NATIONAL QUALITY FORUM
Driving measurable health improvements together

Measure Worksheet

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.
To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3665
Corresponding Measures:
Measure Title: Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood
Measure Steward: American Academy of Hospice and Palliative Medicine
sp.02. Brief Description of Measure : This is a multi-item measure consisting of 4 items: Q1: "I felt heard and understood by this provider and team", Q2: "I felt this provider and team put my best interests first when making recommendations about my care", Q3: "I felt this provider and team saw me as a person, not just someone with a medical problem", Q4: "I felt this provider and team understood what is important to me in my life."
Per the recommendation of our technical expert clinical user and patient panel (TECUPP), survey items refer to "this provider and team" which reflects the interdisciplinary team structure of care delivery in ambulatory palliative care. Providers can be one of many MIPS-eligible provider types, ranging from Doctor of Medicine to clinical nurse specialists. Providers serve as the lead of the palliative care team and are therefore referenced (i.e., named) at the start of the survey instrument. To identify the reference provider named on the survey instrument for each patient, the data set was first filtered to include only visits with MIPS-eligible provider types that occurred in the three months prior to the anticipated start date of survey fielding. We then selected the MIPS-eligible provider whom the patient saw most often within the three-month period, with ties in numbers of visits broken by provider type, giving preference to providers holding primary responsibility for patient care outcomes (e.g., physician or physician-designee over nurse or therapist). If patients had multiple visits, we selected the most recent visit for each patient with the reference provider.
We did not conduct testing to specifically evaluate how patients differentiated between team members in their responses to the survey items.
1b.01. Developer Rationale: Palliative care has expanded rapidly in recent years, and consensus has been growing within the palliative care community regarding the need for measuring the quality of end-of-life care. Yet the quality of care delivered by palliative care providers (and by other clinicians responsible for seriously ill patients) is unknown, particularly among patients who receive their palliative care early in their disease trajectory in ambulatory settings. As a result, stakeholders – including patients and their advocates, as well as providers and health systems – lack actionable measures to guide improvement efforts, as noted by NQF and the CMS Measures Application Partnership (MAP) as well as the 2017 CMS Environmental Scan and Gap Analysis Report (Centers for Medicare & Medicaid Services, 2017). Measures of palliative care quality are also underrepresented in the CMS QPP, with current measures addressing small populations that are often limited to patients with cancer or hospice patients. Furthermore, palliative care quality assessment that incorporates patient preferences (i.e., patient "voice") is noticeably absent despite the patient-centered nature of palliative care (Anhang Price & Elliott, 2018; Anhang Price et al., 2014; Anhang Price et al., 2018; Teno et al., 2017).

Patient-centered measures, and especially patient-reported outcome measures, are an important complement to clinician-reported measurement data.

The patterns of palliative care received in ambulatory clinics differ substantially from palliative care received in other settings. Ambulatory palliative care typically supplements a primary treating service such as oncology, as needed. Patients may have several visits with different members of the palliative care team, or they may only have a single visit. This variability in the patient experience of palliative care raises important measurement challenges (Chen et al., 2020), which this project seeks to address by developing measures that are both broadly applicable to patients with serious illness, and useful to clinicians and health systems in measuring and improving the quality of care that patients with serious illness receive.

It is important to note that the palliative care field is unique in that palliative care patients are seriously ill, and death is not always a negative outcome, though the quality of that death is important. Accordingly, palliative care requires measures that examine whether patients are receiving care that aligns with their goals, rather than meeting clinical outcomes that may be more appropriate to other conditions, such as mortality (Chen et al., 2020).

Existing evidence and expert consensus have highlighted significant unmet need among seriously ill persons and gaps in meaningful communication measures, despite the noted importance of communication to seriously ill patients and their families (CMS Health Services Advisory Group, 2017). These gaps may be particularly pronounced in ambulatory settings, where patients and families have limited access to palliative care services and may struggle to manage their illness and accept their trajectory. Seriously ill persons often report feeling silenced, ignored, and misunderstood in medical institutions (Frosch et al., 2012; Institute of Medicine & Committee on Approaching Death Addressing Key End of Life Issues, 2015; Norton et al., 2003). Many patients with serious illness experience inadequate communication from their health care providers about prognosis and treatment options (Casarett et al., 2008; Covinsky et al., 2000; Dy et al., 2008; Teno et al., 2004; Teno et al., 2009) and receive care that is not consistent with their preferences (Khandelwal et al., 2017; Teno et al., 2004; Teno et al., 2015; Teno et al., 2009). Systematically monitoring, reporting, and responding to how well patients feel heard and understood is crucial to creating and sustaining a health care environment that excels in caring for those who are seriously ill (Gramling et al., 2016). Communication quality in serious illness comprises at least four mutually reinforcing processes: information gathering, information sharing, responding to emotion, and fostering relationships (Street et al., 2009). These elements directly shape patient experience and, when done well, help patients feel known, informed, in control, and satisfied, thus improving well-being and quality of life (Medendorp et al., 2017; Murray et al., 2015; Street et al., 2009).

During information gathering, to assess the meaningfulness of the measured outcome to the target population, AAHPM conducted one-on-one phone interviews with 13 patients, caregivers, and family members (PCFMs). For these interviews, we sought patients who were currently receiving palliative care and/or hospice or who had received these services in the past, patients with advanced illness who were not currently receiving hospice and/or palliative care services, informal caregivers of patients receiving hospice and/or palliative care services, and patient advocates. PCFMs described how the approach to communication, content, and tone were key components of good communication between providers or teams and palliative care patients. PCFMs felt that directly assessing whether patients feel heard and understood is "very, very important," with some commenting that it makes sense to measure this concept after a visit (Chen et al., 2020).

The proposed measure is also valuable for implementation of innovative payment models for palliative care delivery that impact emerging models of community-based palliative care (e.g., embedded clinic models). Interdisciplinary palliative care team services are often unbillable under a fee-for-service model, and value-based payment models may be an alternative for reimbursement (Center to Advance Palliative Care, 2017). However, innovative financial models require quality metrics to ensure accountability for patients as well as payers and providers (Anhang Price et al., 2018; California Health Care Foundation, 2018). Many emerging models of community-based palliative care are delivered in community settings and may not utilize the same interdisciplinary team nor have the same level of training as programs evaluated in the literature (Teno et al., 2017). Palliative care quality measures would hold programs accountable for quality and would allow providers to demonstrate the value of their services (California Health Care Foundation, 2018). Currently available measures are generally limited to end-of-life utilization and process measures and are not consistently used across programs, thus patient reported quality metrics are needed to assess the impact of community-based palliative care and ensure transparency and accountability for these vulnerable patients (California Health Care Foundation, 2018; Teno et al., 2017).

Citations:

Anhang Price, R., & Elliott, M. N. (2018). Measuring Patient-Centeredness of Care for Seriously Ill Individuals: Challenges and Opportunities for Accountability Initiatives. *J Palliat Med*, *21*(Suppl 2), S-28-S-35.

Anhang Price, R., Elliott, M. N., Zaslavsky, A. M., Hays, R. D., Lehrman, W. G., Rybowski, L., Edgman-Levitan, S., & Cleary, P. D. (2014). Examining the role of patient experience surveys in measuring health care quality. *Med Care Res Rev*, *71*(5), 522-554.

Anhang Price, R., Stucky, B., Parast, L., Elliott, M. N., Haas, A., Bradley, M., & Teno, J. M. (2018). Development of Valid and Reliable Measures of Patient and Family Experiences of Hospice Care for Public Reporting. *J Palliat Med*, *21*(7), 924-932.

California Health Care Foundation. (2018). *Lessons Learned from Payer-Provider Partnerships for Community-Based Palliative Care*.

Casarett, D., Pickard, A., Bailey, F. A., Ritchie, C. S., Furman, C. D., Rosenfeld, K., Shreve, S., & Shea, J. (2008). A nationwide VA palliative care quality measure: the family assessment of treatment at the end of life. *J Palliat Med*, *11*(1), 68-75.

Center to Advance Palliative Care. (2017). *Payment Primer: What to Know about Payment for Palliative Care Delivery*.

Centers for Medicare & Medicaid Services, H. S. A. G. (2017). *CMS Quality Measure Development Plan Environmental Scan and Gap Analysis Report (MACRA, Section 102).* . https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/ValueBased-Programs/MACRA-MIPS-and-APMs/MACRA-MIPS-and-APMs.html.

Chen, E. K., Ahluwalia, S. C., Shetty, K., Pillemer, F., Etchegaray, J. M., Walling, A., Kim, A., Martineau, M., Phillips, J., Farmer, C. M., & Ast, K. (2020). *Development of Palliative Care Quality Measures for Outpatients in a Clinic-Based Setting*. RAND Corporation.

Covinsky, K. E., Fuller, J. D., Yaffe, K., Johnston, C. B., Hamel, M. B., Lynn, J., Teno, J. M., & Phillips, R. S. (2000). Communication and decision-making in seriously ill patients: findings of the SUPPORT project. The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *J Am Geriatr Soc*, *48*(5 Suppl), S187-193.

Dy, S. M., Shugarman, L. R., Lorenz, K. A., Mularski, R. A., & Lynn, J. (2008). A systematic review of satisfaction with care at the end of life. *J Am Geriatr Soc*, *56*(1), 124-129.

Frosch, D. L., May, S. G., Rendle, K. A., Tietbohl, C., & Elwyn, G. (2012). Authoritarian physicians and patients' fear of being labeled 'difficult'among key obstacles to shared decision making. *Health Affairs*, *31*(5), 1030-1038.

Gramling, R., Stanek, S., Ladwig, S., Gajary-Coots, E., Cimino, J., Anderson, W., Norton, S. A., Aslakson, R. A., Ast, K., Elk, R., Garner, K. K., Gramling, R., Grudzen, C., Kamal, A. H., Lamba, S., LeBlanc, T. W., Rhodes, R. L., Roeland, E., Schulman-Green, D., & Unroe, K. T. (2016). Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting. *J Pain Symptom Manage*, *51*(2), 150-154.

Institute of Medicine, & Committee on Approaching Death Adressing Key End of Life Issues. (2015). *Dying in America: Improving Quality and Honoring Individual Preferences Near the End of Life*. National Academies Press.

Khandelwal, N., Curtis, J. R., Freedman, V. A., Kasper, J. D., Gozalo, P., Engelberg, R. A., & Teno, J. M. (2017). How Often Is End-of-Life Care in the United States Inconsistent with Patients' Goals of Care? *Journal of palliative medicine*, *20*(12), 1400-1404.

Medendorp, N., Visser, L., Hillen, M., de Haes, J., & Smets, E. (2017). How oncologists' communication improves (analogue) patients' recall of information. A randomized video-vignettes study. *Patient education and counseling*, *100*(7), 1338-1344.

Murray, C. D., McDonald, C., & Atkin, H. (2015). The communication experiences of patients with palliative care needs: A systematic review and meta-synthesis of qualitative findings. *Palliative & supportive care*, *13*(2), 369-383.
Norton, S. A., Tilden, V. P., Tolle, S. W., Nelson, C. A., & Eggman, S. T. (2003). Life support withdrawal: communication and conflict. *American Journal of Critical Care*, *12*(6), 548-555.
Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, *74*(3), 295-301.
Teno, J. M., Clarridge, B. R., Casey, V., Welch, L. C., Wetle, T., Shield, R., & Mor, V. (2004). Family perspectives on end-of-life care at the last place of care. *Jama*, *291*(1), 88-93.
Teno, J. M., Freedman, V. A., Kasper, J. D., Gozalo, P., & Mor, V. (2015). Is care for the dying improving in the United States? *Journal of palliative medicine*, *18*(8), 662-666.
Teno, J. M., Lima, J. C., & Lyons, K. D. (2009). Cancer patient assessment and reports of excellence: reliability and validity of advanced cancer patient perceptions of the quality of care. *J Clin Oncol*, *27*(10), 1621-1626.
Teno, J. M., Price, R. A., & Makaroun, L. K. (2017). Challenges Of Measuring Quality Of Community-Based Programs For Seriously Ill Individuals And Their Families. *Health Affairs*, *36*(7), 1227-1233.

---

sp.12. Numerator Statement: The *Feeling Heard and Understood* measure is calculated using top-box scoring. The top-box score refers to the percentage of patient respondents that give the most positive response. For all four questions in this measure, the top box numerator is the number of respondents who answer “completely true.” An individual’s score can be considered an average of the four top-box responses and these scores are adjusted for mode of survey administration and proxy assistance. Individual scores are combined to calculate an average score for an overall palliative care program.
sp.14. Denominator Statement: All patients aged 18 years and older who had an ambulatory palliative care visit.
sp.16. Denominator Exclusions:
Denominator exclusions include:

- Patients who do not complete at least one of the four items in the multi-item measure;
- Patients who do not complete the patient experience survey within six months of the eligible ambulatory palliative care visit;
- Patients who respond on the patient experience survey that they did not receive care by the listed ambulatory palliative care provider in the last six months (disavowal);
- Patients who were deceased when the survey reached them;
- Patients for whom a proxy completed the entire survey on their behalf for any reason (no patient involvement).

Measure Type: Outcome: PRO-PM
sp.28. Data Source: Electronic Health Records
sp.07. Level of Analysis: Clinician: Group/Practice

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?

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- This is a patient-reported outcome performance measure (PRO-PM) using instrument-based data at the group/practice clinician level to determines ambulatory palliative care patients' experience of feeling heard and understood.
- The developer presented a logic model linking regular and effective patient-provider communication to patients feeling heard and understood by the ambulatory palliative care provider and team.
- The developer assessed the meaningfulness of the measure via 30-to 60-minute phone interviews with patients, caregivers, and family members, in which participants acknowledged the value of good communication between providers or teams and palliative care patients. Additionally, some participants emphasized the value of measuring the concept of *Feeling Heard and Understood* after a visit.
- The developer cited numerous literatures in support of the measure's goal to improve quality of care processes in palliative care settings:
    - Systematically monitoring, reporting, and responding to how well patients feel heard and understood is critical to ensuring a caring environment for seriously ill individuals (Gramling et al., 2016).
    - The quality of provider communication in serious illness is built on at least four mutually reinforcing processes: information gathering, information sharing, responding to emotion, and fostering relationships (Street et al., 2009).
    - Improving provider communication can improve overall satisfaction with care among patients with serious illness (Anhang Price et al., 2018; Dy et al., 2008).
    - The single item Feeling Heard and Understood data element has been tested and used in palliative care populations (Gramling et al., 2016; Ingersoll et al., 2018).
    - Hospitalized patients with serious illness who received palliative care consultations showed improvement in feeling heard and understood the day after the initial consultation (Gramling et al., 2016; Ingersoll et al., 2018).
    - Provision of a question prompt list to promote discussion about end-of-life issues during palliative care consultations with terminally ill patients and their caregivers led to fewer unmet needs for information without increasing patient anxiety or impairing satisfaction (Clayton et al., 2007).
    - Routine collection of PROs in ambulatory oncology settings with timely feedback to providers can improve patient-provider communication about symptoms and quality of life issues and improve patient satisfaction (Berry et al., 2011; Detmar et al., 2002; Taenzer et al., 2000; Takeuchi et al., 2011; Velikova et al., 2004).

Questions for the Committee:

- Is there at least one thing that the provider can do to achieve a change in the measure results?
- If derived from patient report, does the target population value the measured outcome and find it meaningful?

Guidance from the Evidence Algorithm

Does the measure assess performance on a health outcome (Box 1) -> (yes) -> Is there a relationship between the measure and at least one healthcare action is demonstrated by empirical data (Box 2) -> (yes) -> PASS

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- This measure was tested within 44 ambulatory palliative care programs stratified by administrative home type (i.e., hospice, hospital, ambulatory, and other administration) and by geographic location to ensure representation across Census Regions.
- The range in adjusted program scores was from 54.05 to 85.18 with a standard deviation of 7.04 thus suggesting room for improvement. [Lowest Program CI: (42.2, 66.0); Highest Program CI: (77.0, 91.1)].
- The developer cited literature highlighting the variability in care received in ambulatory clinics, which necessitates measures that are both broadly applicable to patients with serious illness, and useful to clinicians and health systems in measuring and improving the quality of care in palliative care settings.
- The developer emphasized the measure's value in implementing innovative financial models for palliative care delivery that impact emerging models of community-based palliative care (e.g., embedded clinic models).

Disparities

- The developer evaluated the relationship of various social risk factors (patient race/ethnicity, education, primary language, urbanicity, median household income, gender, marital status, public insurance use, unemployment, and families below poverty line) to the measure score and the programs.
- After adjusting for multiple comparisons, the developer did not identify a significant relationship between the variables and the measure.
- The developer noted two conflicting studies:
  - Patients belonging to a racial/ethnic minority group, lower income, and higher religiosity reported higher ratings of clinician communication (Coats et al., 2018).
  - Patients from racial/ethnic minority groups, patients with lower income, and patients with lower educational attainment gave physicians in training higher ratings on end-of-life care communication; however, family members of non-white patients gave trainees lower ratings on communication (Long et al., 2014).
- As an explanation for the conflicting findings of the studies above, the developer suggested that various patient characteristics and contextual factors may impact their experience of provider communication.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- Agree that the evidence demonstrates that the target audience values the outcome being measured. The evidence relates directly to the measure. values the measure. The measure relates directly to the desired outcome.
- Patient reported outcome, direct
- Good evidence to support this measure
- The evidence is adequate
- Numerous literatures cited.
- The developer provided a robust literature review that documents the importance of communication between patients and palliative care providers.
- Evidence provided to support measure
- They stated that some participants emphasized the value of measuring the concept of Feeling Heard and understood.

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- There is a performance gap that warrants a national performance measure; this measure assures maintenance of a desired outcome overtime even without a gap
- Yes, large gap from 54.05 to 85.18 with multiple conflicting reasons for gap
- Yes, gap is present, room for improvement
- Yes, the gap was identified, and data provided
- Data provided. Evidence of a performance gap noted. Multiple comparisons made based the social risk factors: patient race/ethnicity, education, primary language, urbanicity, median household income, gender, marital status, public insurance use, unemployment, and families below poverty line. Reported some conflicting data related to racial/ethnic minority groups, and lower income.
- There is wide variability in patient experiences with palliative care. Patient characteristics and the social context can impact expectations of providers and patient interactions with them.
- Would provide data on quality of care from perspective of the patient and permit quality improvement. In alignment with CMS Quality Performance/Payment program
- They cited the fact that there are few if any measures in the ambulatory setting, which demonstrates a need. The developer evaluated several social risk factors, but they did not identify a significant relationship between the variable and the measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Complex measure evaluated by Scientific Methods Panel? ☒ **Yes** ☐ **No**

Evaluators: Alex Sox-Harris, Sam Simon, Zhenqiu Lin, Larry Glance, Matt Austin, Terri Warholak, Jeffrey Geppert, Christie Teigland, Eugene Nuccio, Lacy Fabian, Marybeth Farquhar, Joseph Kunisch

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:
This measure was reviewed by the Scientific Methods Panel (SMP) and passed both reliability and validity evaluations by the SMP Subgroup. It was not discussed during the SMP measure evaluation meeting. A summary of the measure and the Panel subgroup evaluation is provided below.

Specifications

- Two SMP subgroup members requested additional clarification for the specification language. One panel member asked for increased description language, such as the target population, age range, survey questions, time-period.
- Multiple panel members were concerned with potential measure attribution misalignment as an assessment of the provider, rather than the patient outcome in the PRO-PM. They specifically note the clinician/group level of analysis with a numerator stating the accountable entity as an "individual" provider. One noted that patients will see multiple providers within six months of the denominator timeframe. They also note that the survey does not specifically identify an anchor patient visit, as the measure allows patient who transfer to home-based hospice also to reflect on their ambulator hospice care.

Reliability

- Reliability testing conducted at the Patient or Encounter level:
    - The multi-data four-question scale was evaluated with Cronbach's alpha with an acceptable threshold of 0.7. The four-data elements of the Feeling Heard and Understood scale Cronbach's were 0.90.
    - A test-retest reliability coefficients calculation of live phone respondents in high-volume ambulatory programs were given a shortened computer-assisted telephone interview (CATI) survey within 48 hours of the first survey. The result from the polychoric correlation

coefficient was 0.85 for the CATI data collection method. Agreement statistics were not provided.

- Reliability testing conducted at the Accountable Entity level:
  - ○ Using a signal-to-noise analysis, accountable entity testing was conducted to assess between- (i.e., signal) and within- (i.e., noise) subject variability to discriminate provider performance.
  - ○ Developers used hierarchical generalized-linear regressions to decompose variability of binomial outcomes to programs, and to covariates with the data hierarchy as patient observations. The variance of the model can be decomposed using the (adjusted) ICC, which provides a summary of the reliability of the measure as tested, with higher values implying more variability between programs. Using Bayesian generalized mixed-effects models obtained a posterior distribution of the adjusted ICC with estimates of 0.052 (95% CI: 0.027 to 0.089). The SMP members acknowledge that testing during the COVID-19 pandemic may have affected changes in palliative care services and experiences.
  - ○ For projected to observed variance from within each program, Spearman-Brown prophecy formula was used to determine reliability results to future samples. To obtain a result of 0.7 or higher, an average of 45 eligible and complete returned responses were required. Assuming high correlation between the four survey questions, as 3.25 estimated design effect from repeated measure, an average sample size of 37 eligible and complete respondents would be required. The 3.25 estimated design effect method description is not clear.
  - ○ To assess the average adjusted reliability of individual programs, developers estimated a posterior distribution for the overall variability using an Adams-like (2009) approach, which demonstrated an average reliability across programs of approximately r = 0.752.

Validity

- Validity testing conducted at the Patient or Encounter level:
  - ○ Convergent validity testing was used for patient- encounter-level validity testing hypothesizing the relationship so similar constructs, including data elements from other instruments: 1) Consumer Assessment of Healthcare Providers and Systems (CAHPS) Hospice and 2) Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain
    - Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain is a new measure developed by the same developer that is currently being reviewed by NQF. Developers hypothesized feeling heard and understood would correlate to getting help for pain needs from the palliative care team.
    - Interpretation of the bivariate correlation followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50) (Rosnow & Rosenthal, 1989).
    - The Feeling Heard and Understood scale was associated with higher CAHPS communication scores (r = 0.54, p<.001). For Receiving Desired Help for Pain, the correlations were weak/low (r = 0.48, p< .001).
- Validity testing conducted at the Accountable Entity level:
  - ○ To assess accountable entity level validity, measure scores examined the association of the measure scores to 1) Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain, 2) the CAHPS communication measure score, and 3) the individual's overall rating of their palliative care provider and team. The developer hypothesized these scores would be positively associated to Receiving Desired Help for Pain.
    - The measure was positively associated with the CAHPS communication quality measure (r = 0.635, p =0.011), the Receiving Desired Help for Pain quality measure (r = 0.496, p<.001) and the overall rating of the palliative care provider and team (r= 0.768, p=<.001) with associated to other similar measures (r = 0.5 – 0.8). Positive

correlations between 3665 and 3666 Receiving Desired Help for Pain were moderate low.

- Face validity was assessed with a panel of seven palliative care communication experts who assessed the final measure specifications and testing results and rate the measure's ability to distinguish quality palliative care. Face validity ratings were from 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9). The average face validity ratings of the measure score were 8.3 which corresponds to a developer defined average rating of "high."

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- I have concerns. The methodology of the process of identifying the specifications does not seem totally clear or reliable. I have concerns about the denominator as well as the numerator determination. I am unclear how the data is collected
- Passed scientific methods panel
- Concern for communicating with people from cultures that are not so direct.
- Some of the reviewers noted concerns and there seemed to be a wide range of opinion on this that may be hard to reconcile
- Multiple approaches to reliability testing done. Most demonstrate average/acceptable reliability.
- As the SMP indicated, there may be differences in individual vs. group evaluation without an anchor visit. I have no other concerns about reliability.
- Need to define provider and concerns over multiple "providers" of care being involved during the care period. Concern also with not including patients where palliative care is being provided because of a dementia diagnosis because the survey tool excluded patients if a proxy was needed to respond
- None

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- I need more information and clarification on the specifications in order to give an opinion on the reliability of the measure as currently designed
- Passed scientific methods panel
- See above regarding cultures.
- Some reviewers noted concerns here and, again, there seemed a wide range of opinions on this issue

- No
- None
- I share concerns raised by the reviewers as noted above
- No

2b1. Validity -Testing: Do you have any concerns with the testing results?

