity testing met our respective criteria for known-groups validity, floor and ceiling effects, convergent and discriminant validity, and responsiveness. Discussion of each set of validity results for each respective testing sample (calibration or validation testing) follows below.

Known-groups validation

Modality. Known-groups validity was investigated in both the calibration sample and the validation testing sample. For the calibration sample, as expected, patients on a home dialysis modality report greater life goals satisfaction than those on in-center dialysis, supporting known-groups validity for this clinical factor. While we saw a similar trend in the validation testing sample, this finding did not meet conventional levels of significance (i.e., $p=0.07$). The absence of a statistically significant difference may possibly be attributed to a slightly smaller sample size in the validation testing sample. However, the effect size was 0.50 in the calibration sample, which is considered moderate and 0.18 in the validation testing sample, which is considered negligible.

HRQOL. Known-groups validity testing for these analyses was conducted using the validation testing sample only. In all cases results were in accordance with proposed hypotheses. As expected, we found that participants with poor HRQOL reported significantly greater dissatisfaction with life goals discussions than participants with good HRQOL. Specifically, those with poor global physical health, poor global mental health, poor meaning and purpose, poor ability to participate in social roles and activities, higher self-reported depression and poorer self-efficacy reported significantly more dissatisfaction with patient life goals discussions than those with good global physical health, good global mental health, good meaning and purpose, good ability to participate in social roles and activities, lower self-reported depression and good self-efficacy, respectively.

Baseline Rates for Patients with Dissatisfaction with Life Goals discussions. As expected, baseline rates of dissatisfaction with life goals discussions were consistently higher for those on in-center hemodialysis versus those on a home dialysis modality. Baseline rates were also higher for those with poor versus good HRQOL (for all measured PROMIS domains). In addition, rates of dissatisfaction consistently met or exceeded what was expected (16%), both for those on in-center hemodialysis and for those individuals with poor HRQOL.

Effect Sizes. For Cohen's d effect size, we aimed to have moderate or high effect sizes for the different groupings that we examined. These groupings included modality and the different HRQOL domains (as measured by the PROMIS measures). As expected, effect sizes were generally moderate to large; the largest effect size was seen between the PaLS groups for meaning and purpose. The sole exception was the negligible effect size that was seen for the validation testing sample for dialysis modality.

Floor and ceiling effects

We found no evidence of floor or ceiling effects in either results from our calibration sample or the validation testing sample. In the calibration and validation testing samples, floor and ceiling effects were both below 20% (floor effects 0.39% and ceiling effects 6.1% in the calibration sample; floor effects 0.48% and ceiling effects 5.8% in the validation testing sample). This indicates that only a small percentage of participants selected the lowest (worst) or highest (best) responses about their satisfaction with their care team's discussions about life goals, and that most participants had varied responses across the six Likert-type scorable PaLS items.

Convergent and discriminant validity

We expected convergent validity would be supported for the association between patient-level PaLS t -score and PROMIS Meaning and Purpose scores in our validation testing sample. Note that convergent validity was considered supported if the between-score correlation was ≥ 0.36 . Discriminant validity was supported if the correlation was < 0.36 . PROMIS Meaning and Purpose scores and patient-level PaLS t -score had a correlation of 0.46, indicating evidence of convergent validity, as hypothesized. All other correlations between PROMIS measure scores and PaLS scores were < 0.36 , therefore supporting our expectation of providing evidence of discriminant validity.

Responsiveness

3- and 6-month responsiveness was explored in our validation testing sample. We report on 3-month responsiveness below; baseline to 6-months responsiveness was explored but not found to be statistically significant and thus is not reported. As we hypothesized, the number of life events reported by participants yielded greater absolute PaLS t -score changes compared to those below each specified threshold of life events. This indicates that participants reported greater dissatisfaction with their care team's discussion about life goals within 3 months of experiencing a life event. We assume that life events, particularly a culmination of those over a relatively short period of time, may result in individuals re-evaluating their life goals, including whether their

treatment plan needs to be modified. These time periods then present opportunities to revisit these discussions with the dialysis care team.

Participants with five or more life events had a greater absolute change in *t*-score compared to participants with between zero and four life events at 3 months (Table 29, $p=0.03$). Additionally, participants that had greater than the median split of life events (i.e., three life events) had a greater absolute change in the PaLS *t*-score compared to participants with \leq the median split of life events at 3 months (Table 29, $p=0.004$).

When looking at the impact of life events, participants that experienced more than the median split of number of impactful life events (i.e., seven) had a greater change in PaLS *t*-score compared to participants with fewer impactful life events (Table 29, $p=0.003$).

[Response Ends]

2b.05. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified.

Describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]

Calibration and Validation testing sample

Clinically meaningful. We investigated the change in patient-level *t*-scores between baseline, 3-month, and 6-month time points for participants included in the validation testing sample. Time points included 1) validation testing baseline to validation testing 3-month time point, and 2) validation testing baseline to validation testing 6-month time point. Participants were not required to complete 3-month or 6-month follow ups.

Additionally, we investigated the change in patient-level *t*-score for participants that were included in both the calibration sample and the validation testing sample. Time points included 1) calibration testing to validation testing baseline, 2) validation testing baseline to validation testing 3-month time point, and 3) validation testing baseline to validation testing 6-month time point.

Meaningful differences among patient-level *t*-score was used because data were not available to support facility-level *t*-score calculation at this stage. We expected to see variability in a patient-level *t*-score over time, as life goals may change for patients at different times. Differences in a patient-level *t*-score ≥ 5 (i.e., ≥ 0.5 SDs) was considered clinically meaningful (Heaton et al., 2004).

Validation testing sample

Statistical significance. The Guyatt's Responsiveness Statistic (RS) and Standardized Response Mean (SRM) effect sizes were calculated to examine the responsiveness of the PaLS measure scores. RS was calculated by dividing the mean change of PaLS *t*-scores for each PROMIS change group by the standard deviation of change in PaLS *t*-score in the "no change" group (Guyatt et al., 1987). SRMs were calculated by dividing the mean change of PaLS *t*-score for each group by the standard deviation of change of PaLS *t*-score for that group. RS and SRM were calculated relative to PROMIS Global Physical Health, Global Mental Health, and Meaning and Purpose. "No change" in PROMIS *t*-score was defined as a change $< |5|$ *t*-score points (i.e., < 0.5 SDs) between baseline and follow up. "Change" in PROMIS *t*-score is defined as a change $\geq |5|$ *t*-score points (≥ 0.5 SDs). Effect sizes between 0.00 and $|0.19|$ were considered "negligible", $|0.20|$ to $|0.49|$ were small, $|0.50|$ to $|0.79|$ were moderate, and $\geq |0.80|$ were large.

For participants with a "change" in PROMIS *t*-score, we predicted small RS/SRMs. For participants with "no change" in PROMIS *t*-score we predicted "negligible" RS/SRMs.

General linear models (GLMs) were used to examine change over time (from baseline to 3-month and baseline to 6-month time points) for PaLS *t*-score relative to PROMIS Global Physical Health, PROMIS Global Mental Health, and PROMIS Meaning and Purpose scores (Cohen, 1992; Kopjar, 1996). Each model included predictors for

“change” in PROMIS *t*-score. Least-square means and standard errors were calculated for each change group to determine whether change over time was significantly different from zero.

Responsiveness was supported by significant change in PaLS *t*-score relative to change in PROMIS measure *t*-score.

References:

Cohen, J.A. (1992). Power primer. *Psychol Bull*, 112(1), 155–159.

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*, 40(2), 171–178.

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*. Lutz, FL: Psychological Assessment Resources, Inc.

Kopjar, B. (1996). The SF-36 health survey: a valid measure of changes in health status after injury. *Inj Prev*, 2, 135–139.

[Response Ends]

2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined.

[Response Begins]

Calibration and Validation testing sample

Table 30: Difference in PaLS *t*-score between survey time points, calibration testing and validation testing baseline, 3-month, and 6-month time points

*	No change	*	Change >0 and <2	*	Change ≥2 and <5	*	Change ≥5	*
Validation testing	N	%	N	%	N	%	N	%
Baseline to 3-month (N=183)	9	4.9	33	18.0	62	33.9	79	43.2
Baseline to 6-month (N=167)	8	4.8	35	21.0	54	32.3	70	41.9
Calibration testing and Validation testing sample	*	*	*	*	*	*	*	*
Calibration testing to Baseline (N=95)	5	5.3	31	32.6	25	26.3	34	35.8
Baseline to 3-month (N=57)	3	5.3	9	15.8	26	45.6	19	33.3
Baseline to 6-month (N=45)	4	8.9	12	26.7	14	31.1	15	33.3

* Cells intentionally left blank

Validation testing sample

Table 31: Guyatt's Responsiveness statistic for changes in PaLS t-score

Validation testing	Baseline to 3-month	*	*	*	*	*	Baseline to 6-month	*	*	*	*	*
Global Physical Health	No change	*	*	Change	*	*	No change	*	*	Change	*	*
*	N	RS	SRM	N	RS	SRM	N	RS	SRM	N	RS	SRM
PaLS	111	0.25	0.25	66	-0.002	-0.002	106	0.18	0.18	56	0.006	0.004
*	*	*	*	*	*	*	*	*	*	*	*	*
Global Mental Health	No change	*	*	Change	*	*	No change	*	*	Change	*	*
	N	RS	SRM	N	RS	SRM	N	RS	SRM	N	RS	SRM
PaLS	109	0.14	0.14	68	0.17	0.16	93	0.16	0.16	69	0.03	0.03
*	*	*	*	*	*	*	*	*	*	*	*	*
Meaning and Purpose	No change	*	*	Change	*	*	No change	*	*	Change	*	*
*	N	RS	SRM	N	RS	SRM	N	RS	SRM	N	RS	SRM
PaLS	114	0.26	0.26	62	0.07	0.04	86	-0.01	-0.01	74	0.28	0.25

* Cells intentionally left blank

Table 32: Responsiveness relative to change in PROMIS measure score

Validation testing	Baseline to 3-month	*	*	*	*	*	Baseline to 6-month	*	*	*	*	*
Global Physical Health	No change	*	*	Change	*	*	No change	*	*	Change	*	*
*	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value
PaLS	1.78	-0.71	0.01	-0.01	0.92	0.99	1.17	0.74	0.12	0.04	1.02	0.97
*	*	*	*	*	*	*	*	*	*	*	*	*

Validation testing	Baseline to 3-month	*	*	*	*	*	Baseline to 6-month	*	*	*	*	*
Global Mental Health	No change	*	*	Change	*	*	No change	*	*	Change	*	*
	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value
PaLS	1.00	0.72	0.17	1.27	0.92	0.17	1.20	0.79	0.13	0.22	0.92	0.82
*	*	*	*	*	*	*	*	*	*	*	*	*
Meaning and Purpose	No change	*	*	Change	*	*	No change	*	*	Change	*	*
*	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value	Least squared mean	SE	p-value
PaLS	1.58	0.70	0.03	0.43	0.95	0.65	-0.08	0.82	0.93	1.99	0.88	0.02

* Cells intentionally left blank

[Response Ends]

2b.07. Provide your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]

Calibration and Validation testing sample

Clinical significance. As hypothesized, participant PaLS *t*-scores changed over time. In the validation testing, 43.2% of participants had a meaningful change in *t*-score (change in *t*-score ≥ 5) between baseline and 3 months. Additionally, 41.9% of participants had a meaningful change in *t*-score between baseline and 6 months.

For participants in both the calibration sample and validation testing sample, 35.8% of participants had a meaningful change in PaLS *t*-score between the time of response in the calibration sample and validation testing baseline. 33.3% of participants experienced a meaningful change in PaLS *t*-score between validation testing baseline and 3 months as well as between validation testing and 6 months.

While the differences presented above are agnostic to positive or negative changes in patient-level PaLS *t*-score, the results illustrate that a patient's life goals may change over time and should be discussed with the care team regularly. This also shows that the PaLS survey can differentiate patient responses over time.

Validation testing sample

Statistical significance. The SRMs were "negligible" for Global Physical Health from baseline to 6 months, and for Global Mental Health from baseline to 3 months and from baseline to 6 months. The SRM was small for Meaning and Purpose from baseline to 6 months.

Responsiveness was supported in the “no change” baseline to 3-month group of Global Physical Health and Meaning and Purpose. Additionally, responsiveness was supported in the “change” baseline to 6-month group of Meaning and Purpose.

While we expected change in PROMIS measure scores to be predictive of change in PaLS *t*-score, they were not. Participants may not have experienced conversations with their care team initiating changes in life goals between baseline and follow up even if they experienced a change in physical health, mental health, or their feelings of meaning and purpose indicated in the PROMIS measures.

[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]

To identify the extent of missing data, we looked at the count of participants that skipped each of the six Likert-type scorable PaLS items. Because the PaLS *t*-score can be calculated with missing items, low levels of missingness do not bias our results. Thus, due to observed low levels of missingness, differences between participants that responded to all items and participants that did not respond to all items was not investigated.

[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]

Calibration sample

Table 33: Count and percent of participants with missing scorable Likert-type items

Item (N = 517)	Count (%) participants missing
2a. At least one member of my care team knows about my life goals	4 (0.77)
2b. I believe it is important at least one member of my care team talks with me about my life goals	2 (0.39)
2c. My treatment plan is consistent with my life goals	2 (0.39)
3a. At least one member of my care team talks with me about my life goals	3 (0.58)

Item (N = 517)	Count (%) participants missing
3b. I feel comfortable discussing changes in my life goals with at least one member of my care team	1 (0.19)
3c. At least one member of my care team help me meet my life goals	3 (0.58)
Total count of participants that skipped at least one life goals item	7 (1.4)

Validation testing sample

Table 34: Count and percent of participants with missing scorable Likert-type items

Item (N = 420)	Count (%) participants missing
2a. At least one member of my care team knows about my life goals	0 (0)
2b. I believe it is important at least one member of my care team talks with me about my life goals	1 (0.24)
2c. My treatment plan is consistent with my life goals	1 (0.24)
3a. At least one member of my care team talks with me about my life goals	0 (0)
3b. I feel comfortable discussing changes in my life goals with my care team	1 (0.24)
3c. At least one member of my care team helps me meet my life goals	1 (0.24)
Total count of participants that skipped at least one life goals item	4 (0.95)

[Response Ends]

2b.10. Provide your interpretation of the results, in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders), and how the specified handling of missing data minimizes bias.