- No
- No concern
- Only concern regarding different communication styles/cultures.
- Same as above, wide range of concerns by reviewers
- Face validity tested with palliative care communication experts - with average rating of "high". Other validity test comparing result to CAHPS communication measure score, rating of PC provider and team and Ambulatory Palliative care patients' experience of receiving desired help for pain. Moderate low correlation between 3665 and 3666. No major concerns
- No
- Validity appears to be reasonable
- No

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- I need more information. Unclear how it is determined that a proxy has filled out the assessment for the patient rather than the patient filling it out, and is it always inappropriate for a proxy to fill out the form? Patient's level of sophistication and experience with the health care system may be a concern as well
- No concern
- No concerns.
- Again, wide range of concerns by reviewers and I lack expertise to have an informed opinion
- The risk adjustment uses hierarchical generalized-linear model approach using a 12-month period for data collection collected during four three-month periods of Provider-Patient interaction. -- I cannot comment of whether this is the best/appropriate approach.
- The exclusions seem appropriate. I have no concerns with the Risk Adjustment.
- Concerns over patients being excluded if have a proxy/legally responsible party who is intimately involved in their care and could respond to the questions/measure.
- No exclusions are apparent. Yes, everything appears to be present.

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- If I understand it correctly, I would think that missing data would greatly impact the validity of the measure for determining the performance of an ambulatory palliative program; some data points may not be identified. I am unclear about how data is collected consistently
- No concern
- Could have non-response bias. Negative responders might be more likely to complete the survey

- Range of concerns on these points by reviewers, I lack the expertise to determine which ones to focus on
- Only moderate low correlations between 3665 and 3666. Also Receiving Desired Help for Pain, the correlations were weak/low (r = 0.48, p< .001).
- My only concern is that the developer states that feeling heard and understood would correlate to getting help for pain needs. This seems short sighted--there are emotional, psychosocial, existential, and spiritual ways that patients need to be heard and understood.
- Concerns regarding the potential for missing data
- The measure seems to be valid and repeatable, there does not appear to be any missing data.

Criterion 3. feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the 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.

- Patient-reported data is collected via survey instrument. The instrument was developed for this measure and can be completed via web survey, on paper or over telephone in English.
- Patient eligibility is determined based on coded visit information in the electronic health record.
- This is a patient-reported measure of experience that is not currently collected in structured electronic fields or electronic medical records-based clinical data fields.
- The majority of respondents to the 2021 public comment period supported feasibility of the proposed measure; 21.8% of respondents said, “very feasible” and 42.7% said “somewhat feasible” when asked “How feasible would it be to implement these measures (e.g., contracting with a survey vendor, identifying eligible patients through administrative or medical record data, submitting scores to CMS, etc.)?”
- No fees or licensing requirements will be necessary for users to implement the proposed measure. However, implementation costs include the cost of hiring an authorized survey vendor to field surveys and process data.

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form, e.g., EHR or other electronic sources?
- Is the data collection strategy ready to be put into operational use?

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- I need more clarification on this point as this is variable among EHR in different programs
- Cost concern for feasibility for field surveys
- No concerns.
- This seems feasible for the patient population/setting
- The majority of respondents to the 2021 public comment period supported feasibility of the proposed

measure; 21.8% of respondents said, “very feasible” and 42.7% said “somewhat feasible” collected electronically, no fees/licensing
- Data collection seems straightforward using the EHR. The implementation costs may be prohibitive for some organizations.
- Appears to meet feasibility requirement
- Everything seems to be in place, the only concern is the cost of hiring a vendor to field the surveys and process the data, which should not cause a problem.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The measure is not currently in use but has been submitted to the 2021 MUC list for inclusion into CMS’ Quality Payment Programs, including MIPS and APMs.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured, and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- Performance data from the developer’s test has been provided to all participating programs as well as key stakeholder groups, e.g., the technical expert clinical user and patient panel (TECUPP) and project advisory panel.
- All programs that participated in the beta field test will receive a summary report describing their performance on each survey item as well as their performance on the measure.
- Based on feedback from alpha pilot test programs, the summary reports were refined to better suit the needs of programs that participated in the beta field test.
- The developer obtained feedback on potential implementation challenges and usefulness of the proposed measure for quality improvement during the 2021 public comment period.

Additional Feedback:

N/A

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- This is a new measure and not currently in use in any quality improvement programs.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- Repercussions of negative feedback; concerns that some patients may have unrealistic expectations for palliative care, and patients whose expectations are not met may identify as not being heard and understood.
- Providers of palliative care often have to convey bad news to their patients, which may negatively impact patient perceptions of the team.

Potential harms

- The developer did not identify any potential harms.

Additional Feedback: N/A

Questions for the Committee:

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- My understanding is that this is a new measure and is not yet publicly reported; these being measured have been given feedback and help with interpreting the measure results, I am not sure if the feedback from those being measured has been utilized when making changes to the measure.
- Useful to distinguish specialty of palliative care
- Reviewed by small patient/caregiver group; 44 PC programs
- TBD as new measure
- Performance data has been provided to participating programs. Reports refined after alpha pilot test; developer has obtained feedback on use.
- This new measure is not currently in use thus not publicly reported but planned for use in an accountability program.
- Not publicly reported currently. Plans for reporting in future in quality improvement programs
- The measure is not publicly reported, there is planned use in accountability program. Those being measured, and others, have been given the opportunity to provide feedback which is incorporated into the measure.

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- A credible rationale has been provided that describes how the performance results could be used to further the goal of high quality, efficient healthcare. with regards to benefits vs. harms, I think that there could be unintended harms if the specifications of the measure are not refined so as to be reliable measures of quality, and that this measure may not be the only indication of quality and value in a PC program, but currently is the only one we will have, and therefore carries a lot of weight on its shoulders and the value it is given.
- Unintended consequences from hearing bad news.
- Providers need to ensure they are communicating in acceptable ways to patients.
- New measure, not currently in use. Negative feedback may reflect the situation more than the providers. "Top box" scoring is a high bar-but may be warranted.
- The only unintended consequence may occur without the identification of an anchor visit--results could become clouded if patients see multiple providers.
- Agree with reviewer with moderate rating for usability and use.
- No harms were identified but a concern about the patient's expectations and hearing bad news could negatively influence their responses.

Criterion 5: Related and Competing Measures

Related or competing measures

- 2651: CAHPS® Hospice Survey (experience with care)

Harmonization

- The developer states that the measure specifications have been harmonized.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- CAPHS survey related and harmonized.
- Yes, but different population.
- No, this fills an important existing gap.
- 2651: CAHPS® Hospice Survey (experience with care) - developer states that the measure specifications have been harmonized.
- 2651-CAHPS Hospice Survey but the developer states that the specifications have been harmonized.
- CAHPS survey. Measure developer noted harmonization however this measure specific to palliative care ambulatory programs.
- No.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/19/22

- **Of the one NQF member who has submitted a support/non-support choice:**
  - **One supports the measure**
  - **Zero do not support the measure**
- Comment by: **American Academy of Hospice and Palliative Medicine**

These comments are in response to SMP review.

- **Issue 7: Cognitive Testing**

  R6: Was cognitive debriefing done with patients before the measure was tested? I have a few issues with the survey items. Specifically, item #2 is double barreled, and research indicates that this leads to measurement error.

- **Developer Response 7:**

  We conducted 25 one-hour telephone cognitive interviews using a convenience sample of outpatient palliative care patients and caregivers to cognitively test survey items, with positive results. Participants generally understood the intended meaning of the question content. Some changes were made to improve the clarity of specific items. (See published manuscript: Rollison et al, Incorporating the Patient and Caregiver Voice in Palliative Care Quality Measure Development, Journal of Pain and Symptom Management, 2021)

  In particular, the “feeling heard and understood” concept was generally well-understood in its intended meaning as validation and acknowledgement from one's provider. It was determined necessary the two words – “heard” and “understood” together, because when asked separately, interviewees mistakenly understood the terms to refer to hearing (auditory ability) and comprehension (cognitive ability). This confusion also arose in early work to develop the single-item for use in inpatient settings (see: Gramling R et al, Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting, Journal of Pain and Symptom Management, 2016), reinforcing our decision to use both words together to represent the single construct of feeling seen, respected, acknowledged.

- **Issue 8: Communicating scores to providers**

  R9:Will there be any effort to communicate to the Provider the Top Box score on each of the four items so that the Provider can take a targeted intervention? The average score, while informative, does not provide the opportunity to make a targeted intervention.

- **Developer Response 8:** AAHPM will consider this option and if it’s feasible, we will strive to provide the opportunity for targeted intervention

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

- **Issue 3: Risk Adjustment**

  R3, R4: The risk model seems overly simplified, there are many factors that should have been looked into and potentially included, for example, administrative home type, disease status and others; Considered only a small number of patient level risk factors; lack of risk adjustment for patient level factors. Although I understand that this is because of lack of patient-level data on risk factors, this is not an "excuse" for the lack of risk adjustment.

- **Developer Response 3:**

  Using the data available to us (which was limited in terms of what programs were able to provide to us, and how much we could reliably capture via survey-based self-report), we did explore some potential program- and patient-level risk adjustment factors.

  Only survey mode was significant in its relationship with the HU performance measure (p = 0.013) and with programs (p = 0.001) after adjustment for multiple comparisons.

  At the patient-level, a single data element (“I felt this provider and team understood what is important to me out of life”) of the four Feeling Heard and Understood data elements was significantly associated with diagnosis group (p < 0.01), and the raw measure score was significantly associated with diagnosis group. These results held after multiple comparison adjustments. Because of challenges with data quality, we were unable to conduct further analyses within the scope of this effort, but these findings provide preliminary indication that diagnosis might affect responses to the performance measure data elements and overall measure performance. We acknowledge the importance of further research in this area before the measure is used for high-stakes decisions.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

Reliability

- **Issue 3: Measure Score Calculation**

  R3: Additionally, the developer should clearly describe how the measure score is calculated. Specifically, how the results from the hierarchical logistic regression model are rolled up to a measure score. The model specified seems to be a 2-level model, what's unit of analysis? Survey item level response (which is binary) or patient level score (which is not binary)? Measure score reliability was assessed via hierarchical logistic regression model, although it is not clear how it was done, at survey item response level or patient score level?

- **Developer Response 3:**

  The Feeling Heard and Understood measure score is calculated using a hierarchical (two-level) risk-adjusted binomial model (see mathematical details below). The scores are rolled up using a set of baseline characteristics (in this case proxy assistance status and survey mode) such that each provider has the same set of characteristics. The patient is the unit of analysis. We use a binomial model because each respondent contributes two pieces of information: 1) the number of responses provided across the four Feeling Heard and Understood items; and 2) the number of top-box responses. The individual’s score on this measure is the proportion of top-box responses on these four items, i.e., a set of n = 4 trials with probability p of success. The average score is the estimated p (that as the reviewer notes, is not binary).

  More mathematically, our measure assumes that within provider i for each individual j, the k = 1,2,3,4 questions that they respond to are from the following parametric distribution, Y(subscript ij) ~ Binomial(n[subscript ij], p[subscript ij]) where n(subscript ij) = Sigma(subscript k)R(subscript ijk) <= 4 where R(subscript ijk) is one if a question k is responded to and zero otherwise. Thus, the unit of analysis is the patient-level, n(subscript ij) is the number of questions that an individual responded to, and p(subscript ij) represents an individual's average number of top-box responses on the four items. Explicitly, the individual's score arises as a non-continuous value because we have up to four binary outcomes that are contributing to the likelihood function.

Let P(subscript i) represent an indicator that individual j received care from provider i, X(subscript ij) represents the patient's characteristics, then the risk-adjusted model for a provider score assumes the following generalized linear model logit(E[Y(subscript ij)|X(subscript ij), a, standardized beta]) = logit(p[subscript ij]) = (standardized beta[subscript 0] + b[subscript i]P[subscript i]) + X(superscript T, subscript ij)a with an assumption that b(subscript i ~ N(0, lowercase omega[superscript 2, subscript b]). In this model, standardized beta(subscript 0) represents the average score across providers (i.e., grand mean), b(subscript i) is the difference between the average program score across providers (higher values represent better than average care) and a specific provider i's score, and a are risk adjusted coefficients.

To calculate a specific providers score, let X* be a set of "baseline" characteristics to standardize an individual provider's score against, in our example, the characteristics were a fixed survey mode and no proxy assistance. The score for provider i is estimated using the following p-hat(subscript i) = logit(superscript -1)((standardized beta[subscript 0] + b[subscript i]) + X*[superscript T]a)

In our specific submission X* was set to zero (indicating the baseline survey mode and no proxy assistance) and therefore the adjusted score is: p-hat(subscript i) = logit(superscript -1)(standardized beta[subscript 0] + b[subscript i])

Hopefully this clarifies both our model and the estimation of the provider risk-score.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

- **Issue 6: Sampling**

  R6: Also, on page 35 it is indicated that data should be collected from "eligible palliative care patients that are representative of the palliative care provider program." This indicates to me that some sampling technique is used but up to this point in the application I thought the practice would send data on all of the patients who met the criteria - not sample. This is an easy fix and just needs a clarification.

- **Developer Response 6:** Depending on the volume of patients and to support feasibility for programs, palliative care practices may survey all eligible patients or a *random* sample of eligible patients. The target population for sampling includes patients aged 18 years or older who received ambulatory palliative care services from a MIPS-eligible provider within the three months prior to the start of survey fielding. Findings from the alpha pilot test and beta field test support the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. The provider or program will provide a vendor with an extract file of all patients who received care during the measurement period. To prevent gaming and to minimize administration and social desirability bias, the vendor will apply the eligibility criteria to identify the patient sample and field the survey to eligible patients.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

- **Issue 5: Data Element Reliability**

  R3:The developer ascertained both internal consistency and test-retest reliability for data elements. Each survey item has 5 response categories, however, for the measure, top box scoring is used. Therefore, the developer needs to clarify if the testing was consistent with the top box scoring approach; Data element reliability testing needs to be consistent with the top box scoring approach.

- **Developer Response 5:** To clarify, reliability of the four-data element scale using all 5 categorical options was high (Cronbach's alpha = 0.90) and similarly high for the dichotomous top-box option (Cronbach's alpha = 0.84).

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to a concern raised by one of the Scientific Methods Panel reviewers. Reliability

- **Issue 1: Attribution**

  R3: My main concern is with the potential misalignment of provider attribution and patient-reported outcome attribution. Provider was identified based on a three-month period, MIPS-eligible provider who the patient saw most often during the three-month period. However, the attached survey form refers to "the last 6 months". Given that provider who the patient saw most often in the 3-month period may not be the same one in the 6-month period, and it is quite likely that patient might have seen multiple providers during the 6-month period. Therefore, this may potentially cause provider misattribution. To further complicate things, the survey form does not identify the eligible ambulatory palliative care visit, so there is no explicit anchor visit for the patient to refer to even though the developer referred to the eligible ambulatory palliative visit repeatedly in this application, for example, the developer mentioned that patients who had transitioned to hospice could still answer the survey by reflecting on their experience with the visits.

- **Developer Response 1:** Our eligibility and sampling procedures, informed by input from our TECUPP, was designed to reduce the potential for misattribution as much as possible, while enhancing patient recall and their evaluation of the care they received from the palliative care provider and team.

  From the data files outpatient palliative care programs sent us, we first filtered to include only visits with Merit-based Incentive Payment System (MIPS)-eligible provider types that occurred in the three months prior to the anticipated start date of survey fielding (i.e., the planned date for mailing the prenotification letter to patients). We limited to 2019 MIPS-eligible providers so that these measures could be used for MIPS reporting). We limited eligible visits to a three-month period to ensure the recency of the visit patients should consider when responding about their experience. Setting this time frame also allowed each program’s “clock” to start at the same time.

  We then identified a reference provider to be named on the survey instrument for each patient by selecting the MIPS-eligible provider whom the patient saw most often within the three-month period, with ties in numbers of visits broken by provider type, giving preference to providers holding primary responsibility for patient care outcomes (e.g., physician or physician-designee over nurse or therapist). If patients had multiple visits, we selected the most recent visit for each patient with the reference provider.

  The survey instrument included additional protections against misattribution. In both the survey cover letter as well as the instrument itself, we name the provider and team (e.g.: “Dr. Jones and team”). We included mention of the “team” because palliative care is an interdisciplinary team effort, and we anticipated that many patients would have seen the primary provider as well as other palliative care team members across and within visits that they had in the 3-month period. By naming the specific palliative care provider seen most often during the 3-month period, we hoped to avoid confusion with other providers outside palliative care that the patient might have seen.

  The survey instrument refers to a 6-month timeframe rather than the 3-month visit eligibility timeframe to cover potential lags in timing between when the palliative care program sent their data files, and when the survey was fielded and ultimately reached the patient.

  As an example, a program might have submitted a data file to us on September 1st, 2019, covering visits from March 1st through August 31st, 2019. We would sample visits June through August 2019 (the most recent 3 months of data), and field the survey September 25th (once all data files had been cleaned and prepared). The patient might then receive/open the survey on October 1st, 2019. Referring to a 6-month timeframe (rather than a 3-month timeframe) thus covers the full sampling timeframe of June-August 2019.

  Guided by input from our TECUPP, we did not anchor the survey instrument to a specific single visit. Rather, we intentionally wanted patients to reflect on their experience of palliative care as a whole, rather than segmented into what happens in just a single visit, because palliative care as a discipline is intended to be holistic and comprehensive, with a longer-term care relationship. As such, the proposed measures reflect the experience of care over time and cannot be justifiably assessed after a single visit. For example, ensuring that a patient receives the help that they desire for their pain necessarily takes place over time rather than in a single visit.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to an SMP member's concerns.

- **Issue 2: Proxy Response**

  R6: Also, it is stated throughout the application that responses completed by a proxy with our assistance from the patient will be excluded. I'm assuming (perhaps wrongly) that question 10 of the survey (option 3 - Answered the questions for me) will be used to determine this. If that is the case, I have an issue with this as I would not understand that response to indicate no patient involvement. Thus, I feel like this question needs to be re-worked. Also, it is indicated through the application that surveys that were completely filled out by a proxy are excluded. However, it is unclear to me how this would be identified. I'm assuming (perhaps wrongly) that survey question 11 is used for this purpose and that option "answered the questions for me" is used to signify that the patient was not involved. However, I find this option unclear, and I would not have understood it to indicate that the patient was not involved. Thus, I think this item needed to be re-worked to increase clarity before use.

- **Developer Response 2:**

  We excluded from the denominator patients for whom a proxy completed the entire survey on their behalf for any reason i.e., with no patient involvement, (proxy-only responses), but retained proxy-assistance responses, adjusting slightly upward for the latter in our measure scoring procedure, as indicated by our risk adjustment analysis.

  We defined “proxy-only” as the response option “answered the questions for me” to the question “How did that person help you complete the survey?”. This was the only response that indicated that the proxy actually provided the answers to the questions. Based on cognitive interviews and TECUPP input, we felt comfortable that this response option was indicative of no patient involvement. In contrast, we defined “proxy-assistance” as any or all of these responses: “read the questions to me”, “wrote down the answer I gave”, “translated the questions into my language; “helped in some other way”. Further work could reinforce these distinctions and identify slight revisions to increase clarity; the work done to date provides general support for the language currently used.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

- **Issue 3: Measure Score Calculation**

  R3: Additionally, the developer should clearly describe how the measure score is calculated. Specifically, how the results from the hierarchical logistic regression model are rolled up to a measure score. The model specified seems to be a 2-level model, what's unit of analysis? Survey item level response (which is binary) or patient level score (which is not binary)? Measure score reliability was assessed via hierarchical logistic regression model, although it is not clear how it was done, at survey item response level **or patient score level?**

- **Developer Response 3:**

  The Feeling Heard and Understood measure score is calculated using a hierarchical (two-level) risk-adjusted binomial model (see mathematical details below). The scores are rolled up using a set of baseline characteristics (in this case proxy assistance status and survey mode) such that each provider has the same set of characteristics. The patient is the unit of analysis. We use a binomial model because each respondent contributes two pieces of information: 1) the number of responses provided across the four Feeling Heard and Understood items; and 2) the number of top-box responses. The individual’s score on this measure is the proportion of top-box responses on these four items, i.e., a set of n = 4 trials with probability p of success. The average score is the estimated p (that as the reviewer notes, is not binary).

  Please see mathematical calculations and equations provided separately to NQF staff due to inability to copy them in this form.

- Comment by: **American Academy of Hospice and Palliative Medicine**

This comment is in response to SMP review.

- **Issue 4: Measure Score Calculation**

R1:(Re: reliability testing) I was unclear if the hierarchical model accounted for nesting within patient and facility. It looks like items (not patient scores) were the level of analysis, but without accounting for the nesting within patient? If the patient score was the level of analysis, then I don't understand the model form (logistic); Entity-level testing revealed good signal to noise reliability, but I'm a little unclear how the beta binomial distribution was used when the patient score is a proportion (not binomial).

- **Developer Response 4:**

  We believe we understand where the confusion arises. The model is estimated at the patient-level, where a patient's score is a summary of the four binary items. We use a binomial model where the number of trials (i.e., n) are the items and the outcome is the proportion (i.e., p) of top-box (binary) responses to these items and represents an individual's average top-box response. We are considering an individual's score as a simple average and not explicitly modeling an individual effect for items (i.e., no nesting). Where there might be confusion is that under a binomial model, the form of the model is very similar to a standard logistic regression, but the number of trials (i.e., n) is included in the actual estimation (see the probability distribution here https://en.wikipedia.org/wiki/Binomial_distributionand when n = 1 we get back what is normally the logistic regression model for binary outcomes). There are more details in the previous response that explicitly mathematically describes the model that was estimated.

  We also should make clear that the model in the previous response is not exactly the same as the traditional beta-binomial model (https://en.wikipedia.org/wiki/Beta-binomial_distribution) because we do not place beta priors on each individual's probability of success. Beta-binomial models have historical relevance in Bayesian estimation because of computational tractability, but recent software (e.g. the Stan programming language for Bayesian models https://mc-stan.org/) have made alternative models possible. Our model is a hierarchical generalized linear model where we assume a linear form on the probability of success to perform risk-adjustment and differs somewhat in structure from the beta-binomial but achieves the same effect.

- Comment by: **American Academy of Hospice and Palliative Medicine**

These comments are in response to SMP review.

Validity

- **Issue 1: Non-Response**

  R1, R3:am concerned about survey non-response. Although not very large, there is variation in non-response between programs and demographic differences between responders and non-responders. I'm curious is the former is related to the latter. Are there better methods to account for survey non-response than just ignoring it? Nonresponse bias needs to be addressed with known differences between respondents and non-respondents.

- **Developer Response 1:**

  Of the 7,595 surveys we fielded, 2,804 were included as cases for analysis. Another 1,435 were deemed to be ineligible for the measure (e.g.: patient had died or disavowed the reference program or provider) and are thus not considered non-responders.

  Of the remaining 3,356 non-responders (i.e., surveys sent to presumably eligible patients but not returned to us), the majority (80%) were not reachable: 63% were not reachable after the maximum 8 phone call attempts and 17% had non-working phone numbers). Of note, another 14% were reachable but refused to complete the survey.

  As prior survey research has established, it is likely that people who do not return or respond to surveys are systematically different than those who do. This is particularly likely among respondents who explicitly decline or refuse to answer the survey. Our data suggest that survey respondents were slightly older than nonrespondents (mean age 63.4 versus 60.9; $p < 0.01$). The proportion of women was also higher among respondents as compared with nonrespondents (56.2 percent versus 54.5 percent), but the difference was not statistically significant ($p = 0.21$). Although information on patient race was self-reported via the survey instrument, a subset of 12 participating palliative care programs provided patient race for at least 90 percent of their patients in their submitted data files. Among this subset, there was a greater proportion of White patients (88.1 percent versus 80.2 percent) and a lower proportion of Black patients (8.8 percent versus 11.9 percent) in the respondent group compared with the nonrespondent group. The results of a chi-squared test indicate that this difference is statistically significant ($p < 0.01$).