In other words, what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis was conducted, justify the selected approach for missing data.

[Response Begins]

In both the calibration testing and validation testing samples, there was a low level of missingness (i.e., <2% in the calibration sample; <1% in the validation testing sample). Note that, even if one item is missing from the six Likert-type scorable PaLS items, a *t*-score can still be calculated. Low levels of PaLS item response missingness demonstrate that performance results were not biased due to missing data.

[Response Ends]

Note: This item is directed to measures that are risk-adjusted (with or without social risk factors) OR to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eQMs). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]

No, there is only one set of specifications for this measure

[Response Ends]

2b.12. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications.

Describe the steps—do not just name a method. Indicate what statistical analysis was used.

[Response Begins]

[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]

[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]

[Response Ends]

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]

N/A or no exclusions

[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

Describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used?

[Response Begins]

N/A

[Response Ends]

2b.17. Provide the statistical results from testing exclusions.

Include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores.

[Response Begins]

N/A

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion.

[Response Begins]

N/A

[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]

No risk adjustment or risk stratification

[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]

[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]

[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]

[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of $p < 0.10$ or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any “ordering” of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]

[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]

[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]

[Response Ends]

2b.26. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter “N/A” for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]

[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]

[Response Ends]

2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

[Response Begins]

N/A

[Response Ends]

2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]

[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]

[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]

[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]

[Response Ends]

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]

[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.

[Response Begins]

[Response Ends]

3.03. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using data elements not from electronic sources.

[Response Begins]

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

[Response Ends]

3.06. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

[Response Begins]

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, algorithm),

Attach the fee schedule here, if applicable.

[Response Begins]

[Response Ends]

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement, in addition to demonstrating performance improvement.

4a.01. Check all current uses. For each current use checked, please provide:

- ☐ **Name of program and sponsor**
- ☐ **URL**
- ☐ **Purpose**
- ☐ **Geographic area and number and percentage of accountable entities and patients included**
- ☐ **Level of measurement and setting**

[Response Begins]

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]

[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]

[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.

[Response Begins]

[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail how many and which types of measured entities and/or others were included. If only a sample of measured entities were included, describe the full population and how the sample was selected.

[Response Begins]

[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]

[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]

[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]

[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]

[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]

[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included). If no improvement was demonstrated, provide an explanation. If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

[Response Begins]

[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]

[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]

[Response Ends]

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

NOTE: If there are no related measures, please select N/A.

(Can search and select measures.)

[Response Begins]

[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus and target population).

NOTE: If there are no competing measures, please select N/A.

(Can search and select measures.)

[Response Begins]

[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]

[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]

[Response Ends]

5.05. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

[Response Begins]

[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]

[Response Ends]

Appendix

Supplemental materials may be provided in an appendix.:

Available at measure-specific web page URL identified in sp.09

Attachment: PaLS Flowchart

Contact Information

Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Measure Steward Point of Contact: Rawlings, Kimberly, kimberly.rawlings@cms.hhs.gov

Dollar-Maples, Helen, helen.dollar-maples@cms.hhs.gov

Measure Developer if different from Measure Steward: University of Michigan Kidney Epidemiology and Cost Center

Measure Developer Point(s) of Contact: George, Jaclyn, jaclynrg@med.umich.edu

Sardone, Jennifer, jmsto@med.umich.edu

Dahlerus, Claudia, dahlerus@med.umich.edu

Additional Information

1. Provide any supplemental materials, if needed, as an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be collated one file with a table of contents or bookmarks. If material pertains to a specific criterion, that should be indicated.

[Response Begins]

Available at measure-specific web page URL identified in sp.09

[Response Ends]

Attachment: PaLS Flowchart

2. List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

[Response Begins]

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]

[Response Ends]

4. Indicate the month and year of the most recent revision.

[Response Begins]

[Response Ends]

5. Indicate the frequency of review, or an update schedule, for this measure.

[Response Begins]

[Response Ends]

6. Indicate the next scheduled update or review of this measure.

[Response Begins]

[Response Ends]

7. Provide a copyright statement, if applicable. Otherwise, indicate "N/A".

[Response Begins]

[Response Ends]

8. State any disclaimers, if applicable. Otherwise, indicate "N/A".

[Response Begins]

[Response Ends]

9. Provide any additional information or comments, if applicable. Otherwise, indicate "N/A".

[Response Begins]

[Response Ends]