Because the non-responders did not return a survey, we were unable to compare differences in measure scores between them and responders. Although outside the scope of this initial testing effort, future work could attempt to explore other differences between these two groups, for example, to qualitatively understand whether their care experiences differed, in order to shed light on potential response bias.

- **Issue 2: Telehealth**

  R6: I think Telehealth visits should be considered for inclusion in the future. R6, others: Concern about the exclusion of telehealth visits, should be included in the future

- **Developer Response 2:** We strongly agree that telehealth visits should be considered for inclusion in the future. Although we explored the inclusion of telephone and video visits as eligible visits at the outset of our alpha test, we decided not to include those visits because of their low frequency and difficulty identifying these visits. Thus, our initial performance measure eligibility criteria relied on coding in-person office visits. However, because of the COVID-19 pandemic, we were faced with an unexpected situation when participating palliative care programs shifted rapidly to providing telehealth services for their patients. With the input of our TECUPP and project advisory group, as well as input from participating programs, we decided to continue to disallow telehealth visits as eligible for the performance measure when we restarted data collection from September 2020 to February 2021. This ensured consistency in our results (i.e., we were measuring patient experiences with only in-person visits throughout the national beta field test) and avoided any potential confounding effects of the pandemic and telehealth use. However, it is likely that telehealth visits will continue in greater frequency than before the pandemic and should be included in measurement programs in the future. In interviews we conducted with palliative care programs during our testing phase, though most programs had little to no experience with telehealth prior to the pandemic, all programs converted to telehealth after March 2020 and continue to sustain telehealth services in some form. Closer attention to the development and testing of these and other patient experience measures within a telehealth context is warranted prior to widespread use in accountability programs.

- **Issue 3: Risk Adjustment**

  R3, R4: The risk model seems overly simplified, there are many factors that should have been looked into and potentially included, for example, administrative home type, disease status and others; Considered only a small number of patient level risk factors; lack of risk adjustment for patient level factors. Although I understand that this is because of lack of patient-level data on risk factors, this is not an "excuse" for the lack of risk adjustment.

- **Developer Response 3:**

  Using the data available to us (which was limited in terms of what programs were able to provide to us, and how much we could reliably capture via survey-based self-report), we did explore some potential program- and patient-level risk adjustment factors.

  Only survey mode was significant in its relationship with the HU performance measure ($p = 0.013$) and with programs ($p = 0.001$) after adjustment for multiple comparisons.

  At the patient-level, a single data element ("I felt this provider and team understood what is important to me out of life") of the four Feeling Heard and Understood data elements was significantly associated with diagnosis group ($p < 0.01$), and the raw measure score was significantly associated with diagnosis group. These results held after multiple comparison adjustments. Because of challenges with data quality, we were unable to conduct further analyses within the scope of this effort, but these findings provide preliminary indication that diagnosis might affect responses to the performance measure data elements and overall measure performance. We acknowledge the importance of further research in this area before the measure is used for high-stakes decisions.

- Comment by: **American Academy of Hospice and Palliative Medicine**

These comments are in response to SMP review.

Validity

- **Issue 1: Non-Response**

  R1, R3:am concerned about survey non-response. Although not very large, there is variation in non-response between programs and demographic differences between responders and non-responders. I'm curious is the former is related to the latter. Are there better methods to account for survey non-response than just ignoring it? Nonresponse bias needs to be addressed with known differences between respondents and non-respondents.

- **Developer Response 1:**

  Of the 7,595 surveys we fielded, 2,804 were included as cases for analysis. Another 1,435 were deemed to be ineligible for the measure (e.g.: patient had died or disavowed the reference program or provider) and are thus not considered non-responders.

  Of the remaining 3,356 non-responders (i.e., surveys sent to presumably eligible patients but not returned to us), the majority (80%) were not reachable: 63% were not reachable after the maximum 8 phone call attempts and 17% had non-working phone numbers). Of note, another 14% were reachable but refused to complete the survey.

  As prior survey research has established, it is likely that people who do not return or respond to surveys are systematically different than those who do. This is particularly likely among respondents who explicitly decline or refuse to answer the survey. Our data suggest that survey respondents were slightly older than nonrespondents (mean age 63.4 versus 60.9; $p < 0.01$). The proportion of women was also higher among respondents as compared with nonrespondents (56.2 percent versus 54.5 percent), but the difference was not statistically significant ($p = 0.21$). Although information on patient race was self-reported via the survey instrument, a subset of 12 participating palliative care programs provided patient race for at least 90 percent of their patients in their submitted data files. Among this subset, there was a greater proportion of White patients (88.1 percent versus 80.2 percent) and a lower proportion of Black patients (8.8 percent versus 11.9 percent) in the respondent group compared with the nonrespondent group. The results of a chi-squared test indicate that this difference is statistically significant ($p < 0.01$).

  Because the non-responders did not return a survey, we were unable to compare differences in measure scores between them and responders. Although outside the scope of this initial testing effort, future work could attempt to explore other differences between these two groups, for example, to qualitatively understand whether their care experiences differed, in order to shed light on potential response bias.

- **Issue 2: Telehealth**

  R6: I think Telehealth visits should be considered for inclusion in the future. R6, others: Concern about the exclusion of telehealth visits, should be included in the future

- **Developer Response 2:** We strongly agree that telehealth visits should be considered for inclusion in the future. Although we explored the inclusion of telephone and video visits as eligible visits at the outset of our alpha test, we decided not to include those visits because of their low frequency and difficulty identifying these visits. Thus, our initial performance measure eligibility criteria relied on coding in-person office visits. However, because of the COVID-19 pandemic, we were faced with an unexpected situation when participating palliative care programs shifted rapidly to providing telehealth services for their patients. With the input of our TECUPP and project advisory group, as well as input from participating programs, we decided to continue to disallow telehealth visits as eligible for the performance measure when we restarted data collection from September 2020 to February 2021. This ensured consistency in our results (i.e., we were measuring patient experiences with only in-person visits throughout the national beta field test) and avoided any potential confounding effects of the pandemic and telehealth use. However, it is likely that telehealth visits will continue in greater frequency than before the pandemic and should be included in measurement programs in the future. In interviews we conducted with palliative care programs during our testing phase, though most programs had little to no experience with telehealth prior to the pandemic, all programs converted to telehealth after March 2020 and continue to sustain telehealth services in some form. Closer attention to the development and testing of these and other patient experience measures within a telehealth context is warranted prior to widespread use in accountability programs.

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: Items sp.01-sp.30

   NOTE*: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

2. **Briefly summarize any concerns about the measure specifications.**

   For example: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

Reviewer 3: My main concern is with the potential misalignment of provider attribution and patient-reported outcome attribution. Provider was identified based on a three-month period, MIPS-eligible provider who the patient saw most often during the three-month period. However, the attached survey form refers to "the last 6 months". Given that provider who the patient saw most often in the 3-month period may not be the same one in the 6-month period, and it is quite likely that patient might have seen multiple providers during the 6-month period. Therefore, this may potentially cause provider misattribution. To further complicate things, the survey form does not identify the eligible ambulatory palliative care visit, so there is no explicit anchor visit for the patient to refer to even though the developer referred to the eligible ambulatory palliative visit repeatedly in this application, for example, the developer mentioned that patients who had transitioned to hospice could still answer the survey by reflecting on their experience with the visits. Additionally, the developer should clearly describe how the measure score is calculated. Specifically, how the results from the hierarchical logistic regression model are rolled up to a measure score. The model specified seems to be a 2-level model, what's unit of analysis? Survey item level response (which is binary) or patient level score (which is not binary)?

Reviewer 5: I may have missed something, but shouldn't the denominator be: "All patients aged 18 years and older who had an ambulatory palliative care visit [who completed the survey]"?

Reviewer 6: * The issue I have with the specifications (and it is easily fixable) is that the conversion of the rate to a 1-100 score is not transparent. It may be simple and assumed that most people know but I think this calculation should be spelled out. An example calculation would be helpful for the users as well. * Also, on page 35 it is indicated that data should be collected from "eligible palliative care patients that are representative of the palliative care provider program." This indicates to me that some sampling technique is used but up to this point in the application I thought the practice would send data on all of the patients who met the criteria - not sample. This is an easy fix and just needs a clarification.

Reviewer 9: Top Box scoring may be challenging, but this does push the Provider to deliver outstanding care to patients. Will there be any effort to communicate to the Provider the Top Box score on each of the four items so that the Provider can take a targeted intervention? The average score, while informative, does not provide the opportunity to make a targeted intervention.

Reviewer 11: I am reluctantly indicating "yes" on the specification because the developer needs to do a better job at the description. As I went through the form, things became clearer things like target population, age range, etc. should be included in the broad description.

RELIABILITY: TESTING

Submission document: Questions 2a.01-09

3. **Reliability testing level**

   For example: for some types of measures, if patient/encounter level validity is demonstrated, additional reliability testing is not required. Please review table above.

   ☒ **Accountable-Entity Level** ☒ **Patient/Encounter Level** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**

   NOTE: "level of analysis" reflects which entity is being assessed or held accountable by the measure.
   For example: If a measure is specified for a clinician level of analysis, but facility-level testing is provided, then testing does NOT match level of analysis. Or, if two levels of analysis are specified (e.g., clinician and facility) but testing is conducted for only one, then testing does NOT match level of analysis. Or, if claims data are selected as a data source, but testing data doesn't include claims data, then testing does NOT match data source.
   Also, check "NO" if only descriptive statistics are provided or submitter only describes process for data management/cleaning/computer programming.

   ☒ **Yes** ☐ **No**

5. If accountable-entity level and/or patient/encounter level reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical *VALIDITY* testing** of patient-level data conducted?

   According to current guidance patient/encounter level validity testing can be used for patient/encounter level reliability testing. Answer ONLY if you responded "Neither" on question #3 and/or "No" to question #4. Note that for some types of measures, additional reliability testing is not required IF patient/encounter level validity is demonstrated.

   ☒ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

   Submission document: Question 2a.10

   For example: Is the method(s) appropriate? If not, please explain (and offer potential alternatives if possible). Does the testing conform to NQF criteria and guidance? Was testing was conducted with the data source and level of analysis indicated for this measure? Address each level of testing provided, and each analysis under each method.

   Reviewer 1: Patient-level testing methods were strong. Entity-level testing was also generally strong. I was unclear is the hierarchical model accounted nesting within patient and facility. It looks like items (not patient scores) were the level of analysis, but without accounting for the nesting within patient? If the patient score was the level of analysis, then I don't understand the model form (logistic).
   Reviewer 3: The developer ascertained both internal consistency and test-retest reliability for data elements. Each survey item has 5 response categories, however, for the measure, top box scoring is used. Therefore, the developer needs to clarify if the testing was consistent with the top box scoring approach. Measure score reliability was assessed via hierarchical logistic regression model, although it is not clear how it was done, at survey item response level or patient score level?
   Reviewer 4: Data element reliability was tested using Cronbach's alpha = 0.95 and test-retest reliability (0.85) SNR = 0.752.
   Reviewer 5: Data element: Appropriate methods. Evaluated using both internal consistency and test-retest reliability coefficients. Score level: Appropriate method. Conducted "signal-to-noise" analysis.

Reviewer 6: Acceptable methods were used for the most part, but I question the use of a 2-day window for the test re-test. Two days seems to be too close to the original measure and may be influenced by memory of the first survey answers given.

Reviewer 8: Reliability of data elements was evaluated using both Cronbach's alpha as a measure of internal consistency and obtained data from a subset of respondents at two timepoints to conduct test-retest reliability. A final subsample of 197 respondents was used for the test-retest analysis. Only patient respondents who completed the original CATI survey without proxy assistance were invited to participate in the retest.

Reliability of the quality measure score at the program level was evaluated using "signal-to-noise" analysis. Hierarchical generalized-linear regression models with binomial outcomes to decompose the variability were used to relate outcome measures to programs and covariates, where the hierarchy of data is patient observations within the program.

The reliability from the measure test was then projected out based on observed variances and sample sizes from each program, using the Spearman-Brown prophecy formula to estimate a required within-program sample size to achieve a desired reliability for the measure.

Reviewer 9: The developer indicates that t-tests and chi-square tests were used, but the results are not presented clearly (e.g., in tabular format). Additionally, while the developer attempted to sample a diverse group based on race, this proved to be impossible given the data were collected during the Covid period and the strong homogeneity of patients (white, not Hispanic) in the palliative care sites that agreed to participate in the beta testing. The data collection instrument is new (developed by Rand?). The Cronbach's alpha tests methodology was presented. Similarly, the developer described the methodology (S-t-N) for the measure score was described as well as the computational method for applying the risk adjustment to the measure score.

Reviewer 10: Cronbach's alpha and signal to noise

Reviewer 11: Appropriate methods for testing. Should check to see if mode administration has any effect on the results.

Reviewer 12: The testing methods at the data element level and the score level is appropriate. Concerned that the developer only tested the telephone mode when other modes are available for the respondents. test-retest reliability calculation, we obtained data from a subset of respondents at two timepoints.

7. **Assess the results of reliability testing**

Submission document: Question 2a.11

For example: Is the test sample adequate to generalize for widespread implementation? Is there high or moderate confidence that the measure results and/or the data used in the measure are reliable? Address each level of testing provided, and each analysis under each method.

Reviewer 1: Patient-level testing revealed good internal reliability and test-retest reliability. Entity-level testing revealed good signal to noise reliability, but I'm a little unclear how the beta binomial distribution was used when the patient score is a proportion (not binomial).

Reviewer 3: Data element reliability testing needs to be consistent with the top box scoring approach. Further clarification in how the developer derived the reliability estimate from the hierarchical model would help with interpretation of the reported results. It is puzzling that accounting for design effect would lower the sample size needed from 45 to 37. The developer also reported the results based on an approach similar to Adams's with average reliability of 0.75.

Reviewer 4: Data element reliability was tested using Cronbach's alpha = 0.95 and test-retest reliability (0.85) SNR = 0.752.

Reviewer 5: Data element: Reliability of scores was high (Cronbach's alpha = 0.90; polychoric correlation coefficient = 0.85). Measure score: Average reliability across programs was approximately r = 0.752.

Reviewer 6: * Results for data element testing were 0.9 (alpha) and 0.85 (test-retest) and I believe this is acceptable. * Results for measure score were 0.052 (adjusted ICC with 95% CI 0.027 to 0.089). This is, according to the references I use for ICCs, very low and concerning.

Reviewer 8: Data Element Reliability: Results provide support for the reliability of scores derived from the four-data element Feeling Heard and Understood scale (Cronbach's alpha = 0.90). Test-retest reliability score was 0.85. Reliability was high indicating that scores can be reliably used in the quality measure. Quality Measure Score Reliability: The estimate of the Bayesian adjusted ICC was 0.052 (95% CI: 0.027 to 0.089) indicating a moderate level of between-program variability as compared to the within-program variability. However, using the S-B prophecy formula they estimate that to obtain a nominal reliability of 0.7, an average sample size of 37 would be required.

Reviewer 9: The data element reliability results based on a polychoric correlation coefficient) was 0.85; an acceptable value. The measure score reliability using a Bayesian generalized mixed effects models using the ICC distribution was 0.052 (95% CI: 0.027 – 0.089).

Reviewer 10: 0.90 and ICC of .052

Reviewer 11: Test sample is adequate (with the reference to the lack of testing of other modes of administration). It would be interesting to see if there was any difference in responses based upon the mode of administration.

Reviewer 12: the reliability of the quality measure score at the program level, we used a traditional "signal-to-noise" analysis that decomposes variability in the measure score into a) between-subject variability and b) within-subject variability

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? **NOTE:** If multiple methods used, at least one must be appropriate.

   Submission document: Question 2a.10-12

   For example: Appropriate signal-to-noise analysis; random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒ **Yes**

   ☐ **No**

   ☒ **Not applicable**

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Question 2a.10-12

   For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

   ☒ **Yes**

   ☐ **No**

   ☒ **Not applicable (patient/encounter level testing was not performed)**

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and **all** testing results):

    ☒ **High** (NOTE: Can be HIGH only if accountable-entity level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity level testing has not been conducted)

☒ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

☒ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of *OVERALL RATING OF RELIABILITY* and any concerns you may have with the approach to demonstrating reliability.**

Reviewer 1: High reported reliability but I had a few small questions about the methods.
Reviewer 3: For data element reliability testing, it needs to be done consistent with the top box scoring approach. For measure score reliability, the developer needs to clarify how the results were derived from the model.
Reviewer 4: I did not give it a "high" because of NSR = 0.75, which may be falsely elevated due to absence of meaningful risk adjustment
Reviewer 5: Strong methods with strong results.
Reviewer 6: * Was cognitive debriefing done with patients before the measure was tested? * I have a few issues with the survey items. Specifically, item #2 is double barreled, and research indicates that this leads to measurement error. Also, it is indicated through the application that surveys that were completely filled out by a proxy are excluded. However, it is unclear to me how this would be identified. I'm assuming (perhaps wrongly) that survey question 11 is used for this purpose and that option "answered the questions for me" is used to signify that the patient was not involved. However, I find this option unclear, and I would not have understood it to indicate that the patient was not involved. Thus, I think this item needed to be re-worked to increase clarity before use.
Reviewer 8: Reliability results were generally very strong. ASSUMING the minimum sample size of 37 is applied when implementing the measure; otherwise, would score moderate.
Reviewer 9: This is a “benefit of the doubt” rating. The measure score reliability was very low. However, given that this is a new measure and has a small, restricted sample due to Covid, the measure score does have potential to assess the patient’s experience with palliative care.
Reviewer 10: 0.90 and ICC of .052
Reviewer 11: Based on the results of testing.
Reviewer 12: Results provide support for the reliability of scores derived from the four-data element Feeling Heard and Understood scale (Cronbach’s alpha = 0.90). Test-retest reliability (i.e., polychoric correlation coefficient) for the Feeling Heard and Understood total score was 0.85. Reliability of the four-data element scale was high indicating that scores obtained are in fact reliable and can, therefore, be used in the construction of the quality measure. Reliability of the four-data element scale was high indicating that scores obtained are in fact reliable and can, therefore, be used in the construction of the quality measure. Results of the “signal-to-noise” analysis of quality measure reliability suggest there is a reasonable level of reliability based on the observed between-program variability and the within-program variability

VALIDITY: TESTING

12. **Validity testing level (check all that apply):**

☒ **Accountable-Entity Level** ☒ **Patient or Encounter-Level** ☐ **Both**

13. **Was the method described and appropriate for assessing the accuracy of ALL *critical data elements?***
NOTE that data element validation from the literature is acceptable.

Submission document*:* Questions 2b.01-02.

For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.
Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

☒ **Yes**

☐ **No**

☐ **Not applicable (patient/encounter level testing was not performed)**

14. **Method of establishing validity at the *accountable-entity level*:**

NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.
Submission document: Questions 2b.01-02
☐ **Face validity**
☒ **Empirical validity testing at the accountable-entity level**
☐ **N/A (accountable-entity level testing not conducted)**

15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Question 2b.02
For example: Correlation of the accountable-entity level on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

☒ **Yes**

☐ **No**

☐ **Not applicable (accountable-entity level testing was not performed)**

16. **Assess the method(s) for establishing validity**

Submission document: Question 2b.02

For example:

- *If face validity the only testing conducted: Was it accomplished through a systematic and transparent process, by identified experts, explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality, and the degree of consensus and any areas of disagreement provided/discussed?*
- *If a maintenance measure, but no empirical testing conducted, was justification provided?*
- *If construct validation conducted, was the hypothesized relationship (including strength and direction) described and does it seem reasonable?*

Reviewer 1: Validity test were appropriate in that the tested hypothesized relationships with stated standards for how the results should be interpreted.
Reviewer 3: To assess data element validity, the developer compared the key item to other relevant survey items to assess convergent validity. The developer also examined the association between this measure and other related measures. However, using "Receiving desired help for pain" for this testing is questionable as this is currently being evaluated at NQF now. The developer also surveyed seven experts for face validity.
Reviewer 4: data element validity – assessed convergent validity with survey data elements that are currently in use (r values 0.48-0.54) quality measure level – associated with other similar measures (r = 0.5 – 0.8)

Reviewer 5: Data element: Appropriate method. Included survey questions from other validated surveys, with the hypothesis that responses would move together. Measure score: Examined both convergent validity and face validity of the measure.

Reviewer 6: Convergent validity is an appropriate measure for this measure, and it seems that appropriate comparators were used.

Reviewer 8: Data element validity was assessed by including additional survey data elements from other instruments with expected relationship to the Feeling Heard and Understood items. These included CAHPS item “in the last 3 months, how often did this provider and team listen carefully to you?” and “in the last 6 months did you get as much help as you wanted for your pain from this provider and team?”.

Quality measure score validity was assessed by examining the relationship of the overall measure score with the “receiving desired help for pain” overall measure score, the CAHPS communication measure score and individuals’ overall rating of their palliative care provider and team. Face validity was determined using panel of experts. Advisors were asked to consider how well the measure scoring approach distinguishes between programs with high, medium, and low performance and how useful it is to quality improvement efforts. Advisors rated face validity on a scale of 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9).

Reviewer 9: The presentation of the methodologies for both data element and measure score was clear."

Reviewer 10: Construct validity, face validity

Reviewer 11: Appropriate testing.

Reviewer 12: higher scores on the Feeling Heard and Understood scale were associated with higher CAHPS communication scores (r = 0.54, p<.001) as well as with Receiving Desired Help for Pain (r = 0.48, p< .001). Taken together, these results support the convergent validity of the Feeling Heard and Understood data elements.

17. **Assess the results(s) for establishing validity**

Submission document: Questions 2b.03-04

For example: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient validity so that conclusions about quality can be made? Do you agree that the score from this measure as specified is an indicator of quality?

Reviewer 1: Convincing

Reviewer 3: Without using the results based on "receiving desired help for pain", the results were somewhat supportive. Face validity based on survey seems to be very good.

Reviewer 4: data element validity – assessed convergent validity with survey data elements that are currently in use (r values 0.48-0.54) quality measure level – associated with other similar measures (r = 0.5 – 0.8)

Reviewer 5: Data element: Tool items were positively correlated with hypothesized items from other surveys. Measure score: Overall scores were positively correlated with other measures of palliative care quality. Strong results on face validity.

Reviewer 6: * At the data element level the correlations were 0.54 with CAHPS communication and 0.48 with help for pain which, although in the positive direction, are a bit low. * at the measure score level - the correlations were 0.635 for CAHPS communication, 0.496 for help with pain, and 0.0768 for the overall rating of the provider. Two of the three of these are at an acceptable level. * Face validity was also assessed and seems acceptable.

Reviewer 8: Data Element level: Higher scores on the Feeling Heard and Understood scale were associated with higher CAHPS communication scores (r = 0.54, p<.001) as well as with Receiving Desired Help for Pain (r = 0.48, p< .001).
Quality Measure level: The measure scores were significantly and positively associated with the CAHPS communication quality measure (r = 0.635, p =0.011), the Receiving Desired Help for Pain quality measure (r = 0.496, p<.001) and the overall rating of the palliative care provider and team (r= 0.768, p=<.001). Seven expert advisors rated face validity of the measure score a mean of 8.3 on a scale of 1-9, corresponding with an average rating of "high." These ratings reflect strong support for face validity of the proposed quality measure from experts in palliative care and quality measurement.
Reviewer 9: Test both a "summed score" with a "threshold value" used for including the result, and simply aggregating (averaging?) the top box percentages and applying a binomial regression approach. The latter performed better for reliability based on ICC values computed using the Spearman-Brown Prophesy model. The Top Box/binomial approach was chosen for the measure score. Note: Proxy assistance is limited to reading the questions to the patient and/or translating the questions into the patient's primary language. The responses must be the patient's and questions to this effect are embedded in the instrument. The results for both data element and measure score were comparisons (correlation) to the existing CAHPS scores for these facilities. The results were positive, but not terribly impressive. The expert panel results for the measure score were high (8.3 out of 9).
Reviewer 10: .54, .48; 0.635 and .496 Face validity 8.3
Reviewer 11: Adequate results.
Reviewer 12: Yes

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**

Submission document: Questions 2b.15-18.

For example: Are there exclusions? If so, are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? Are any patients or patient groups inappropriately excluded from the measure? If patient preference (e.g., informed decision-making) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent? If you have concerns based on a clinical rationale, please note here as well as in question #29.

Reviewer 3: Potential misalignment between provider attribution and outcome attribution is a real threat.
Reviewer 5: None. Excluding "proxy only' completed surveys makes sense.
Reviewer 6: I am a bit concerned about the exclusion of tele-health visits and think this should be included in future revisions of this measure.
Reviewer 9: The exclusions were explained clearly and are reasonable.
Reviewer 11: No concerns.
Reviewer 12: No concerns, sample size was still adequate despite fairly low response rate. Also, as indicated the homogeneity of the response group threatens the validity across diverse populations

19. **Risk Adjustment**

Submission Document: Questions 2b.19-32

Applies to all outcome, cost, and resource use measures. Please answer all checkbox questions (19a - 19d), then elaborate on your answers in your response to 19e.

19a. **Risk-adjustment method**

☐ None ☒ Statistical model ☐ Stratification

☐ Other method assessing risk factors (please specify)

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☒ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☐ Yes ☒ No ☐ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☒ Yes ☒ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☒ No

19d.**Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☒ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☒ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☒ No

19d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

19e**. Assess the risk-adjustment approach**

For example: If measure is risk adjusted:

- *If the developer asserts there is no conceptual basis for adjusting this measure for social risk factors, do you agree with the rationale?*
- *How well do social risk factor variables that were available and analyzed align with the conceptual description provided?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)?*
- *If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision?*
- *Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)?*
- *Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

If measure is NOT risk-adjusted:

- *Is a justification for not risk adjusting provided (conceptual and/or empirical)?*
- *Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

Reviewer 3: The risk model seems overly simplified, there are many factors that should have been looked into and potentially included, for example, administrative home type, disease status and others.
Reviewer 4: Considered only a small number of patient level risk factors – education, race/ethnicity, proxy level, survey mode. severely constrained in our ability to explore potential risk adjustment by diagnosis because of the inadequacy of diagnosis data we received from programs in their submitted files. They performed univariate analyses to evaluate patient-level risk factors but did not seem to have attempted to test the significance of these risk factors in a multivariable regression model (likely due to data quality issues).

Reviewer 5: Appropriate risk adjusters: mode of administration - and - whether a proxy helped complete the survey.
Reviewer 6: No concerns.
Reviewer 9: The risk adjustment uses hierarchical generalized-linear model approach using a 12-month period for data collection collected during four three-month periods of Provider-Patient interaction.
Reviewer 11: Appropriate risk adjustment approach. Should provide additional testing regarding social determinants in the future as well as testing (and perhaps risk adjusting) for difference of mode administration.
Reviewer 12: Developer's analysis concluded no social risk factors were statistically significant

20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Questions 2b.05-07

    For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

    Reviewer 3: The range of measure scores seems to be quite wide, from 54 to 85 and the measure can identify programs at both ends in particular.
    Reviewer 5: No concerns. See variation in scores within a program and across programs.
    Reviewer 6: * Figure 5 on page 64 gives me pause. There is wide variation in many of the program scores and, from my reading of the figure, there are not many programs whose 95% CIs do not overlap.   * Also, it seems to me that the difference between programs, as evidenced by table 6 on page 65 are small and I'm concerned they are not meaningful.
    Reviewer 9: Because this is a new measure data, meaningful difference comparisons are difficult to assess empirically. The developer used preliminary data to extrapolate to create a larger hypothetical data set. The results of this hypothetical data set suggest that meaningful differences among palliative care facilities can be identified.
    Reviewer 10: None
    Reviewer 11: No concerns.

21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Questions 2b.11-14.

    ***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**. 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.*

    Note if not applicable. Note if applicable but not addressed. If multiple sets of specification (e.g., due to different data sources or methods of data collection): Do analyses indicate they produce comparable results?

    Reviewer 5: Not applicable.

Reviewer 6: N/A
Reviewer 10: None
Reviewer 11: Single specification.

22. **Please describe any concerns you have regarding missing data.**

    Submission document: Questions 2b.08-10.

    For example: Are there any sources of missing data not considered? Is it clear how missing data are handled? Is missing data more of a problem for some providers or patients than others? Does the extent of missing data impact the validity of the measure?

    Reviewer 1: I am concerned about survey non-response. Although not very large, there is variation in non-response between programs and demographic differences between responders and non-responders. I'm curious is the former is related to the latter. Are there better methods to account for survey non-response than just ignoring it?
    Reviewer 3: Nonresponse bias needs to be addressed with known differences between respondents and non-respondents.
    Reviewer 5: None. Overall, they had good survey response rates with levels comparable to those typical to CAHPS surveys; there were no clear outliers in terms of program nonresponse; and there were low levels of missingness in completed surveys.
    Reviewer 6: No concerns
    Reviewer 9: Measure score is designed to minimize missing data.
    Reviewer 10: None
    Reviewer 11: No concerns.

For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

23. **Are the specifications in alignment with the stated measure intent?**

    Consider these specific aspects of the measure specifications: attribution, cost categories, target population.

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

    Attribution: Does the accountable entity have reasonable control over the costs/resources measured? Is this approach aspirational (intending to drive change) or was it developed based on current state?
    Costing Approach: Do the cost categories selected align with the measure intent, target population and care settings? Is the approach for assigning dollars to resources
    Carve Outs: Has the developer addressed how carve outs in the data source are handled (or should be handled for other users)? For example, if pharmacy data is carved out (missing) from the data set, can a measure that focuses on cost of care for asthmatics still be valid?
    Truncation (approach to outliers): What is the threshold for outliers (i.e., extremely high cost or low-cost cases) and how are they handled?

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☒ **High** (NOTE: Can be HIGH only if accountable-entity level testing has been conducted)

☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity level testing has NOT been conducted)

☒ **Low** (NOTE: Should rate LOW if you believe that there **are** threats to validity and/or relevant threats to validity were **not assessed OR** if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the accountable-entity level and the patient/encounter level **is required**; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of *OVERALL RATING OF VALIDITY* and any concerns you may have with the developers' approach to demonstrating validity.**

Reviewer 1: Overall comforting validity testing, but I have lingering concerns over non-response bias.
Reviewer 3: Potential misalignment of provider attribution and patient-reported outcome attribution is a key concern. Inadequate risk adjustment and potential nonresponse bias issue are two additional concerns.
Reviewer 4: lack of risk adjustment for patient level factors. Although I understand that this is because of lack of patient-level data on risk factors, this is not an "excuse" for the lack of risk adjustment.
Reviewer 5: Used multiple, appropriate testing methods. All methods produced strong results.
Reviewer 6: I have this score based on the fact that I do not think this measure can identify meaningful differences between programs.
Reviewer 8: Validity tests consistently showed strong validity of the data elements and measure score.
Reviewer 9: As stated previously, this rating is based on the fact that this is a new measure with reasonable potential to be useful in assessing a patient's experience with palliative care.
Reviewer 10: Multi-method approach, moderate statistics
Reviewer 11: Based on testing and face validity results.
Reviewer 12: no concerns

For composite measures ONLY

If not composite, please skip this section.
Submission documents: Questions 2c.01-08
Examples of analyses:
1) If components are correlated - analyses based on shared variance (e.g., factor analysis, Cronbach's alpha, item-total correlation, mean inter-item correlation).
2) If components are not correlated - analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure).
3) Ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.
4) Overall frequency of missing data and distribution across providers. Ideally, sensitivity analysis of the effect of various rules for handling missing data and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

For example: Do the component measures fit the quality construct and add value? Are the objectives of parsimony and simplicity achieved while supporting the quality construct? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

- ☐ **High**
- ☐ **Moderate**
- ☐ **Low**
- ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

Reviewer 12: only concern was identified by developer and that was the homogeneity of the respondents

Developer Submission

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:

2021 Submission:
Updated evidence information here.

2018 Submission:
Evidence from the previous submission here.

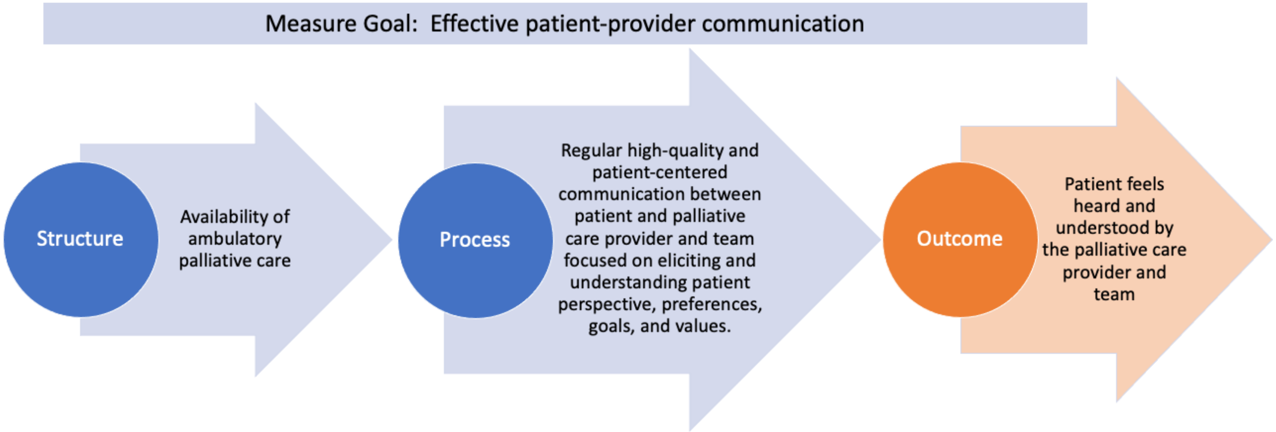
1a.01. Provide a logic model.

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]
The importance of the proposed measure for feeling heard and understood is predicated on existing guidelines and conceptual models of the quality of palliative care, including the National Consensus Project Clinical Practice Guidelines for Quality Palliative Care (2018) supported by a systematic review (Ahluwalia et al., 2018), the National Quality Forum Preferred Practices of Palliative and Hospice Care (National Quality Forum, 2006) (i.e. Preferred Practice 7 and 9 and 24), a consensus building process from the National Coalition for Hospice and Palliative Care, and input from qualitative inquiry of patients and providers.

There are various conceptual frameworks and theoretical models to help explain how patient-provider communication in the palliative care context can improve patient experience. The Integrative Framework of Appraisal and Adaptation in Serious Medical Illness (Bickel et al., 2020) posits that patients are constantly appraising and adapting to serious illness, and their health and emotional needs change throughout the course of their illness. The role of palliative care is to assess and respond to patient needs by providing expert medical knowledge tailored to the patient's specific informational needs, facilitating disease understanding and prognostic awareness, discussing options for management, and suggesting active coping strategies (Bickel et al., 2020). Good communication and interpersonal skills are thus a core competency for palliative care providers, and the patient's experience in this domain is central to the overall quality of palliative care (American Academy of Hospice and Palliative Care Medicine, 2009; National Quality Forum, 2006). By evaluating the extent to which patients feel heard and understood by their palliative care team, the proposed measure is also expected to capture the overall quality of palliative care.

The logic model that describes the relevant healthcare and structures to the patient's health care outcome is described below. The goal of the proposed measure is to facilitate and improve effective patient-provider communication that engenders trust, acknowledgement, and a whole-person orientation to the care that is provided. The outcome that is the focus of the proposed quality measure is that the patient feels heard and understood by the ambulatory palliative care provider and team. The proposed measure is related to three NQF Preferred Practices for Palliative and Hospice Care Quality (National Quality Forum, 2006): #7 – Ensure that upon transfer between healthcare settings, there is timely and thorough communication of the patient's goals, preferences, values, and clinical information so that continuity of care and seamless follow-up are assured.; #9 - Patients and caregivers should be asked by palliative and hospice care programs to assess physicians'/healthcare professionals' ability to discuss hospice as an option; and #24 - Incorporate cultural assessment as a component of comprehensive palliative and hospice care assessment, including but not limited to locus of decision making, preferences regarding disclosure of information, truth telling and decision making, dietary preferences, language, family communication, desire for support measures such as palliative therapies and complementary and alternative medicine, perspectives on death, suffering, and grieving, and funeral/burial rituals.


Measure Goal: Effective patient-provider communication
Structure
Availability of ambulatory palliative care
Process
Regular high-quality and patient-centered communication between patient and palliative care provider and team focused on eliciting and understanding patient perspective, preferences, goals, and values.
Outcome
Patient feels heard and understood by the palliative care provider and team

Citations:

Ahluwalia, S. C., Chen, C., Raaen, L., Motala, A., Walling, A. M., Chamberlin, M., O'Hanlon, C., Larkin, J., Lorenz, K., Akinniranye, O., & Hempel, S. (2018). A Systematic Review in Support of the National Consensus Project Clinical Practice Guidelines for Quality Palliative Care, Fourth Edition. *J Pain Symptom Manage*, *56*(6), 831-870.

American Academy of Hospice and Palliative Care Medicine. (2009). *Hospice and Palliative Care Medicine Core Competencies Version 2.3*.

Bickel, K. E., Levy, C., Macphee, E., Brenner, K. O., Temel, J. S., Arch, J. J., & Greer, J. A. (2020). An Integrative Framework of Appraisal and Adaptation in Serious Medical Illness. *J Pain Symptom Manage*, *60*(3), 657-677.e656.

National Consensus Project for Quality Palliative Care (2018). Clinical Practice Guidelines for Quality Palliative Care, 4th edition. Richmond, VA: National Coalition for Hospice and Palliative Care. https://www.nationalcoalitionhpc.org/ncp

National Quality Forum. (2006). *A National Framework and Preferred Practices for Palliative and Hospice Care Quality: A Consensus Report*. NQF.

[Response Ends]

1a.02. Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful.

Describe how and from whom input was obtained.

[Response Begins]
The meaningfulness of the measured outcome to the target population was assessed via 30- to 60-minute phone interviews with patients, caregivers, and family members. For these interviews, we sought patients who were currently receiving palliative care and/or hospice or who had received these services in the past, patients with advanced illness who were not currently receiving hospice and/or palliative care services, informal caregivers of patients receiving hospice and/or palliative care services, and patient advocates. The National Coalition for Hospice and Palliative Care sent outreach emails with information on this research to partners at the National Patient Advocate Foundation (NPAF), American Cancer Society, Family Caregiver Alliance, and National Alliance for Caregiving, and solicited nomination forms. NPAF also identified and provided the contact information of individual patients, caregivers, and family members, whom RAND directly contacted via phone. Our final sample of interview participants consisted of 4 patients with advanced illness who had never received hospice and/or palliative care (one of whom was joined by his wife during the interview), 8 patients who were receiving or had previously received palliative care, and 1 caregiver.

The interview protocols included the following topics: what it is like to receive palliative care, what information sharing between patients and providers would look like in an ideal situation, unmet symptom need and communication, and preferences for responding to mail versus in-person surveys. Participants mentioned the value of measuring heard and understood as a concept, with some commenting that it makes sense to measure this concept after a visit. One patient highlighted the importance of the heard and understood concept in the patient-provider relationship:

Heard and understood ... What does it mean to me? It means everything. If you're not heard, if you're not understood, you might as well walk away and find another doctor, because then ... you have nothing. That's the basis of ... the relationship. What's the point [otherwise]? Just so that he can tell you, “Oh, the test was okay.” [or] “The test wasn't okay.” Well, I need more than that from a doctor.
[Response Ends]

1a.03. Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

[Response Begins]
Seriously ill persons often report feeling silenced, ignored, and misunderstood in medical institutions (Frosch et al., 2012; Institute of Medicine Committee on Approaching Death, 2015; Norton et al., 2003). Systematically monitoring, reporting, and responding to how well patients feel heard and understood are crucial to creating and sustaining a health care environment that excels in caring for those who are seriously ill (Gramling et al., 2016). The quality of provider communication in serious illness is built on at least four mutually reinforcing processes: information gathering, information sharing, responding to emotion, and fostering relationships (Street et al., 2009). These elements directly shape patient experience and, when done well, help patients feel known, informed, in control, and satisfied, thus improving well-being and quality of life (Medendorp et al., 2017; Murray et al., 2015; Street et al., 2009).

Available evidence supports that the proposed patient-reported, patient experience measure can improve quality of care processes in palliative care settings. Improving provider communication can improve overall satisfaction with care among patients with serious illness (Anhang Price et al., 2018; Dy et al., 2008). The single-item *Feeling Heard and Understood* data element has been tested and used in palliative care populations (Gramling et al., 2016; Ingersoll et al., 2018). In multiple studies, hospitalized patients with serious illness who received palliative care consultations showed improvement in feeling heard and understood the day after the initial consultation (Gramling et al., 2016; Ingersoll et al., 2018). Gramling et al. (2016) found that patients’ baseline scores on the *Feeling Heard and Understood* data element varied, but 36% of patients improved their rating following palliative care consultation. Over 80% of respondents who reported feeling “not at all” heard and understood at baseline improved their rating at the post-consultation assessment, with 23% increasing their rating to “completely.” Similarly, Ingersoll et al. (2018) found that among the two-thirds of palliative care patients who felt incompletely heard and understood at baseline, more than half (56%) improved their rating post-consultation.

While no randomized controlled trials (RCTs) have assessed the newly developed *Feeling Heard and Understood* measure as an outcome, multiple RCTs demonstrate that interventions implemented in ambulatory palliative care settings can improve the quality of patient-provider communication (a different yet related concept to feeling heard and understood) (Dy et al., 2008). In one RCT, provision of a question prompt list to promote discussion about end-of-life issues during palliative care consultations with terminally ill patients and their caregivers led to fewer unmet needs for information without increasing patient anxiety or impairing satisfaction (Clayton et al., 2007). In another RCT, a computer-based training program to improve oncologist responses to patient expressions of negative emotion led to greater empathic responses to negative emotions by intervention oncologists as compared with controls, and patients of intervention oncologists reported greater trust in their oncologists than did patients of control oncologists (Tulsky et al., 2011). In addition, in a systematic review Chen et al. (2013) found strong evidence that routine collection of PROs in ambulatory oncology settings with timely feedback to providers can improve patient-provider communication about symptoms and quality of life issues and improve patient satisfaction (Berry et al., 2011; Detmar et al., 2002; Taenzer et al., 2000; Takeuchi et al., 2011; Velikova et al., 2004). Successful interventions occurred in the context of sufficient intensity of feedback (multiple times over a sustained period of time) targeting multiple stakeholders (doctors, nurses, interdisciplinary team members, and patients) (Chen et al., 2013). This evidence suggests that interventions implemented in ambulatory palliative care settings can lead to improved patient report of feeling heard and understood.

Palliative care itself is associated with improved outcomes including improved quality of life. A central goal of palliative care is to ensure that patients receive goal-concordant care (National Consensus Project for Quality Palliative Care, 2018). High quality serious illness communication is necessary to ensure goal-concordant care and feeling heard and understood is a key component of measuring this from the patient perspective (Sanders et al., 2018).

Citations:

Anhang Price, R., & Elliott, M. N. (2018). Measuring Patient-Centeredness of Care for Seriously Ill Individuals: Challenges and Opportunities for Accountability Initiatives. J Palliat Med, 21(Suppl 2), S-28-S-35.

Berry, D. L., Blumenstein, B. A., Halpenny, B., Wolpin, S., Fann, J. R., Austin-Seymour, M., Bush, N., Karras, B. T., Lober, W. B., & McCorkle, R. (2011). Enhancing patient-provider communication with the electronic self-report assessment for cancer: a randomized trial. *J Clin Oncol*, *29*(8), 1029-1035.

Chen, J., Ou, L., & Hollis, S. J. (2013). A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Serv Res*, *13*, 211.

Clayton, J. M., Butow, P. N., Tattersall, M. H., Devine, R. J., Simpson, J. M., Aggarwal, G., Clark, K. J., Currow, D. C., Elliott, L. M., Lacey, J., Lee, P. G., & Noel, M. A. (2007). Randomized controlled trial of a prompt list to help advanced cancer patients and their caregivers to ask questions about prognosis and end-of-life care. *J Clin Oncol*, *25*(6), 715-723.

Detmar, S. B., Muller, M. J., Schornagel, J. H., Wever, L. D., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial. *Jama*, *288*(23), 3027-3034.

Dy, S. M., Shugarman, L. R., Lorenz, K. A., Mularski, R. A., & Lynn, J. (2008). A systematic review of satisfaction with care at the end of life. J Am Geriatr Soc, 56(1), 124-129.

Frosch, D. L., May, S. G., Rendle, K. A., Tietbohl, C., & Elwyn, G. (2012). Authoritarian physicians and patients' fear of being labeled 'difficult'among key obstacles to shared decision making. *Health Affairs*, *31*(5), 1030-1038.

Gramling, R., Stanek, S., Ladwig, S., Gajary-Coots, E., Cimino, J., Anderson, W., Norton, S. A., Aslakson, R. A., Ast, K., Elk, R., Garner, K. K., Gramling, R., Grudzen, C., Kamal, A. H., Lamba, S., LeBlanc, T. W., Rhodes, R. L., Roeland, E., Schulman-Green, D., & Unroe, K. T. (2016). Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting. *J Pain Symptom Manage*, *51*(2), 150-154.

Ingersoll, L. T., Saeed, F., Ladwig, S., Norton, S. A., Anderson, W., Alexander, S. C., & Gramling, R. (2018). Feeling Heard & Understood in the Hospital Environment: Benchmarking Communication Quality Among Patients with Advanced Cancer Before and After Palliative Care Consultation. *J Pain Symptom Manage*, *56*(2), 239-244.

Institute of Medicine Committee on Approaching Death: Addressing Key End of Life Issues. (2015). *Dying in America: Improving Quality and Honoring Individual Preferences Near the End of Life*. National Academies Press.

Medendorp, N., Visser, L., Hillen, M., de Haes, J., & Smets, E. (2017). How oncologists' communication improves (analogue) patients' recall of information. A randomized video-vignettes study. *Patient education and counseling*, *100*(7), 1338-1344.

Murray, C. D., McDonald, C., & Atkin, H. (2015). The communication experiences of patients with palliative care needs: A systematic review and meta-synthesis of qualitative findings. *Palliative & supportive care*, *13*(2), 369-383.

National Consensus Project for Quality Palliative Care. Clinical Practice Guidelines for Quality Palliative Care, 4th edition. Richmond, VA: National Coalition for Hospice and Palliative Care; 2018. https://www.nationalcoalitionhpc.org/ncp

Norton, S. A., Tilden, V. P., Tolle, S. W., Nelson, C. A., & Eggman, S. T. (2003). Life support withdrawal: communication and conflict. *American Journal of Critical Care*, *12*(6), 548-555.

Sanders, J. J., Curtis, J. R., & Tulsky, J. A. (2018). Achieving Goal-Concordant Care: A Conceptual Model and Approach to Measuring Serious Illness Communication and Its Impact. *J Palliat Med*, *21*(S2), S17-s27.

Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, *74*(3), 295-301.

Taenzer, P., Bultz, B. D., Carlson, L. E., Speca, M., DeGagne, T., Olson, K., Doll, R., & Rosberger, Z. (2000). Impact of computerized quality of life screening on physician behaviour and patient satisfaction in lung cancer outpatients. *Psychooncology*, *9*(3), 203-213.

Takeuchi, E. E., Keding, A., Awad, N., Hofmann, U., Campbell, L. J., Selby, P. J., Brown, J. M., & Velikova, G. (2011). Impact of patient-reported outcomes in oncology: a longitudinal analysis of patient-physician communication. *J Clin Oncol*, *29*(21), 2910-2917.

Tulsky, J. A., Arnold, R. M., Alexander, S. C., Olsen, M. K., Jeffreys, A. S., Rodriguez, K. L., Skinner, C. S., Farrell, D., Abernethy, A. P., & Pollak, K. I. (2011). Enhancing communication between oncologists and patients with a computer-based training program: a randomized trial. *Annals of internal medicine*, *155*(9), 593-601.

Berry, D. L., Blumenstein, B. A., Halpenny, B., Wolpin, S., Fann, J. R., Austin-Seymour, M., Bush, N., Karras, B. T., Lober, W. B., & McCorkle, R. (2011). Enhancing patient-provider communication with the electronic self-report assessment for cancer: a randomized trial. *J Clin Oncol*, *29*(8), 1029-1035.

Chen, J., Ou, L., & Hollis, S. J. (2013). A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Serv Res*, *13*, 211.

Clayton, J. M., Butow, P. N., Tattersall, M. H., Devine, R. J., Simpson, J. M., Aggarwal, G., Clark, K. J., Currow, D. C., Elliott, L. M., Lacey, J., Lee, P. G., & Noel, M. A. (2007). Randomized controlled trial of a prompt list to help advanced cancer patients and their caregivers to ask questions about prognosis and end-of-life care. *J Clin Oncol*, *25*(6), 715-723.

Detmar, S. B., Muller, M. J., Schornagel, J. H., Wever, L. D., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial. *Jama*, *288*(23), 3027-3034.

Dy, S. M., Shugarman, L. R., Lorenz, K. A., Mularski, R. A., & Lynn, J. (2008). A systematic review of satisfaction with care at the end of life. J Am Geriatr Soc, 56(1), 124-129.

Frosch, D. L., May, S. G., Rendle, K. A., Tietbohl, C., & Elwyn, G. (2012). Authoritarian physicians and patients’ fear of being labeled ‘difficult’among key obstacles to shared decision making. *Health Affairs*, *31*(5), 1030-1038.

Gramling, R., Stanek, S., Ladwig, S., Gajary-Coots, E., Cimino, J., Anderson, W., Norton, S. A., Aslakson, R. A., Ast, K., Elk, R., Garner, K. K., Gramling, R., Grudzen, C., Kamal, A. H., Lamba, S., LeBlanc, T. W., Rhodes, R. L., Roeland, E., Schulman-Green, D., & Unroe, K. T. (2016). Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting. *J Pain Symptom Manage*, *51*(2), 150-154.

Ingersoll, L. T., Saeed, F., Ladwig, S., Norton, S. A., Anderson, W., Alexander, S. C., & Gramling, R. (2018). Feeling Heard & Understood in the Hospital Environment: Benchmarking Communication Quality Among Patients with Advanced Cancer Before and After Palliative Care Consultation. *J Pain Symptom Manage*, *56*(2), 239-244.

Institute of Medicine Committee on Approaching Death: Addressing Key End of Life Issues. (2015). *Dying in America: Improving Quality and Honoring Individual Preferences Near the End of Life*. National Academies Press.

Medendorp, N., Visser, L., Hillen, M., de Haes, J., & Smets, E. (2017). How oncologists’ communication improves (analogue) patients’ recall of information. A randomized video-vignettes study. *Patient education and counseling*, *100*(7), 1338-1344.

Murray, C. D., McDonald, C., & Atkin, H. (2015). The communication experiences of patients with palliative care needs: A systematic review and meta-synthesis of qualitative findings. *Palliative & supportive care*, *13*(2), 369-383.

National Consensus Project for Quality Palliative Care. Clinical Practice Guidelines for Quality Palliative Care, 4th edition. Richmond, VA: National Coalition for Hospice and Palliative Care; 2018. https://www.nationalcoalitionhpc.org/ncp

Norton, S. A., Tilden, V. P., Tolle, S. W., Nelson, C. A., & Eggman, S. T. (2003). Life support withdrawal: communication and conflict. *American Journal of Critical Care*, *12*(6), 548-555.
Sanders, J. J., Curtis, J. R., & Tulsky, J. A. (2018). Achieving Goal-Concordant Care: A Conceptual Model and Approach to Measuring Serious Illness Communication and Its Impact. *J Palliat Med*, *21*(S2), S17-s27.
Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, *74*(3), 295-301.
Taenzer, P., Bultz, B. D., Carlson, L. E., Speca, M., DeGagne, T., Olson, K., Doll, R., & Rosberger, Z. (2000). Impact of computerized quality of life screening on physician behaviour and patient satisfaction in lung cancer outpatients. *Psychooncology*, *9*(3), 203-213.
Takeuchi, E. E., Keding, A., Awad, N., Hofmann, U., Campbell, L. J., Selby, P. J., Brown, J. M., & Velikova, G. (2011). Impact of patient-reported outcomes in oncology: a longitudinal analysis of patient-physician communication. *J Clin Oncol*, *29*(21), 2910-2917.
Tulsky, J. A., Arnold, R. M., Alexander, S. C., Olsen, M. K., Jeffreys, A. S., Rodriguez, K. L., Skinner, C. S., Farrell, D., Abernethy, A. P., & Pollak, K. I. (2011). Enhancing communication between oncologists and patients with a computer-based training program: a randomized trial. *Annals of internal medicine*, *155*(9), 593-601.
Velikova, G., Booth, L., Smith, A. B., Brown, P. M., Lynch, P., Brown, J. M., & Selby, P. J. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. *J Clin Oncol*, *22*(4), 714-724.

[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]
Palliative care has expanded rapidly in recent years, and consensus has been growing within the palliative care community regarding the need for measuring the quality of end-of-life care. Yet the quality of care delivered by palliative care providers (and by other clinicians responsible for seriously ill patients) is unknown, particularly among patients who receive their palliative care early in their disease trajectory in ambulatory settings. As a result, stakeholders – including patients and their advocates, as well as providers and health systems – lack actionable measures to guide improvement efforts, as noted by NQF and the CMS Measures Application Partnership (MAP) as well as the 2017 CMS Environmental Scan and Gap Analysis Report (Centers for Medicare & Medicaid Services, 2017). Measures of palliative care quality are also underrepresented in the CMS QPP, with current measures addressing small populations that are often limited to patients with cancer or hospice patients. Furthermore, palliative care quality assessment that incorporates patient preferences (i.e., patient "voice") is noticeably absent despite the patient-centered nature of palliative care (Anhang Price & Elliott, 2018; Anhang Price et al., 2014; Anhang Price et al., 2018; Teno et al., 2017). Patient-centered measures, and especially patient-reported outcome measures, are an important complement to clinician-reported measurement data.

The patterns of palliative care received in ambulatory clinics differ substantially from palliative care received in other settings. Ambulatory palliative care typically supplements a primary treating service such as oncology, as needed. Patients may have several visits with different members of the palliative care team, or they may only have a single visit. This variability in the patient experience of palliative care raises important measurement challenges (Chen et al., 2020), which this project seeks to address by developing measures that are both broadly applicable to patients with serious illness, and useful to clinicians and health systems in measuring and improving the quality of care that patients with serious illness receive.

It is important to note that the palliative care field is unique in that palliative care patients are seriously ill, and death is not always a negative outcome, though the quality of that death is important. Accordingly, palliative care requires measures that examine whether patients are receiving care that aligns with their goals, rather than meeting clinical outcomes that may be more appropriate to other conditions, such as mortality (Chen et al., 2020).

Existing evidence and expert consensus have highlighted significant unmet need among seriously ill persons and gaps in meaningful communication measures, despite the noted importance of communication to seriously ill patients and their families (CMS Health Services Advisory Group, 2017). These gaps may be particularly pronounced in ambulatory settings, where patients and families have limited access to palliative care services and may struggle to manage their illness and accept their trajectory. Seriously ill persons often report feeling silenced, ignored, and misunderstood in medical institutions (Frosch et al., 2012; Institute of Medicine & Committee on Approaching Death Adressing Key End of Life Issues, 2015; Norton et al., 2003). Many patients with serious illness experience inadequate communication from their health care providers about prognosis and treatment options (Casarett et al., 2008; Covinsky et al., 2000; Dy et al., 2008; Teno et al., 2004; Teno et al., 2009) and receive care that is not consistent with their preferences (Khandelwal et al., 2017; Teno et al., 2004; Teno et al., 2015; Teno et al., 2009). Systematically monitoring, reporting, and responding to how well patients feel heard and understood is crucial to creating and sustaining a health care environment that excels in caring for those who are seriously ill (Gramling et al., 2016). Communication quality in serious illness comprises at least four mutually reinforcing processes: information gathering, information sharing, responding to emotion, and fostering relationships (Street et al., 2009). These elements directly shape patient experience and, when done well, help patients feel known, informed, in control, and satisfied, thus improving well-being and quality of life (Medendorp et al., 2017; Murray et al., 2015; Street et al., 2009).

During information gathering, to assess the meaningfulness of the measured outcome to the target population, AAHPM conducted one-on-one phone interviews with 13 patients, caregivers, and family members (PCFMs). For these interviews, we sought patients who were currently receiving palliative care and/or hospice or who had received these services in the past, patients with advanced illness who were not currently receiving hospice and/or palliative care services, informal caregivers of patients receiving hospice and/or palliative care services, and patient advocates. PCFMs described how the approach to communication, content, and tone were key components of good communication between providers or teams and palliative care patients. PCFMs felt that directly assessing whether patients feel heard and understood is “very, very important,” with some commenting that it makes sense to measure this concept after a visit (Chen et al., 2020).

The proposed measure is also valuable for implementation of innovative payment models for palliative care delivery that impact emerging models of community-based palliative care (e.g., embedded clinic models). Interdisciplinary palliative care team services are often unbillable under a fee-for-service model, and value-based payment models may be an alternative for reimbursement (Center to Advance Palliative Care, 2017). However, innovative financial models require quality metrics to ensure accountability for patients as well as payers and providers (Anhang Price et al., 2018; California Health Care Foundation, 2018). Many emerging models of community-based palliative care are delivered in community settings and may not utilize the same interdisciplinary team nor have the same level of training as programs evaluated in the literature (Teno et al., 2017). Palliative care quality measures would hold programs accountable for quality and would allow providers to demonstrate the value of their services (California Health Care Foundation, 2018). Currently available measures are generally limited to end-of-life utilization and process measures and are not consistently used across programs, thus patient reported quality metrics are needed to assess the impact of community-based palliative care and ensure transparency and accountability for these vulnerable patients (California Health Care Foundation, 2018; Teno et al., 2017).

Citations:

Anhang Price, R., & Elliott, M. N. (2018). Measuring Patient-Centeredness of Care for Seriously Ill Individuals: Challenges and Opportunities for Accountability Initiatives. *J Palliat Med*, *21*(Suppl 2), S-28-S-35.

Anhang Price, R., Elliott, M. N., Zaslavsky, A. M., Hays, R. D., Lehrman, W. G., Rybowski, L., Edgman-Levitan, S., & Cleary, P. D. (2014). Examining the role of patient experience surveys in measuring health care quality. *Med Care Res Rev*, *71*(5), 522-554.

Anhang Price, R., Stucky, B., Parast, L., Elliott, M. N., Haas, A., Bradley, M., & Teno, J. M. (2018). Development of Valid and Reliable Measures of Patient and Family Experiences of Hospice Care for Public Reporting. *J Palliat Med*, *21*(7), 924-932.

California Health Care Foundation. (2018). *Lessons Learned from Payer-Provider Partnerships for Community-Based Palliative Care*.

Casarett, D., Pickard, A., Bailey, F. A., Ritchie, C. S., Furman, C. D., Rosenfeld, K., Shreve, S., & Shea, J. (2008). A nationwide VA palliative care quality measure: the family assessment of treatment at the end of life. *J Palliat Med*, *11*(1), 68-75.

Center to Advance Palliative Care. (2017). *Payment Primer: What to Know about Payment for Palliative Care Delivery*.

Centers for Medicare & Medicaid Services, H. S. A. G. (2017). *CMS Quality Measure Development Plan Environmental Scan and Gap Analysis Report (MACRA, Section 102).* . https://www.cms.gov/Medicare/Quality-

Initiatives-Patient-Assessment-Instruments/ValueBased-Programs/MACRA-MIPS-and-APMs/MACRA-MIPS-and-APMs.html.

Chen, E. K., Ahluwalia, S. C., Shetty, K., Pillemer, F., Etchegaray, J. M., Walling, A., Kim, A., Martineau, M., Phillips, J., Farmer, C. M., & Ast, K. (2020). *Development of Palliative Care Quality Measures for Outpatients in a Clinic-Based Setting*. RAND Corporation.

Covinsky, K. E., Fuller, J. D., Yaffe, K., Johnston, C. B., Hamel, M. B., Lynn, J., Teno, J. M., & Phillips, R. S. (2000). Communication and decision-making in seriously ill patients: findings of the SUPPORT project. The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *J Am Geriatr Soc*, *48*(5 Suppl), S187-193.

Dy, S. M., Shugarman, L. R., Lorenz, K. A., Mularski, R. A., & Lynn, J. (2008). A systematic review of satisfaction with care at the end of life. *J Am Geriatr Soc*, *56*(1), 124-129.

Frosch, D. L., May, S. G., Rendle, K. A., Tietbohl, C., & Elwyn, G. (2012). Authoritarian physicians and patients' fear of being labeled 'difficult'among key obstacles to shared decision making. *Health Affairs*, *31*(5), 1030-1038.

Gramling, R., Stanek, S., Ladwig, S., Gajary-Coots, E., Cimino, J., Anderson, W., Norton, S. A., Aslakson, R. A., Ast, K., Elk, R., Garner, K. K., Gramling, R., Grudzen, C., Kamal, A. H., Lamba, S., LeBlanc, T. W., Rhodes, R. L., Roeland, E., Schulman-Green, D., & Unroe, K. T. (2016). Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting. *J Pain Symptom Manage*, *51*(2), 150-154.

Institute of Medicine, & Committee on Approaching Death Adressing Key End of Life Issues. (2015). *Dying in America: Improving Quality and Honoring Individual Preferences Near the End of Life*. National Academies Press.

Khandelwal, N., Curtis, J. R., Freedman, V. A., Kasper, J. D., Gozalo, P., Engelberg, R. A., & Teno, J. M. (2017). How Often Is End-of-Life Care in the United States Inconsistent with Patients' Goals of Care? *Journal of palliative medicine*, *20*(12), 1400-1404.

Medendorp, N., Visser, L., Hillen, M., de Haes, J., & Smets, E. (2017). How oncologists' communication improves (analogue) patients' recall of information. A randomized video-vignettes study. *Patient education and counseling*, *100*(7), 1338-1344.

Murray, C. D., McDonald, C., & Atkin, H. (2015). The communication experiences of patients with palliative care needs: A systematic review and meta-synthesis of qualitative findings. *Palliative & supportive care*, *13*(2), 369-383.

Norton, S. A., Tilden, V. P., Tolle, S. W., Nelson, C. A., & Eggman, S. T. (2003). Life support withdrawal: communication and conflict. *American Journal of Critical Care*, *12*(6), 548-555.

Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, *74*(3), 295-301.

Teno, J. M., Clarridge, B. R., Casey, V., Welch, L. C., Wetle, T., Shield, R., & Mor, V. (2004). Family perspectives on end-of-life care at the last place of care. *Jama*, *291*(1), 88-93.

Teno, J. M., Freedman, V. A., Kasper, J. D., Gozalo, P., & Mor, V. (2015). Is care for the dying improving in the United States? *Journal of palliative medicine*, *18*(8), 662-666.

Teno, J. M., Lima, J. C., & Lyons, K. D. (2009). Cancer patient assessment and reports of excellence: reliability and validity of advanced cancer patient perceptions of the quality of care. *J Clin Oncol*, *27*(10), 1621-1626.

Teno, J. M., Price, R. A., & Makaroun, L. K. (2017). Challenges Of Measuring Quality Of Community-Based Programs For Seriously Ill Individuals And Their Families. *Health Affairs*, *36*(7), 1227-1233.

[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]
A total of 44 ambulatory palliative care programs (i.e., the accountable clinical groups) participated in the beta field test (defined as providing at least one sample file during the field-testing period). A detailed description of palliative care program characteristics is provided in section 2a.05. We sought national representation in the beta field test by oversampling larger programs (i.e., those with more patients), stratifying recruitment efforts for ambulatory palliative

care programs by administrative home type (i.e., hospice, hospital, ambulatory, and other administration) and by geographic location to ensure representation across Census Regions. Among the ambulatory palliative care programs in our sample, administrative home types included ten hospice sites, 24 hospitals, and ten ambulatory or other administrative sites. Patient sampling was conducted each month between November 2019 and February 2021, with a pause between March and September 2020 due to the COVID-19 pandemic. We fielded a total of 7,595 surveys to eligible patients in the beta field test, of which 2,804 are completed surveys, or "cases," that were used for analysis. Completed surveys are defined as any survey returned within six months of lookback start date that was not excluded due to ineligibility (e.g., surveys sent to patients who were later identified as deceased, surveys completed entirely by a proxy respondent, or surveys to patients who disavowed the receipt of care). See section 2a.06 for a detailed description of the patient sample.

Based on the testing sample (n=44 palliative care programs), the average adjusted measure score is 72.1 (see sp.13 for explanation of measure scoring). The standard deviation in average program scores is 7.04. Analyses from the beta field test demonstrate room for improvement in the *Feeling Heard and Understood* measure:

- The observed variability across programs (adjusted ICC point estimate = 0.052) supports the potential of the measure to distinguish among programs with high, medium, and low performance.
- Across the 44 palliative care programs in our sample, adjusted program scores range from 54.05 to 85.18 with a standard deviation of 7.04. Confidence intervals for the highest and lowest program scores do not overlap: Lowest Program CI: (42.2, 66.0); Highest Program CI: (77.0, 91.1).
- When programs are ranked by their measure performance, we calculated that a program at the median of measure performance would need a large increase of 4.19 points in their measure score to improve to the 20th top-ranked program. A program at the bottom of the ranking (e.g., the 10th lowest ranked program) would need a 7-point increase measure score to improve to the median.

The mean, std dev, min, max, and interquartile range of the program adjusted scores are provided below:

Mean Score	S.D.	Min Score	Max Score	25th Percentile	75th Percentile	IQR
72.1%	7.0	54.1%	85.2%	67.8%	76.8%	9.1%

The deciles of the observed program adjusted scores (N=44) are:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
54.1	62.6	66.6	68.0	70.4	73.3	75.1	76.6	78.1	79.6	85.2

[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]
See 1b.02 above.
[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]
The measure is not currently in use.

To understand if and to what extent disparities in measure reporting and patient experience exist, we evaluated the relationship of various social risk factors to the measure score and the programs. These included patient race/ethnicity, education, and primary language, as well as multiple census-level variables such as race/ethnicity, urbanicity, median household income, gender, marital status, public insurance use, unemployment, and families below poverty line (see sections 2b.23 and 2b.24 for details). After adjustment for multiple comparisons, none of these variables were significant in their relationship with the measure.
[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]
The *Feeling Heard and Understood* measure has been used in palliative care research studies – but not in measurement programs - in different formats (Gramling et al., 2016; Ingersoll et al., 2018). The exact wording and structure for the proposed quality measure has not yet been used among ambulatory palliative care populations.

Among seriously ill patients, several patient-level characteristics have been associated with feeling incompletely heard and understood by health care providers. In a cross-sectional analysis of advanced cancer patients aged 21 and older at two U.S. medical centers, Ingersoll et al. (2018) evaluated the extent to which patients felt heard and understood by providers immediately before and the day after receiving an initial inpatient palliative care consultation. Patient-level factors associated with feeling incompletely heard and understood included lower levels of financial security, younger age, high emotional distress, patient uncertainty regarding one-year prognosis, and not endorsing a preference for comfort over longevity. Patient gender, race, ethnicity, and educational attainment were not significantly associated with feeling incompletely heard and understood prior to the palliative care consultation. Among those not feeling completely heard and understood prior to the consultation, more than half (55.6%) improved upon re-assessment the day after inpatient palliative care consultation. All patient-level factors associated with feeling incompletely heard and understood were substantially attenuated in the post-assessment, suggesting that palliative care consultation helps to improve health care communication globally (Ingersoll et al., 2018).

Other studies have examined disparities by race and ethnicity in satisfaction with provider communication among patients with serious illness (a different but related concept to feeling heard and understood), but findings have varied. In a national survey of bereaved family members, surrogates of Black patients reported lower satisfaction with the quality of end-of-life care and reported more concerns about provider communication (Welch et al., 2005). Another study found that terminally ill Black patients reported lower satisfaction than White patients with the overall quality of patient-physician relationships, with greater disparities observed in racially discordant patient-provider relationships (Smith et al., 2007). Other studies have reported conflicting findings. In a multicenter study to examine quality of communication and trust in patients with serious illness, patient characteristics associated with higher ratings of clinician communication included belonging to a racial/ethnic minority group, lower income, and higher religiosity (Coats et al., 2018). Another study found that patients from racial/ethnic minority groups, patients with lower income, and patients with lower educational attainment gave physicians in training higher ratings on end-of-life care communication; however, family members of non-white patients gave trainees lower ratings on communication (Long et al., 2014). A possible explanation for these conflicting findings is that various patient characteristics and contextual factors may impact patients' experience of provider communication. For example, socioeconomic status has also been associated with quality of patient-provider communication and may mediate associations by race and ethnicity (Coats et al., 2018; Siminoff et al., 2006).

Citations:

Coats, H., Downey, L., Sharma, R. K., Curtis, J. R., & Engelberg, R. A. (2018). Quality of Communication and Trust in Patients With Serious Illness: An Exploratory Study of the Relationships of Race/Ethnicity, Socioeconomic Status, and Religiosity. *J Pain Symptom Manage*, *56*(4), 530-540.e536.

Gramling, R., Stanek, S., Ladwig, S., Gajary-Coots, E., Cimino, J., Anderson, W., Norton, S. A., Aslakson, R. A., Ast, K., Elk, R., Garner, K. K., Gramling, R., Grudzen, C., Kamal, A. H., Lamba, S., LeBlanc, T. W., Rhodes, R. L., Roeland, E., Schulman-Green, D., & Unroe, K. T. (2016). Feeling Heard and Understood: A Patient-Reported Quality Measure for the Inpatient Palliative Care Setting. *J Pain Symptom Manage*, *51*(2), 150-154.

Ingersoll, L. T., Saeed, F., Ladwig, S., Norton, S. A., Anderson, W., Alexander, S. C., & Gramling, R. (2018). Feeling Heard & Understood in the Hospital Environment: Benchmarking Communication Quality Among Patients with Advanced Cancer Before and After Palliative Care Consultation. *J Pain Symptom Manage*, *56*(2), 239-244.

Long, A. C., Engelberg, R. A., Downey, L., Kross, E. K., Reinke, L. F., Cecere Feemster, L., Dotolo, D., Ford, D. W., Back, A. L., & Curtis, J. R. (2014). Race, income, and education: associations with patient and family ratings of end-of-life care and communication provided by physicians-in-training. *Journal of palliative medicine*, *17*(4), 435-447.

Siminoff, L. A., Graham, G. C., & Gordon, N. H. (2006). Cancer communication patterns and the influence of patient characteristics: disparities in information-giving and affective behaviors. *Patient Educ Couns*, *62*(3), 355-360.

Smith, A. K., Davis, R. B., & Krakauer, E. L. (2007). Differences in the quality of the patient-physician relationship among terminally ill African-American and white patients: impact on advance care planning and treatment preferences. *Journal of general internal medicine*, *22*(11), 1579-1582.

Welch, L. C., Teno, J. M., & Mor, V. (2005). End-of-life care in black and white: race matters for medical care of dying patients and their families. *J Am Geriatr Soc*, *53*(7), 1145-1153.

[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]
Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood
[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]
This is a multi-item measure consisting of 4 items: Q1: "I felt heard and understood by this provider and team", Q2: "I felt this provider and team put my best interests first when making recommendations about my care", Q3: "I felt this provider and team saw me as a person, not just someone with a medical problem", Q4: "I felt this provider and team understood what is important to me in my life."

Response to NQF request for clarification: Per the recommendation of our technical expert clinical user and patient panel (TECUPP), survey items refer to "this provider and team" which reflects the interdisciplinary team structure of care delivery in ambulatory palliative care. Providers can be one of many MIPS-eligible provider types, ranging from doctors of medicine to clinical nurse specialists. Providers serve as the lead of the palliative care team and are therefore referenced (i.e., named) at the start of the survey instrument. To identify the reference provider named on the survey instrument for each patient, the data set was first filtered to include only visits with MIPS-eligible provider types that occurred in the three months prior to the anticipated start date of survey fielding. We then selected the MIPS-eligible provider whom the patient saw most often within the three-month period, with ties in numbers of visits broken by provider type, giving preference to providers holding primary responsibility for patient care outcomes (e.g., physician or physician-designee over nurse or

therapist). If patients had multiple visits, we selected the most recent visit for each patient with the reference provider.
We did not conduct testing to specifically evaluate how patients differentiated between team members in their responses to the survey items.

[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]
Palliative Care and End-of-Life Care
[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)
[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]
Clinician: Group/Practice
[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]

Ambulatory Care

[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]

To be developed if appropriate following NQF endorsement.

[Response Ends]

sp.11. 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]

For the question below: state the outcome being measured. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.12. 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 *Feeling Heard and Understood* measure is calculated using top-box scoring. The top-box score refers to the percentage of patient respondents that give the most positive response. For all four questions in this measure, the top box numerator is the number of respondents who answer "completely true." An individual's score can be considered an average of the four top-box responses and these scores are adjusted for mode of survey administration and proxy assistance. Individual scores are combined to calculate an average score for an overall palliative care program.

[Response Ends]

For the question below: describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.13. 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]
Data for this measure is collected by survey. For the *Feeling Heard and Understood* 4-item scale, "top box scoring" takes each data element and defines "Completely true" (i.e., the top-box response) as passing for that item, where 1 = passing and 0 = not passing. The number of top-box responses (up to four) can be averaged across the number of responded items to provide an individual's estimate for their proportion of *Feeling Heard and Understood.* The within-individual estimates are averaged at the program level to provide the measure score. Missing data in the outcome is naturally accommodated among the four response items by the modeling procedure for adjusted score estimation; thus no outcome imputation is necessary and should not be performed. Risk-adjusted program level measure scores are estimated using hierarchical generalized-linear models that relate the proportion of top-box patient-level outcome responses to program scores, conditioned on risk adjustment covariates (survey mode and proxy assistance).
[Response Ends]

For the question below: state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]
All patients aged 18 years and older who had an ambulatory palliative care visit.
[Response Ends]

For the question below: describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.15. 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]
Denominator criteria:

All patients aged 18 years and older on date of encounter.

AND

Ambulatory palliative care visit[1] defined as:

- ICD-10 Z51.5 (Encounter for Palliative Care), ***OR***
- Provider Hospice and Palliative Care Specialty Code 17; ***AND***
- CPT 99201-99205 (New Office Visit); OR CPT 99211-99215 (Established Office Visit); or Place of service (POS) Code 11 – Office.

WITH

An eligible provider type: Physicians (including doctors of medicine, osteopathy, dental surgery, dental medicine, podiatric medicine, and optometry); osteopathic practitioners; chiropractors; physician assistants; nurse practitioners; clinical nurse specialists; certified registered nurse anesthetists; physical therapists; occupational therapists; clinical psychologists; qualified speech-language pathologists; qualified audiologists; registered dietitians or nutrition professionals.[2]

[1] Telehealth visits were not included in testing.

[2] Based on 2019 Merit-Based Incentive Program (MIPS) eligible clinician types

Response to NQF request for clarification:

- Yes, we intend for CPT 99211 to be included in the denominator. The list of CPT codes is meant to be as inclusive as possible to ensure that any new or established office visit is allowable in the denominator.
- We used this list of eligible clinicians in measure testing because we were developing the measure specifically for use in MIPS, and we thought it helpful to specify eligible provider types because palliative care is provided by an interdisciplinary team and a wide range of providers may see patients in the ambulatory setting. However, per feedback from CMS, we intend to remove the list of MIPS eligible providers from the denominator statement, replacing it with the statement "with a MIPS eligible provider."
- Telehealth visits are excluded from the denominator because our TECUPP emphasized variability and incompleteness in coding for telehealth visits among palliative care programs. This was verified at the outset of data collection by programs in our test sample. The COVID-19 pandemic and public health emergency provided new reimbursement policies for telehealth which resulted in improved coding practices however this improvement began in mid- to late-2020 when our national field test was nearing completion. Future work should explore inclusion of telehealth visits in the denominator; however we do not currently have testing data to support inclusion of these visits.
- We will consider adding this statement, given additional guidance on where in the denominator statement to include it.

[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]

Denominator exclusions include:

- Patients who do not complete at least one of the four items in the multi-item measure;
- Patients who do not complete the patient experience survey within six months of the eligible ambulatory palliative care visit;
- Patients who respond on the patient experience survey that they did not receive care by the listed ambulatory palliative care provider in the last six months (disavowal);
- Patients who were deceased when the survey reached them;
- Patients for whom a proxy completed the entire survey on their behalf for any reason (no patient involvement).

[Response Ends]

sp.17. 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]

Based on technical expert clinical user and patient panel (TECUPP) and advisor feedback, we propose that for programs to be eligible to participate in this measure that they demonstrate an ability to field the survey (i.e., deploy the survey per protocol by email, mail, and telephone) to ambulatory palliative care patients within three-months of eligible visits. Per discussion with the TECUPP, constraining the implementation to ensure that patients are sent surveys within 3-months of their eligible visit provides a sufficiently large pool of eligible patients with visits recent enough to avoid recall bias or loss to follow-up. Surveys must be completed by patients within 6 months of the visit to avoid challenges with recall or loss-to-follow-up which would make findings less actionable. During the alpha pilot test, we confirmed the feasibility of this implementation guidance.

Patients who have already completed the patient experience survey in a given reporting period should not be fielded the survey again to avoid response bias due to priming effects and minimize patient burden. Patients who do not complete the item set measuring *Feeling Heard and Understood* will be excluded from the denominator as no data will be available on the proposed measure. Providers and programs will not be penalized for non-response.

Patients who have died or are unable to complete the patient experience survey due to cognitive impairment will be excluded. Proxy assistance with the survey is allowed; however, following discussion with the project advisory board, we decided to exclude surveys that were completed solely by a proxy with no patient involvement for conceptual reasons. We elected to include proxy-assisted surveys and to add an adjustment for proxy assistance to account for small differences in measure components due to the proxy involvement.

Response to NQF request for clarification: As relevant background to our response, this measure was developed for use in ambulatory palliative care settings where patients can receive interventions to promote quality of life over the course of serious illness. Ambulatory palliative care is not the same as hospice, where many patients are not admitted until the last days of life. Ambulatory palliative care can be provided at any stage of serious illness, starting from diagnosis.

It should be noted that we looked for eligible outpatient visits within a 3-month lookback period from the date of the program's data pull. For example, a participating program could run a data query on August 1, 2020, covering all visits occurring for a patient in May, June, and July of 2020. The program would then send this file to RAND. Once we cleaned the file and identified the eligible visit in that 3-month timeframe we would field the survey to the patient. Given data processing times and the need to field surveys to all patients in participating programs at the same time (we did this on a quarterly basis through the fielding period), there was often a data lag between the receipt of each program's data and the survey fielding start date. In addition, there was often a 1–2-month data lag between when a program pulled their data and the timeframe they referenced (e.g.: a data pull on August 1, 2020, would most likely include visits occurring during the months of April, May, and June of 2020). Because of these data lags, although we identified visits within a 3-month period, to ensure that patients who received a survey were including that eligible visit in their consideration of their care experience, we used a 6-month reference timeframe in the wording of our survey questions (e.g.: "In the last 6 months, did you get as much help as you wanted for your pain from this provider and team?").

We worked closely with our technical expert, clinical user, and patient panel (TECUPP) to establish all these parameters prior to testing, and our alpha test provided additional support for the feasibility and face validity of this approach. Specifically, the TECUPP discussed and acknowledged that patients would likely (and ideally) have more than a single palliative care visit – potentially with different members of the palliative care interdisciplinary team - in the reference timeframes. They felt strongly that palliative care was a team-based discipline, and the eligible provider was accountable for the care provided by the team overall. They also acknowledged that patients would reflect on their care experience as a whole, which could include experiences with other providers seen during this timeframe, which is a challenge for patient experience measurement in general. We attempted to mitigate this by clearly specifying the palliative care provider seen by the patient in the eligible visit in the survey materials so as to orient the respondent to the care experience associated with that provider.

Specific to the reference timeframes, TECUPP members also discussed the challenges with a 6-month reference timeframe for patients to consider (e.g.; potential loss to follow-up if the patient became too ill to answer the survey or was moved to hospice by the time it was fielded, and patient recall) but acknowledged the error that data lags could introduce, and ultimately agreed that ensuring the eligible visit was captured in the timeframe referenced in the survey was of utmost importance. We selected the final time frame parameters based on discussion with palliative care experts from our technical expert clinical user and patient panel (TECUPP) and advisory board.

We confirmed the feasibility of these time frame parameters in testing. In the national field test, we found that the median number of days from the start of the eligible visit period to date of survey return was 124 days (about four months), with a minimum of 88 days (about three months) and a maximum of 167 days (about 5.5 months). Programs seeking to implement this performance measure should send the patient experience survey to patients within three months of their eligible visit to reasonably satisfy the six-month lookback time frame referenced in the performance measure. In testing we excluded patients who did not return the survey within the six-month time frame because of concerns regarding recall bias and because of their likely minimal impact (patient who returned a survey outside the six-month time frame n = 61 out of 3,356 nonrespondents, or 1.8 percent).

We are not aware of industry standards for other ambulatory palliative care surveys. In our information gathering activities we identified a gap in quality measures that have been designed for use in, and tested among, patients with serious illness receiving ambulatory palliative care services. The CAHPS Hospice survey evaluates palliative care experience from the perspective of bereaved caregivers, which is conceptually different from the proposed measure. Reference and lookback timeframes for that survey varies by mode of administration but data collection for sampled decedents/caregivers must be initiated two months following the month of patient death.

Response to NQF request for clarification, 8/30/21: We did consider whether to exclude hospice patients and it was indeed a very early exclusion. However, we later realized that since eligibility was based on an ambulatory palliative care visit, hospice patients would rarely be included. If they were included because they were receiving both types of care, that would be okay – we are still asking about the ambulatory palliative care provider and team, and we assume that patients are receiving other health care services; hospice should be no different. The pre-notification letter, the cover letter, and the wording at the start of the survey are intended to orient the patient to the specific provider and team.

We also considered that some patients may be in hospice by the time they receive the survey. If a patient entered hospice during the six-month period following the eligible visit but was able to reflect on their experiences with ambulatory palliative care (the referenced provider and team) and complete the survey, then they should have the opportunity to provide feedback on their experience of care. If the patient was too ill to complete the survey, had passed away, or was no longer living in the community we had processes in place to address these cases. Our data collection approach was to first send eligible patients a letter notifying them of the upcoming survey with a stamped postcard that could be returned in the event of death or a move/new address. If the patient had moved to a residential hospice, this could be indicated in the returned postcard noting they had moved. If they were still at home, but had discontinued their prior outpatient palliative care, they should still be eligible and able to respond about their experience with their ambulatory palliative care provider and team.

[Response Ends]

sp.18. 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.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.
[Response Begins]
Statistical risk model
[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.
[Response Begins]
Rate/proportion

[Response Ends]

sp.21. 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.22. 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]
Information for the measure calculation is collected via a survey data collection instrument, which will be provided to CMS, to be made available to CMS-approved survey vendors and palliative care programs. The below steps should be completed by an authorized survey vendor to minimize bias and reduce workload burden on programs. The survey vendor will be responsible for identifying eligible cases using electronic/automated queries, fielding the survey in the appropriate timeframes, receiving, cleaning, and summarizing survey data for program-level quality improvement (if requested by the program), and submitting a final program-level data set to CMS for measure scoring. This last step may include the submission of both program-level data as well as unadjusted program scores to CMS, for risk-adjustment once data are aggregated across programs.

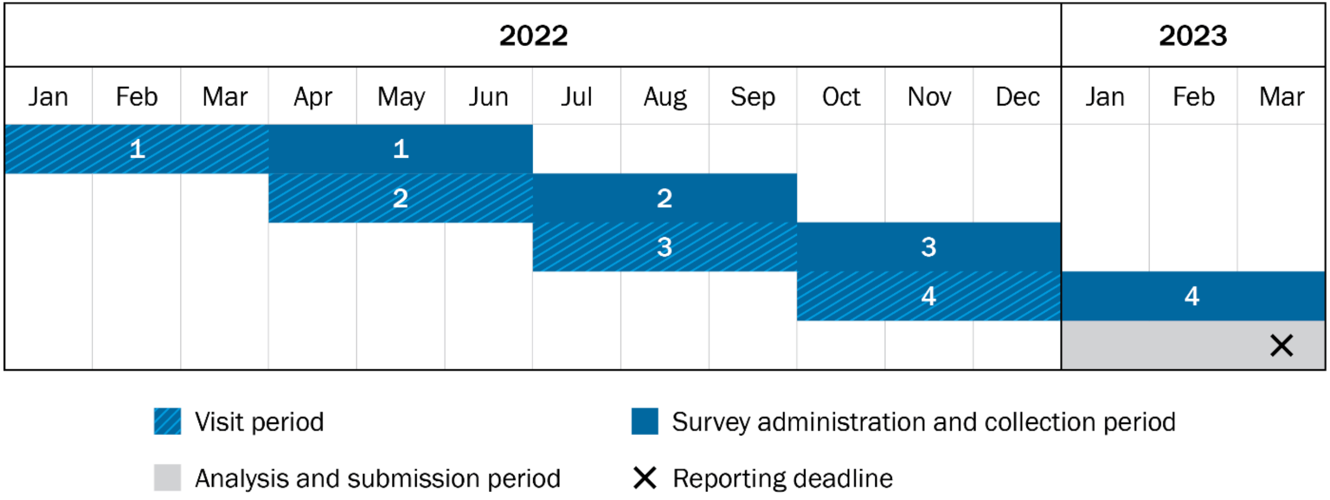
1. Identify eligibility within target respondent population
   a. Check patient age – 18 years or older?
      i. Yes - Eligible
      ii. No - Not eligible
   b. Check patient most recent disposition – alive?
      i. Yes - Eligible
      ii. No - Not eligible
   c. Check whether patient received in-person ambulatory palliative care visit with a MIPS-eligible provider within the past 3 months (see Figure 1 for example fielding and data collection timeframes). The reference provider named on the survey instrument for each patient is the MIPS-eligible provider who the patient saw most often within the three-month period, with ties in numbers of visits broken by provider type, giving preference to providers holding primary responsibility for patient care outcomes (e.g., physician or physician-designee over nurse or therapist).
      i. Yes - Eligible
      ii. No - Not eligible
   i. Check whether patient has already been fielded a survey in the current 12-month performance period
      i. Yes - Not eligible
      ii. No - Eligible
   ii. Check whether US-based contact information is available for patient
      i. Yes - Eligible
      ii. No - Not eligible
2. Field survey to all eligible cases using enhanced mixed-mode administration (web to mail to phone, i.e., live telephone interview)
3. Receive all returned survey data.
4. Identify any denominator exclusions.
   a. Survey completed (i.e., returned) within six months of the eligible ambulatory palliative care visit?
      i. Yes - Include
      ii. No - Exclude
   b. Patient participated in survey completion, with or without proxy assistance?
      i. Yes - Include
      ii. No - Exclude

        c. Patient responds in the patient experience survey that they received care by the listed ambulatory palliative care provider in the last six months?
            i. Yes - Include
            ii. No - Exclude

5. Score program-level quality measure using top-box scoring methodology (*See below for analysis of alternative measure scoring approaches)
    a. For the *Feeling Heard and Understood* 4-item scale, "top box scoring" takes each data element and defines "Completely true" (i.e., the top-box response) as passing for that item, where 1 = passing and 0 = not passing. The number of top-box responses (up to four) can be averaged to provide an individual's estimate for their proportion of *Feeling Heard and Understood*.
6. Eligible group, or survey vendor on behalf of the eligible group, submits clean program-level dataset (including unadjusted program score if applicable) to CMS for aggregation with other program datasets and measure scoring.
7. Risk adjustment calculation (to be performed by CMS or its third-party intermediary)
    a. To estimate risk-adjusted program level measure scores, we utilize hierarchical generalized-linear models that relate the proportion of top-box patient-level outcome responses to program scores, conditioned on risk adjustment covariates (survey mode and proxy assistance).
8. Scores are reported at the program level aggregating results over a 12-month, or calendar year reporting period. See O'Malley et al. (2005) for an example of risk adjusted scoring.

Figure 1 shows an example data collection and reporting schedule that reflects the process used during testing: identification of all eligible visits during a 3-month or quarterly time frame, and a subsequent 3-month survey administration/data collection time frame, with data from all participating programs aggregated over a 12-month, or calendar year reporting period.

Figure 1. Example Data Collection and Reporting Schedule for Measure Performance Year 2022



Citations:

O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., & Cleary, P. D. (2005). Case-mix adjustment of the CAHPS Hospital Survey. *Health services research*, *40*(6 Pt 2), 2162-2181. https://doi.org/10.1111/j.1475-6773.2005.00470.x

Analysis of Alternative Measure Scoring Approaches

We considered two distinct approaches to scoring the *Feeling Heard and Understood* quality measure based on the four "heard and understood" items. The first approach, which we refer to as the summed score approach, summed together the four items and then a threshold was selected that, if reached or exceeded, defined that an individual was "heard and understood." For example, each of the four items that compose our multi-item measure range from a value of 0 – "Not at All" to a top-box value of 4 – "Completely True" , thus the total score could range from 0 to 16 and a value could be selected that defines “passing” (e.g., the average score across programs). This approach was not optimal, however, as the observed responses were skewed heavily toward the top response on each item and the average summed score across individuals was high (mean=14.4, median=16). These high values meant that we were limited in our options for selecting a threshold. We could define a threshold at the mean or median, but from a program perspective, that value would seem difficult to achieve as it implies that all individuals would only “pass” the *Feeling Heard and Understood* measure if they responded in a top-box manner on all items. Additionally, preliminary ICC analyses suggested a requirement of very high minimum sample sizes under this approach (approximately 100 returned surveys to achieve an ICC of 0.7 using the Spearman-Brown Prophecy formula).

In light of these challenges, we considered another approach in which we treat each of the four data elements of the *Feeling Heard and Understood* scale as reflecting the construct of feeling heard and understood. Under this approach, we use a binomial regression approach that aggregates the four individual responses and takes full advantage of all of the information in the sample. This binomial regression approach performed much better than the summed score approach, defining a less stringent threshold for passing and requiring a much smaller minimum sample size (approximately 40 completed surveys per provider) to achieve sufficient reliability. This is due to the fact that the binomial model allows us to more efficiently include information on each item from each individual by considering all observations together in scoring, as opposed to collapsing them into a single observation per person as would be done with a summed score approach.

[Response Ends]

sp.23. 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: Patient Experience Survey - Feeling Heard and Understood.pdf

sp.24. Indicate the responder for your instrument.

[Response Begins]
Patient
[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]
The target population for sampling includes patients aged 18 years or older who received ambulatory palliative care services from a MIPS-eligible provider within the three months prior to the start of survey fielding. Findings from the alpha pilot test and beta field test support the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. The provider or program will provide a vendor with an extract file of all patients who received care during the measurement period. To prevent gaming and to minimize administration and social desirability bias, the vendor will apply the eligibility criteria to identify the patient sample and field the survey to eligible patients. Survey administration will be mixed-mode, including web (emailed link to online survey), mail (hard-copy of the survey) followed by telephone (Computer Assisted Telephone Interviewing) survey if needed.

Assessments of measure reliability based on the intraclass correlation coefficient (ICC) suggest that programs will need a sufficient sample to have at least 37 completed responses to the *Feeling Heard and Understood* items over the 12-month reporting period.

Response to NQF request for clarification: We will add a minimum sample size requirement of 37 respondents per palliative care group to our measure specifications. However, this estimate is based on our current testing sample and should be revisited in future years.

[Response Ends]

sp.26. Identify whether and how proxy responses are allowed.

[Response Begins]
Proxy assistance is allowed. However, the patient must be involved in survey completion. Patients for whom a proxy completed the entire survey on their behalf for any reason (i.e., with no patient involvement, including patients who are deceased by the time the survey reaches them) are excluded.
[Response Ends]

sp.27. 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]
The measure is composed of survey data representing patient report of care over a reporting period of one calendar year (January 1st to December 31st). A copy of the survey instrument can be found in the Appendix. Programs will need to contract with a survey vendor to field surveys and process data. The data should be collected from eligible palliative care patients that are representative of the palliative care provider or program within the designated timeframe. Additional details of sampling are described in sp.22. An enhanced mixed-mode survey administration design, web to mail with telephone follow-up, is recommended.

As this measure is a multi-item measure, response rates at the program level should be calculated with respect to key items and reported to determine the sufficiency of the data to calculate the measure. We found that measure reliability is sensitive to smaller (i.e., lower patient volume) programs. For a reliable *Feeling Heard and Understood* measure we recommend an average sample size, at the program-level, of 37 participants responding to the *Feeling Heard and Understood* data elements. Based on an estimated response rate range of 37% to 46% as was found in the measure testing process, ambulatory palliative care programs with annual visit volume of between approximately 80 and 100 adult patients could achieve the minimum average sample size required for a reliable measure.

Missing data in the outcome is naturally accommodated among the four response items by the modeling procedure for adjusted score estimation; thus no outcome imputation is necessary and should not be performed. Missing values for proxy assistance should be imputed as “No Proxy Assist.”
[Response Ends]

sp.28. Select only the data sources for which the measure is specified.

[Response Begins]
Electronic Health Records
Instrument-Based Data
[Response Ends]

sp.29. 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]

Patient-reported data is collected via survey instrument. The instrument was developed for this measure and can be completed via web survey, on paper or over telephone in English. Patient eligibility is determined based on coded visit information in the electronic health record.
[Response Ends]

sp.30. 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

OR
- 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 measures 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 Importance to Scientific Acceptability sections. For example:

2021 Submission:
Updated testing information here.

2018 Submission:
Testing from the previous submission here.

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

[Response Begins]

Electronic Health Records
Instrument-Based 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]
Does not use existing data.
[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]
11-01-2019 – 02-28-21

Analyses include data collected between November 2019 and February 2021, with a pause between March and September 2020 due to the COVID-19 pandemic.

> **Response to NQF request for clarification:** We faced an important and unusual threat to the validity of the *Feeling Heard and Understood* data element and performance measure when the COVID-19 pandemic began in March 2020, during our national beta field test period. We realized that the pandemic could alter the provision and experience of receiving outpatient palliative care and thus disrupt the relationship between quality of care and patient responses. To evaluate the potential impact of the pandemic at the data element level, we compared scores for the *Feeling Heard and Understood* survey data elements across time. Specifically, we conducted statistical tests to see whether the individual data elements, total sum score, or the distribution of individual proportion of top boxes differed before the pandemic versus after the onset of the pandemic. Mean scores were compared using a generalized linear model, and chi-squared tests were performed to assess whether there were significant differences in the distributions of responses for the *Feeling Heard and Understood* data elements comparing the pre-COVID-19 pandemic (data collection rounds one through four) and after the height of the pandemic (rounds seven through ten).
>
> To evaluate the potential impact of the pandemic on the validity of the performance measure, we analyzed *Feeling Heard and Understood* performance measure scores across time to see whether measure scores were substantively affected by the pandemic. Specifically, we conducted statistical tests to see whether the performance measure scores differed before the pandemic versus after the onset of the pandemic. Mean performance measure scores were analyzed using paired t-tests (i.e., the same programs were included in the pre-pandemic and mid-pandemic groups) to assess whether there were significant differences pre-COVID-19 pandemic (data collection rounds one through four) and after the height of the pandemic (rounds seven through ten).
>
> The results indicate that there was a significant (α = 0.05 level) difference in the overall distribution of responses for one data element (Q1: "I felt heard and understood by this provider and team"). However, when the statistical tests were adjusted for multiple comparisons (using the Benjamini-Hochberg method to control the false discovery rate), the effect was no longer statistically significant. When we compared the *Feeling Heard and Understood* individual score (i.e., considering the aggregate proportion of top-box responses) pre-COVID-19 and during the pandemic, the performance measure score distributions were similar across the two periods. Although the proportion of respondents with all top-box responses fell by 5 percentage points in the mid-pandemic rounds, the difference in distributions was not statistically significant (p = 0.06).
>
> Because the chi-squared test is sensitive to small differences in the distributions, we also examined the average individual measure score using a generalized linear model (binomial regression adjusting for an indicator of mid-pandemic). We found that the mean of the *Feeling Heard and Understood* individual measure score does not differ significantly between groups (p = 0.21).

Finally, we found no significant difference in performance measure scores pre- or mid-pandemic (t(24) = 0.268, p = 0.791). This suggests that the proposed performance measure was largely robust to the dramatic changes caused by the COVID-19 pandemic, which is promising for the future validity of the performance measure.

[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]
Clinician: Group/Practice
[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]
We aimed to recruit a total of 45 ambulatory palliative care programs (i.e., the accountable clinical groups) for the beta field test, based on assumptions informed by our alpha test regarding providers per program and total patient volume. We sought national representation by oversampling larger programs (i.e., those with more patients), stratifying recruitment efforts by administrative home type (i.e., hospice, hospital, ambulatory, and other administration) and by geographic location to ensure representation across Census Regions (Table 1). Using the list of 395 ambulatory palliative care programs, we sorted programs into recruitment queues according to these criteria. RAND's Survey Research Group (SRG) began contacting programs to discuss participation in the beta field test. Recruitment focused on ensuring the program provided ambulatory palliative care, had sufficient established patient volume in the ambulatory setting to ensure a timely contribution to the testing sample, and were MIPS-eligible at both the program and provider levels. Program recruitment contact began with an introductory email followed by a telephone call from SRG staff members. Contact continued until quotas for each queue (i.e., setting type and geographic region) were reached and Data Use Agreements (DUAs) were executed. A total of 238 palliative care programs were contacted about the test, at which point we met desired target numbers and discontinued outreach/recruitment efforts. Of these 238 contacted programs, 70 programs were deemed ineligible to participate in the field tests for one or more of the following reasons:

- did not provide ambulatory care
- were less than six-months old and thus had little established experience providing ambulatory palliative care
- saw fewer than 20 patients in an ambulatory setting over the prior six months and thus would be unlikely to contribute adequate sample size for testing purposes
- were a PACE (Program of All-Inclusive Care for the Elderly) or VA (Veterans Administration) program and thus not eligible for the MIPS program; or
- had no MIPS-eligible practitioners (as of the 2019 MIPS-eligible provider list) who provided ambulatory palliative care to patients.

Of the remaining eligible programs, 44 programs participated (defined as providing at least one sample file during the field-testing period) (Table 1).

Table 1. Beta Field Test Recruitment Targets and Final Recruitment Data

*	*	Midwest	Northeast	South	West	TOTAL
Hospice	Targeted sites to recruit	3	1	3	1	8
	Sites recruited	2	4	3	1	10
Hospital	Targeted sites to recruit	5	9	7	7	28
	Sites recruited	5	6	6	7	24
Ambulatory/Other	Targeted sites to recruit	3	2	5	5	15
	Sites recruited	2	3	3	2	10
All Settings	Targeted sites to recruit	11	11	15	13	50
All Settings	Sites recruited	9	13	12	10	44

Cell left intentionally blank.

[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]
Our target patient sample was guided by several assumptions based in part on our alpha pilot test findings, including: 1) there would be an average of three MIPS-eligible providers per program, 2) providers would see on average ten unique eligible patients per three-month period, and 3) a 40 percent survey response rate would be achieved, based on existing literature (Parast, 2018). Operating under these assumptions, we planned to field between 6,000 and 7,500 surveys to patients receiving ambulatory palliative care at the participating programs and expected between 2,400 and 3,000 completed surveys. These assumptions/criteria were used solely to establish basic parameters for the fielding design. Minimum requirements for participation in the measure were determined based on reliability analyses conducted after the field period (see sections 2a.10-2a.11).

To maximize the chance that we would achieve adequate sample size, we used data from all eligible providers belonging to a program and all eligible patients cared for by these providers. If a program submitted too many patients for SRG to contact within a given period due to capacity constraints, a random sample of eligible patients up to the number SRG could field in that period was selected.

We fielded the survey to 7,595 sampled patients across 10 rounds. Of these, 3,356 were not returned, 1,435 were excluded from any analyses due to ineligibility for the larger study, and 2,804 were returned and included in analyses (37% raw response rate; 46% response rate excluding ineligible patients) (Table 2). Completed surveys are defined as any survey returned within six months of the lookback start date that was not excluded due to ineligibility (e.g., surveys sent to patients who were later identified as deceased, surveys completed entirely by a proxy respondent, or surveys to patients who disavowed the receipt of care).

Table 2. Final Disposition of Fielded Surveys (n=7,595)

Category	Description	Counts	Percentage	Count	Percentage
Completed Surveys	Mail	1298	17.09%	2804	36.92%
	Phone	980	12.90%		
	Web	526	6.93%		
Survey Nonresponse		3356	44.19%	3356	44.19%
Excluded Surveys Due to Ineligibility		1435	18.89%	1435	18.89%
				Overall Total	7595

Cell left intentionally blank.

Table 3 provides descriptive information on survey nonresponse and surveys that were otherwise ineligible for measure analysis. Survey nonresponse includes patient refusals to complete a phone interview, bad or disconnected phone numbers, or inability to reach patient after maximum attempts.

Table 3. Survey Nonresponse (n=3,356)

Category	Counts
Maximum Calls (8) Reached Without Response	2128
Not Found (i.e., Bad Phone Number)	583
Refusals	463
Final Refusal (Patient Reached but Refused Participation)	348
Informant Refusal (Someone Other Than Patient or Proxy Declined)	87
Breakoff (Respondent Discontinued During CATI)	28
Patient Unable and No Available Proxy	121
Late Complete	61
Mail	58
Web	3
Overall Total	3356

Surveys with other ineligibility factors include patients who were indicated as deceased by the time the survey reached them (a proxy could return a stamped postcard to indicate this, or may have notified us when reached by telephone), returned surveys completely solely by a proxy without patient involvement, surveys where the respondent disavowed the program (i.e., indicated that they had not received care in the past 6 months from the stated palliative care provider and team), or for whom we had bad contact information, such as when a patient had moved or the mailed packet was returned to sender (Table 4).

Table 4. Excluded Surveys Due to Ineligibility (n=1,435)

Ineligibility Category	N (% of total excludes)
Patient deceased	748 (52%)
Proxy-only response	435 (30%)
Disavowed program/provider	146 (10%
Bad contact information; moved	35 (2%)
Multiple ineligibilities (e.g., patient deceased and proxy-only response)	71 (5%)
Overall Total	1435

Survey respondents, i.e., those who comprised our analytic sample (n=2,804), generally reflected the larger patient sample (n=7,595) (Table 5). However, there were slight differences between the respondents and the larger patient sample in terms of age and race. Survey respondents were slightly older than nonrespondents (mean age 63.4 vs 60.9; p<0.01). The proportion of women was also higher among respondents compared to nonrespondents (56.2% vs 54.5%), but the difference was not statistically significant (p = 0.21). Among the subset of 12 programs who provided patient race for at least 90% of their patients, respondents were more likely to identify as White (88.1% vs 80.2%) and less likely to identify as Black (8.8% vs 11.9%) or another race (3.1% vs 8%) compared to nonrespondents. The results of a chi-squared test indicate that this difference is statistically significant (p < 0.01).

Table 5. Patient Respondent Characteristics

Characteristic	Summary	S.D.	Percent Missing
Number of Observations	N=2804		
Age (Mean)	63.36	13.32	0.04%
Gender (Male %)	43.81%		0.04%
Race	N=2753		1.82%
White	87.61%		
Black or African American	5.88%		
Asian	0.91%		
Multi-racial	2.76%		
American Indian or Alaska Native	0.44%		
Native Hawaiian or other Pac. Islander	0.25%		
Other	2.14%		
Education	N=2782		0.78%
More than 4-year college degree	15.74%		
4-year college graduate	15.46%		
Some college or 2-year degree	34.94%		
High school graduate or GED	25.63%		
Some high school but did not graduate	6.40%		
8th grade or less	1.83%		
Hispanic	N=2743		2.18%
Yes, Hispanic or Latino	4.67%		
No, not Hispanic or Latino	95.33%		

Cell left intentionally blank.

Citations:

Parast, L., Elliott, M. N., Hambarsoomian, K., Teno, J., & Anhang Price, R. (2018). Effects of Survey Mode on Consumer Assessment of Healthcare Providers and Systems (CAHPS) Hospice Survey Scores. *J Am Geriatr Soc*, *66*(3), 546-552.

Response to NQF request for clarification: We purposively sampled programs of all sizes from all regions of the U.S. Despite our geographically diverse sample of palliative care programs, the survey data collected during the national beta field test largely reflect the experiences of a non-Hispanic White patient population. The homogeneity of the sample is similar to that in other studies in palliative care populations and may be a function of multiple factors (Temel, Greer, Admane, et al., 2011; Temel, Greer, El-Jawahri, et al., 2017). Such factors include the limited reach of palliative care services into diverse communities and the inaccessibility of palliative care to certain patient groups or survey response, as other studies have also shown that non-White participants may be less likely to respond to mailed surveys (Elliott, Edwards, et al., 2005; Link et al., 2006; Couper et al., 2007). As palliative care groups in our sample did not consistently capture or provide data on patient race in their sampling files, we were unable to evaluate potential response bias (i.e., whether non-White patients were less likely to respond to the survey).

However, to better understand these factors and to capture the experiences of racial and ethnic minority patients with ambulatory palliative care and with receiving desired help for pain, we queried participating programs about the population they serve and interviewed racial and ethnic minority patients and family caregivers who had experience with palliative care. Most programs that participated in interviews reported that the majority of patients in their ambulatory palliative care practice were White (with estimates ranging from 75 percent to 95 percent White). Only one program described its patient population as "pretty diverse." For many programs, their patient population represented the demographics of the larger communities and areas served in the programs' geographical reach. For a few programs, however, interviewees noted that their outpatient palliative care patient population was not representative of their larger communities. Program interviewees observed that access to ambulatory palliative care, a lack of diversity in palliative care providers, cultural mistrust in the medical system, and an overall misperception of palliative care as end-of-life care may be barriers to engaging patients from racial and ethnic minorities in ambulatory palliative care. Some interviewees described institutional outreach efforts to engage patients of diverse racial and ethnic backgrounds, including by establishing a diversity committee or education for providers about perceptions of palliative care among patients from racial or ethnic minorities. Some of this was echoed in the interviews with non-White patient and family caregivers who had experience with palliative care. Interviewees described highly positive experiences with palliative care and noted feeling heard and understood by their palliative care providers but also noted certain challenges they faced accessing palliative care. Though our data cannot discern whether the homogeneity of the sample reflects a response bias or a care disparity or some combination of both, these interview data as well as input from our TECUPP suggest that non-White patients likely face various systems- and individual-level barriers to accessing palliative care that reflect an important care disparity.

References:

Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter, "Noncoverage and Nonresponse in an Internet Survey," Social Science Research, Vol. 36, No. 1, March 2007, pp. 131–148.

Elliott, Marc N., Carol Edwards, January Angeles, Katrin Hambarsoomian, and Ron D. Hays, "Patterns of Unit and Item Nonresponse in the CAHPS Hospital Survey," Health Services Research, Vol. 40, No. 6, Part 2, December 2005, pp. 2096–2119.

Temel, Jennifer S., Joseph A. Greer, Sonal Admane, Emily R. Gallagher, Vicki A. Jackson, Thomas J. Lynch, Inga T. Lennes, Connie M. Dahlin, and William F. Pirl, "Longitudinal Perceptions of Prognosis and Goals of Therapy in Patients with Metastatic Non–Small-Cell Lung Cancer: Results of a Randomized Study of Early Palliative Care," Journal of Clinical Oncology, Vol. 29, No. 17, June 2011, pp. 2319–2326.

Temel, Jennifer S., Joseph A. Greer, Areej El-Jawahri, William F. Pirl, Elyse R. Park, Vicki A. Jackson, Anthony L. Back, Mihir Kamdar, Juliet Jacobsen, Eva H. Chittenden, Simone P. Rindaldi, Emily R. Gallagher, Justin R. Eusebio, Zhigang Li, Alona Muzikansky, and David P. Ryan, "Effects of Early Integrated Palliative Care in Patients with Lung and GI Cancer: A Randomized Clinical Trial," Journal of Clinical Oncology, Vol. 35, No. 8, March 2017, pp. 834–841.

[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]

The same sample was used for all aspects of testing with the exception of a small subset of respondents who also participated in a test-retest design to provide additional reliability evidence for data elements. For the test-retest reliability calculation, we obtained data from a subset of respondents at two timepoints. We invited patients who completed the survey by phone (i.e., the Computer-Assisted Telephone Interview [CATI] survey) to be re-administered a shortened survey, including the *Feeling Heard and Understood* data elements, at a second timepoint within two days of the original CATI interview. Once patients were invited, participation in Time 2 was based on willingness and availability of the identified patient respondents. By restricting to telephone-only patients, the time interval could be minimized and

standardized (i.e., two days), and the mode of administration would be the same at both data collection time points. Only patient respondents who completed the original CATI survey without proxy assistance were invited to participate in the retest. This subset included 437 respondents at Time 1, with 235 of these respondents also providing responses at Time 2. Respondents with data at both timepoints were included in these analyses. For all other analyses, we used the first responses collected from the test-retest participants. Our analysis of test-retest reliability was intended to establish data element level reliability; future work may need to examine test-retest reliability across survey modes and populations.
[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]
Based on input from our project advisory group and technical expert clinical user and patient panel (TECUPP), we determined it was not appropriate to adjust this measure for social risk factors. However, to understand if and to what extent disparities in measure reporting and patient experience exist, we evaluated the relationship of various social risk factors to the measure score and the programs. These included patient race/ethnicity, education, and primary language, as well as multiple census-level variables such as race/ethnicity, urbanicity, median household income, gender, marital status, public insurance use, unemployment, and families below poverty line (see section 2b.23 for details). After adjustment for multiple comparisons, none of these variables were significant in their relationship with the measure.
[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter “see validity testing section of data elements”; and enter “N/A” for 2a.09 and 2a.10.

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)
Accountable Entity Level (e.g., signal-to-noise analysis)
[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]
We first evaluated Data Element Reliability:

The reliability of the *Feeling Heard and Understood* multi-data element scale was evaluated using both internal consistency and test-retest reliability coefficients. We used Cronbach’s alpha as a measure of internal consistency reliability; acceptable reliability was based on conventional criteria of Cronbach’s alpha of 0.70 (Nunnally, 1978). For the test-retest reliability calculation, we obtained data from a subset of respondents at two timepoints. We invited patients who completed the survey by phone (i.e., the Computer-Assisted Telephone Interview [CATI] survey) to be re-administered a shortened survey, including the *Feeling Heard and Understood* data elements, at a second timepoint within two days of the original CATI interview. Once patients were invited, participation in Time 2 was based on willingness and availability of the identified patient respondents. This subset included 437 respondents at Time 1, with 233 of these also providing responses at Time 2. From these 233 respondents, 36 were ultimately determined to be ineligible for the measure analyses, resulting in a final subsample of 197 respondents with complete data for the test-retest analysis. By restricting to telephone-only patients, the time interval could be minimized and standardized (i.e., two days), and the mode of administration would be the same at both data collection time points. Only patient respondents who completed the original CATI survey without proxy assistance were invited to participate in the retest. We evaluated test-retest reliability with a stability coefficient (i.e., correlation coefficient) wherein scores from the initial administration (Time 1) were compared against scores from a second administration (Time 2). Scores from both administrations were compared using a polychoric correlation coefficient, with a correlation of at least 0.70 demonstrating acceptable reliability.

We then evaluated Quality Measure Score Reliability:

To assess the reliability of the quality measure score at the program level, we used a traditional "signal-to-noise" analysis that decomposes variability in the measure score into a) between-subject variability and b) within-subject variability. If there is a large amount of between-subject variability (i.e., "signal") compared to within-subject variability (i.e., "noise"), then there is more evidence that it is possible to discriminate performance among providers. To measure variability, we used hierarchical generalized-linear regression models to relate our outcome measures to our programs and their covariates, where the hierarchy of data is patient observations within the program. We performed hierarchical regressions with binomial outcomes to decompose the variability. The random effects model for analysis across providers is

$$logit\left(E\left[Y_{ij}\middle|P_i, X_{ij}\right]\right) = (\beta_0 + b_i P_i) + X_{ij}^T\alpha$$

where we assume that $b_i \sim N(0, \sigma_b^2)$. In this model, the term β_0 represents the overall average performance, each term P_i is an indicator of provider i and therefore b_i represents the provider-specific offset from the overall performance and ${X_{ij}}^T\alpha$ captures risk adjustment (specifically for survey mode and proxy measures). The variance of the model can be decomposed using the (adjusted) intraclass correlation coefficient (ICC), which provides a summary of the reliability of the measure as tested, with higher values implying more variability between programs (Rodríguez & Elo, 2003; Wu et al., 2012):

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi}{3}}$$

We incorporate risk adjustment variables into our models to provide fair comparisons among programs and to provide a best effort to ensure that the observed differences from programs are truly from differences in performance and not due to baseline differences in risk variables (including survey mode) that represent the programs. The reliability from the measure test is then projected out based on observed variances and sample sizes from each program, using the Spearman-Brown prophecy formula. This allows us to estimate a required within-program sample size to achieve a desired reliability for the measure. Reliability values of approximately 0.7 reflect an acceptable level of reliability and guided determination of required within-provider sample sizes (Nunnally, 1978).

Citations:

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed. ed.). McGraw-Hill.

Rodríguez, G., & Elo, I. (2003). Intra-class Correlation in Random-effects Models for Binary Data. *The Stata Journal*, *3*(1), 32-46.

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp Clin Trials*, *33*(5), 869-880.

Response to NQF request for clarification: All recommended modes of data collection (i.e., web, mail, and phone) were tested in the national field test, though we did not conduct a full mode experiment. A subset of respondents were also asked to participate in a test-retest design to establish data element reliability. To analyze test-retest reliability, we restricted to telephone-only so that the time interval and mode of administration could be standardized, and telephone surveys could be completed close together so that the patient was reflecting on the same time period/visits.

Upon full submission, we will clarify that the recommended mode of implementation is enhanced mixed-mode administration (web to mail to phone, i.e., live telephone interview).

[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]

Data Element Reliability:

Results provide support for the reliability of scores derived from the four-data element *Feeling Heard and Understood* scale (Cronbach's alpha = 0.90). Test-retest reliability (i.e., polychoric correlation coefficient) for the *Feeling Heard and Understood* total score was 0.85. Reliability of the four-data element scale was high indicating that scores obtained are in fact reliable and can, therefore, be used in the construction of the quality measure.

Quality Measure Score Reliability:

We conducted a formal measure score reliability analysis using Bayesian generalized mixed-effects models to obtain a posterior distribution of the ICC. The estimate of the adjusted ICC is approximately 0.052 (95% CI: 0.027 to 0.089) (Figure 2). This implies that there is a reasonable level of between-program variability as compared to the within-program variability.

Figure 2. Estimated Posterior Distribution of the Adjusted ICC

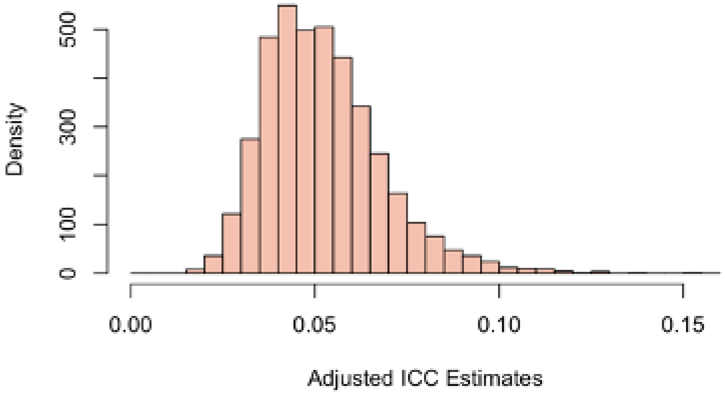


Figure 3. Estimate Posterior Distribution of Reliability at Fixed Sample Sizes

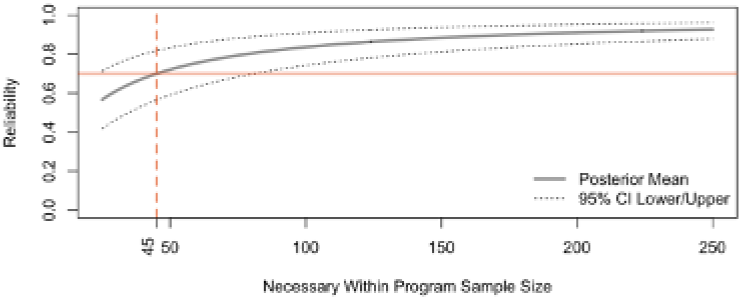
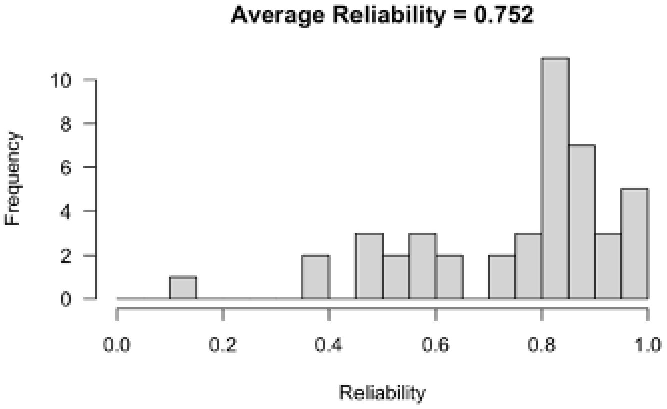


Figure 4. Distribution of Program-Specific Reliability



Adams, J. L. (2009). *The Reliability of Provider Profiling: A Tutorial*. https://www.rand.org/pubs/technical_reports/TR653.html

[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]
Testing results provide support for the reliability of scores derived from the four-data element *Feeling Heard and Understood* scale. Test-retest reliability for the *Feeling Heard and Understood* total score was high. Reliability of the four-data element scale was high indicating that scores obtained are in fact reliable and can, therefore, be used in the construction of the quality measure. Results of the “signal-to-noise” analysis of quality measure reliability suggest there is a reasonable level of reliability based on the observed between-program variability and the within-program variability, given our sample sizes.
[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)
Accountable Entity Level (e.g. hospitals, clinicians)
Empirical validity testing
Systematic assessment of face validity of performance measure score as an indicator of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)
[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 first assessed Data Element Validity:

As there is currently no gold standard to compare with our *Feeling Heard and Understood* multi-data element scale, our data element validity analyses did not address sensitivity and specificity but instead focused on assessing the convergent validity of the proposed scale. We hypothesized that the *Feeling Heard and Understood* scale is theoretically related to similar constructs and thus sought to assess whether they were indeed related. As part of the study design we included additional survey data elements taken from other instruments in use (e.g., the Consumer Assessment of Healthcare Providers and Systems [CAHPS] Hospice Survey) expected to be related to *Feeling Heard and Understood*. Selection of additional data elements was based on theory, prior literature, and clinical practice and/or expert feedback. For example, the four-item CAHPS Communication composite measure (e.g., "In the last 3 months, how often did this provider and team listen carefully to you?") was hypothesized to be associated with the proposed *Feeling Heard and Understood* data element. We also examined the association between *Feeling Heard and Understood* and *Receiving Desired Help for Pain* (Included data element: "In the last 6 months, did you get as much help as you wanted for your pain from this provider and team?" [Yes, definitely; Yes, somewhat; No]). We hypothesized that receiving the help one wanted for their pain from their palliative care provider and team would be linked with feeling heard and understood by that same palliative care provider and team. Associations between the proposed data elements and validity items were evaluated using bivariate correlations. Interpretation of correlations followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50) (Rosnow & Rosenthal, 1989).

We then assessed Quality Measure Score Validity:

To evaluate the validity of the *Feeling Heard and Understood* quality measure, we examined the association of the *Feeling Heard and Understood* measure score with the *Receiving Desired Help for Pain* measure score, the CAHPS communication measure score, and individual's overall rating of their palliative care provider and team, with the hypothesis that scores would be positively associated. Associations between the quality measures were evaluated using bivariate correlations. Interpretation of correlations followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50).

Face validity of the quality measure score was determined through a systematic and transparent process by convening experts who explicitly addressed whether scores resulting from the measure, as specified, can be used to distinguish good from poor quality. In May 2021, following completion of testing, a panel of seven advisors with expertise in palliative care and clinical quality measurement were asked to review the final measure specifications and testing results and rate face validity of the measure score. The expert panel consisted of six palliative care physicians and a researcher with expertise in palliative care communication. Six of the seven advisors also had expertise in clinical quality measurement. Advisors were asked to consider how well the measure scoring approach distinguishes between programs with high, medium, and low performance and how useful it is to quality improvement efforts. Advisors rated face validity on a scale of 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9).

Citations:

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276-1284.

Response to NQF request for clarification: None of the seven experts who completed the face validity exercise were part of the core measure development and testing team. The advisory board is an external group of subject matter experts, including measure developers and palliative care clinicians, who provided input on measure specification and field-testing decisions. We met with the advisory board at specific points during measure development to collect their input. After testing was completed and measure specifications were finalized, advisors were asked to provide their objective ratings of face validity based on their review of the final measure specifications and testing results.

[Response Ends]

2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]
Statistical results include correlations.

Data Element level:

Results of validity testing at the data element level provide evidence supporting the use of the *Feeling Heard and Understood* scale score. As hypothesized, higher scores on the *Feeling Heard and Understood* scale were associated with higher CAHPS communication scores (r = 0.54, p<.001) as well as with *Receiving Desired Help for Pain* (r = 0.48, p< .001). Taken together, these results support the convergent validity of the *Feeling Heard and Understood* data elements.

Quality Measure level:

Results of validity testing at the quality measure level provide evidence supporting the use of the *Feeling Heard and Understood* quality measure as constructed. As hypothesized, the *Feeling Heard and Understood* quality measure was significantly and positively associated with the CAHPS communication quality measure (r = 0.635, p =0.011), the *Receiving Desired Help for Pain* quality measure (r = 0.496, p<.001) and the overall rating of the palliative care provider and team (r= 0.768, p=<.001).

Seven expert advisors rated face validity of the *Feeling Heard and Understood* measure score. On average, advisors rated face validity of the measure score 8.3 on a scale of 1-9, corresponding with an average rating of “high.” These ratings reflect strong support for face validity of the proposed quality measure from experts in palliative care and quality measurement.
[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]
Testing results support the convergent validity of both the *Feeling Heard and Understood* 4-data element scale and the *Feeling Heard and Understood* quality measure. Further expert ratings reflect strong support for the face validity of the proposed quality measure.
[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]
We guided our analyses based on the literature and lessons learned from the CAHPS program (Quigley et al., 2018). For example, we tested for statistically significant differences among programs using techniques similar to analysis of variance (ANOVA) that aim to compare a “full model” and a “reduced model” (or “nested model”). The reduced (or nested) model assumes that there are no differences among the programs, and the full model assumes that there is at least one difference. Under a generalized linear model, the above null and alternative hypotheses can be tested using a likelihood ratio statistic.

As the concept of clinically meaningful or practical significance difference is a notion without perfect agreement among researchers on how it should be defined, we assessed such difference with a ranking approach that uses all programs that participated in the national beta field test. To that end, we estimated equivalence of a difference in measure score to ranking of program performance. This equivalence method is intended to relate the magnitude of difference in the program’s score to its ranking and potentially gives patients and decisionmakers a magnitude that can have practical choice implications for them.

Citations:

Quigley, D. D., Elliott, M. N., Setodji, C. M., & Hays, R. D. (2018). Quantifying Magnitude of Group-Level Differences in Patient Experiences with Health Care. *Health services research*, *53 Suppl 1*(Suppl Suppl 1), 3027-3051.

[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]
We found evidence of statistically significant differences in program scores; however, interpreting the meaning of those differences requires information about both the score and the rank order. For score differences, Figure 5 shows the adjusted measure scores for each program with the dot indicating the mean score and the extended lines to left and right of the dot indicating the variability in scores within each program. The lines extend to reflect a 95 percent confidence interval; thus, programs that are statistically different from one another are represented by non-overlapping lines. From the figure we see that there is diversity in program measure scores; for example, the lowest scoring program does not overlap the highest scoring program. This provides evidence that there are significant differences in program performance on the measure.

Figure 5. Posterior Distribution of Adjusted Program Scores

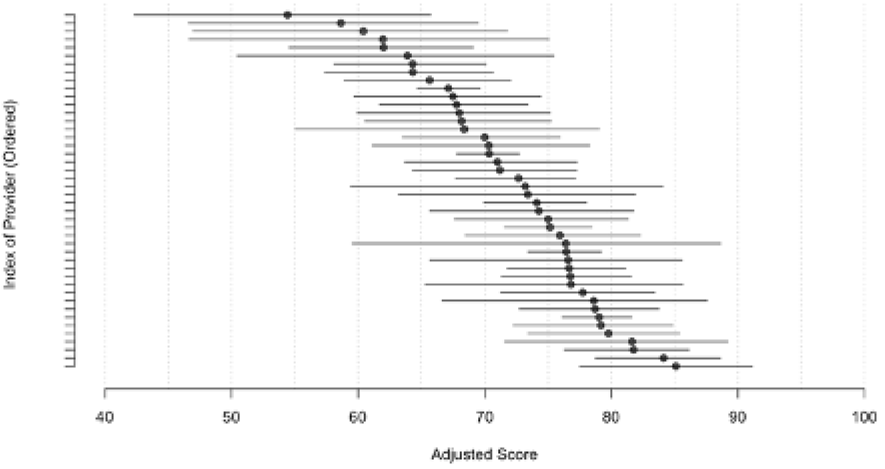


Table 6: Change in Measure Score, Assuming 100 Ranked Programs

Program Rank Difference	Change in Score
5th top to 10th top	2.05 points drop
10th top to 20th top	2.83 points drop
Median to 20th top	4.19 points increase
20th lowest to Median	7.28 points increase
10th lowest to Median	10.90 points increase

[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]
When programs are ranked by their measure performance, we calculated that a program at the median of measure performance would need a large increase of 4.19 points in their measure score to improve to the 20th top-ranked program. A program at the bottom of the ranking (e.g., the 10th lowest ranked program) would need a 7-point increase measure score to improve to the median. These changes in score by program rank suggest that measure performance is actionable; i.e., there is room for programs to improve their score.

Face validity of the quality measure scores was determined through a systematic and transparent process by convening experts who explicitly addressed whether scores resulting from the measure, as specified, can be used to distinguish good from poor quality. In May 2021, following completion of testing, a panel of seven advisors with expertise in palliative care and clinical quality measurement were asked to review the final measure specifications and testing results and rate face validity of the measure score. Providers were asked to consider how well the approach distinguishes between programs with high, medium, and low performance and how useful it is to quality improvement efforts. Advisors rated face validity on a scale of 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9). On average, advisors rated face validity of the measure score for *Feeling Heard and Understood* 8.3 on a scale of 1-9, corresponding with an average rating of "high." These ratings reflect strong support for face validity of the proposed quality measure from experts in palliative care and quality measurement.
[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]
Missing data in the outcome is naturally accommodated among the four response items by the modeling procedure for adjusted score estimation; thus no outcome imputation is necessary and should not be performed. Missing values for proxy assistance should be imputed as "No Proxy Assist."

Within palliative care programs in the beta field test, we assessed the distribution of missing data (i.e., not responding to specific questions) and nonresponse (i.e., not responding to the survey) to assess their impact on utilizing the proposed measure using statistical tests (e.g., t-tests to compare distribution means, Kolmogorov-Smirnoff statistics to assess cumulative distribution functions) that provide quantifications of the discrepancies between distributions.

To better understand the potential for bias due to nonresponse, we used available patient data to characterize the differences between respondents (n=2,804) and non-respondents (3,356). Age and gender were available for all patients as they were included in the data files provided to us by participating programs; we compared mean age using a two-sample t-test and gender using a chi-squared test. Patient race was collected via self-report on the survey instrument, but a subset of participating programs provided race for at least 90% of their patients in their submitted data files. We compared patient race within this subset of programs between respondents and nonrespondents using a chi-squared test.

We also examined missing data among completed surveys, such as not responding to individual items or demographic questions. To handle this, we again assessed the distributions and patterns of missing data, including an assessment to see if a missing-at-random assumption seemed plausible.
[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]
We fielded the survey to 7,595 sampled patients in the beta test across 10 rounds. Of these, 3,356 were not returned, 1,435 were excluded from any analyses due to ineligibility for the larger study, and 2,804 were returned and included in analyses (37% raw response rate; 46% response rate excluding ineligible patients). Completed surveys are defined as any survey returned within six months of lookback start date that was not excluded due to ineligibility (e.g., surveys sent to patients who were later identified as deceased, surveys completed entirely by a proxy respondent, or surveys to patients

who disavowed the receipt of care). Completed surveys may still have item-level missingness. The 2,804 completed surveys reflect a patient sample that was largely female (56%), White (88%) and non-Hispanic or Latino (95%), and very educated, with 66% having some college or more. The overall level of nonresponse to fielded surveys (after removing exclusions) was approximately 54.4%.

We also examined the distribution of nonresponse across programs, comparing responders to non-responders (but removing those ineligible responders due to exclusions). There were no clear outliers in terms of program nonresponse. The proportion of women was higher among respondents, compared to nonrespondents (56.2% vs 54.5%); the results from a chi-squared test indicates that this difference is not statistically significant (p = 0.21). Additionally, survey respondents were slightly older than patients who did not complete a survey (mean age 63.4 vs 60.9; p<0.01). Though significant, patient age was not significantly related to response patterns (see section 2b.24, Table 9) and therefore we do not believe that non-response related to age should significantly bias results.

Finally, among the subset of 12 programs who provided patient race for at least 90% of their patients, respondents were more likely to identify as White (88.1% vs 80.2%) and less likely to identify as Black (8.8% vs 11.9%) or another race (3.1% vs 8%) compared to nonrespondents. The results of a chi-squared test indicate that this difference is statistically significant (p < 0.01). Though significant, as with age, patient race was not significantly related to the *Feeling Heard and Understood* data element responses (see section 2b.24, Table 9), and therefore we do not think non-response related to race should significantly bias results. However, due to inconsistency in reporting of the race variable across programs for non-responders, nonresponse due to race will need to be evaluated in the future to explore the need for adjustment with nonresponse weights.

Among the 2,804 completed surveys from the beta field test, the mean item-level missingness was 0.8% across the entire survey. Appropriately skipped survey items are not counted as missing. Among the four data elements that comprise the *Feeling Heard and Understood* measure, the data element-level missingness among responders was relatively low ranging from 1.07% (Q3: "I felt this provider and team saw me as a person, not just someone with a medical problem") to 1.53% (Q4: "I felt this provider and team understood what is important to me in my life"). We also examined patterns in missingness among responders for each grouping of the four *Feeling Heard and Understood* data elements. Among responders who have any missing data across the four data elements (N=60), approximately 43% (n=26) skipped all of the measure data elements (N=26), while 47% (n=28) skipped only individual questions). This suggests there are low levels of missingness and no discernible patterns.
[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]
Overall, we had good survey response rates with levels comparable to those typical to CAHPS surveys; there were no clear outliers in terms of program nonresponse; there were low levels of missingness in completed surveys; and our non-response analysis did not identify any evidence of systematic bias (see 2b.09 for details). This conclusion is largely supported by the fact that many of the candidate variables with differences in non-response are unrelated to outcomes (see section 2b.24, Table 9).
[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 eCQMs). 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]
Yes, the measure uses 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]
We considered five exclusions from the proposed denominator of all adult patients with an ambulatory palliative care visit:

1. Patients who do not complete at least one of the four items in the multi-item measure (n=26).
2. Patients who do not complete the patient experience survey within six months of the eligible ambulatory palliative care visit (n=3,356).
3. Patients who respond on the patient experience survey that they did not receive care by the listed ambulatory palliative care provider in the last six months (disavowal; n=146).
4. Patients who were deceased when the survey reached them (n=748).

5. Patients for whom a proxy completed the entire survey on their behalf for any reason (no patient involvement; n=435).

Patients who did not respond to least one of the four items in the multi-item measure, return the survey at all, those who disavowed the program, or those who died before the survey could be completed were necessary exclusions because no survey data for the quality measure would be available. We further excluded the small number of patients who did not return the survey within the 6-month timeframe because of concerns regarding recall bias and due to their likely minimal impact (n=61 out of 3,356 nonrespondents, or 1.8%). Although we could not analyze the impact on measure outcomes of excluding these groups because of the absence of survey data, we did compare respondent and nonrespondent (n=3,356) characteristics.

Our exclusion analysis primarily focused on exploring the impact of proxy-involved survey data. Existing CAHPS surveys (e.g., Medicare CAHPS, Prescription Drug Plan CAHPS) use proxy response as a case-mix adjustment variable; despite evidence that proxy response contributes only weakly to differences in measure scores it is retained to alleviate ongoing concerns about the potential for impact. Respondents were categorized into three distinct groups based on proxy assistance as follows: respondent only (no proxy assistance at all), proxy assisted (proxy helped patient complete the survey but patient supplied answers e.g., proxy read questions and wrote down answers), proxy only (proxy answered all questions and patient was not involved) (CMS, 2020; National Cancer Institute, 2021). We compared descriptive statistics for the measure components for each of these three groups to inform the impact of proxy assistance and to determine whether to include/exclude proxy responses.

Citations:

Centers for Medicare & Medicaid Services. (2020). *Medicare Advantage and Prescription Drug Plan CAHPS® Survey: Quality Assurance Protocols & Technical Specifications V11.0*.

National Cancer Institute. (2021). *Case Mix Adjustment Guidance*. Retrieved May 20, 2021, from https://healthcaredelivery.cancer.gov/seer-cahps/researchers/adjustment_guidance.html

[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]

Survey respondents were slightly older than patients who did not complete a survey (mean age 63.4 vs 60.9; p<0.01). The portion of women was also higher among respondents, compared to nonrespondents (56.2% vs 54.5%); the results from a chi-squared test indicates that this difference is not statistically significant (p = 0.21). Among the subset of 12 programs who provided patient race for at least 90% of their patients, respondents were more likely to identify as White (88.1% vs 80.2%) and less likely to identify as Black (8.8% vs 11.9%) or another race (3.1% vs 8%) compared to nonrespondents. The results of a chi-squared test indicate that this difference is statistically significant (p < 0.01). As age and race may differentially impact patient experiences of care, future work to improve response rates among specific demographic groups, such that measure performance may more robustly reflect the experiences of all patients, is important, though out of scope of the current testing effort.

We had a total of 2548 completed surveys by patients without proxy assistance, 224 completed by patients with proxy assistance, and 430 completed by proxies alone with no patient involvement. Table 7 shows the mean (SD) *Feeling Heard and Understood* measure score according to these three groups (responses restricted to those patients who answered at least 3 out of the 4 *Feeling Heard and Understood* data elements for analytic purposes).

Table 7. Mean *Feeling Heard and Understood* Measure Score According to Proxy Group

Group	N*	Mean (SD)
Patient only	2548	0.71 (0.37)
Proxy assisted	224	0.77 (0.34)
Proxy only	430	0.69 (0.37)

*N reflects patients who responded to at least 3 out of the 4 *Feeling Heard and Understood* data elements comprising the multi-item measure

A one-way ANOVA test for differences among these three means was significant ($F_{(2, 3199)}$=3.80, *p*=.023), and follow-up pairwise mean comparisons revealed no difference between patient only and proxy only ($t_{(581)}$=1.22, *p*=0.22). However, proxy assisted was significantly different from both patient only ($t_{(271)}$=-2.48, *p*=0.01) and proxy only ($t_{(487)}$=-2.86, *p*=0.004).

[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]
Despite the lack of a significant difference in *Feeling Heard and Understood* measure score means between the proxy only and patient only groups, after discussing these results with our advisory board, we decided to exclude surveys that were completed solely by a proxy with no patient involvement for conceptual reasons. As a patient-reported measure of palliative care experience, we wanted to ensure that at least some direct patient report was reflected in the measure response, a rationale for excluding proxy-only responses that was endorsed by the advisory board. We elected to include proxy-assisted surveys and to add an adjustment for proxy assistance to account for small differences in measure components due to the proxy involvement. This allowed us to retain as much patient-reported data as possible, while acknowledging that patients in this population will likely need some assistance with survey completion.
[Response Ends]

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

[Response Begins]
Statistical risk model with risk factors (specify number of risk factors)
2 risk factors: 1) survey mode and 2) an indicator of proxy assistance
[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]
The measure is risk adjusted for 1) survey mode and 2) an indicator of proxy assistance. To estimate risk-adjusted quality measure scores, we utilize hierarchical generalized-linear models that relate the proportion of top-box patient-level outcome responses to provider scores (conditioned on risk adjustment covariates). The hierarchy of data is patient observations within the designated accountable health care entity, i.e., programs. The model is calculated at all baseline covariate values of the model (i.e., with risk adjustment indicators set to 0).
[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]
N/A
[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]
Published literature
Internal data analysis
[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]
As suggested by our technical expert clinical user and patient panel (TECUPP) and project advisory group, we sought to develop a measure for which high performance would reflect the intrinsic quality of care. We considered that adjustment of the measure calculation might be appropriate to account for differences in performance due to extrinsic factors beyond the control of the palliative care program or provider. Relevant factors are those that systematically differ across programs and also are related to the measure score. To identify the latter, we conducted a broader literature review of palliative and serious illness care assessments to identify patient, provider, and practice factors that could impact a patient's experiences of *Feeling Heard and Understood.* We examined the following factors:

- **Patient Age, Education, Financial, Physical, and Mental Health**: Research suggests that age, education, and self-reported health can impact experiences of care and/or response tendencies (Elliott et al., 2001; Elliott et al., 2009; Ingersoll et al., 2018; Zaslavsky et al., 2001). For example, Ingersoll et. al. found that among individuals hospitalized with metastatic cancer, older age, financial security, and low emotional distress were some of the factors associated with feeling completely heard and understood prior to palliative care consultation (Ingersoll et al., 2018). Studies using CAHPS Hospice Survey data show that respondents with lower education, older age, and better self-reported mental health tended to report higher patient experience ratings (Anhang Price et al., 2014). For pain specifically, patient age, sex, comorbid anxiety/depression, and health insurance have been found to impact unmet needs including pain management (John et al., 2014).
- **Patient Race and Ethnicity:** The relationship between race and ethnicity and experiences of patient-provider communication is mixed. For example, some studies have found that Black patients and their family members report having greater communication issues and give lower communication ratings of their health care providers than their White counterparts (Welch et al., 2005), while other studies have found the opposite relationship (Coats et al., 2018).
- **Proxy Response:** Prior CAHPS survey research has also demonstrated that proxy respondents tend to give lower ratings than non-proxy respondents (O'Malley et al., 2005).

Based on our literature review and guided by TECUPP and project advisory group feedback, we selected a set of potential risk factors, shown below, to be evaluated in our final models. Data were selected from three sources: patient information provided by programs in their submitted data files, responses from completed surveys (e.g., proxy respondent characteristics), and Census data matched to the ZIP code of patient residence. Variables were assessed for inclusion in a risk model using statistical testing to determine whether each variable was associated with the measure and whether it differed substantially across programs, at the a=0.05 level of significance. Binary or categorical variables were tested for association with the *Feeling Heard and Understood* measure using Fisher's exact test. Fisher's exact test was also used to assess if there were differences in these variables across programs. For continuous variables, a Z-test from a generalized linear model was used to test for an association with the *Feeling Heard and Understood* measure, and an ANOVA F-test was used to test for differences between programs. P-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons to control the false discovery rate (Benjamini & Hochberg, 1995).

Potential risk adjusters:

- Survey data or program information
  - Patient education
  - Patient Hispanic
  - Patient language
  - Patient race
  - Proxy level
  - Survey mode
- Provider data
  - Patient age
  - Patient female
- Census data
  - Female residents (%)
  - Marriage status (% of residents age 15+)
  - Disabled residents (%)
  - Labor force participation (% of residents age 16+)
  - Employed residents (% of residents age 16+)
  - Unemployed residents (% of residents age 16+)
  - Median household income ($)
  - Families below the poverty line (%)
  - Urban residents (%)
  - Owner-occupied housing units (%)
  - Residents with health insurance (%)
  - Private health insurance only (% of non-institutionalized residents)
  - Public health insurance only (% of non-institutionalized residents)
  - Race = American Indian or Alaska Native (%)
  - Race = Asian (%)
  - Race = Black (%)
  - Race = White (%)
  - Ethnicity = Hispanic (%)

We also considered the potential for patient diagnoses to vary across programs and impact measure scores. In the development of the Hospice CAHPS measures, investigators found that primary diagnosis varied across hospice programs and were significantly and strongly associated with assessments of experience and were thus included as a case-mix adjustment variable (Parast et al., 2018). Although the target respondent population and setting were different than ours (bereaved caregiver vs. patient; hospice vs. ambulatory palliative care), both the Hospice CAHPS measures and our proposed measure seek to assess the patient's experience of palliative care. However, we were severely constrained in our ability to explore potential risk adjustment by diagnosis because of the inadequacy of diagnosis data we received from programs in their submitted files:

- Not all programs consistently provided diagnostic information across the 10 rounds of fielding;
- Programs that did provide any data typically did not clarify primary, secondary, or other diagnosis, but instead provided multiple diagnoses per patient;
- We received different types of diagnostic information in submitted data files, both within programs and across programs, including ICD-10 codes, CPT codes, problem codes, reasons for visit codes, and program- or software-specific codes, and sometimes free text.

Instead, we undertook an exploratory descriptive analysis to identify any signals that measure performance might vary by diagnosis and thus that future work should assess the role of diagnosis as a risk-adjustment variable.

Using our full survey respondent sample of 2,804, we assigned primary diagnoses in a two-step process: 1) where primary diagnosis was indicated in the program file, we used that; and 2) if primary diagnosis was not indicated but usable diagnosis data was provided, we assigned primary diagnosis by applying a condition hierarchy based on prevalence in our sample, prior research (Keating et al., 2016; Wachterman et al., 2016; Wennberg et al., 2004), and our research team physicians' expert opinion:

a. cancer (both solid and liquid);
b. non-neurologic end-organ disease (e.g. heart failure, end-stage liver disease, renal failure, chronic obstructive pulmonary disease);
c. dementia (e.g. Alzheimer's, vascular, and others);
d. movement disorders (e.g. stroke, Parkinson's, multiple sclerosis, ALS); and
e. other diagnosis (e.g. fibromyalgia, sequelae of diabetes, AIDS, symptom only diagnoses such as dyspnea)

All other data was categorized as missing (e.g., non-interpretable diagnostic information). A physician then reviewed these group to ensure clinical accuracy. See Table 8 for counts of each assigned diagnosis group.

Table 8. Diagnosis Groupings

Diagnosis grouping	**Count**	**Percentage** (numbers do not add to 100 due to rounding)
Cancer (solid and liquid)	1685	60%
End-organ disease (non-neurological)	225	8
Dementia	14	0.5
Movement disorders (e.g. Parkinson's, multiple sclerosis)	29	1
Other	561	20
Missing	290	10
Total = 2804		

Cell left intentionally blank.

Because of the low numbers of respondents with assigned primary diagnosis of dementia or movement disorders, we focused on comparing differences between the cancer, end-organ disease, and other diagnosis groups for our analysis (i.e., ignoring the dementia, movement disorders, and missing groups). For each of the four individual *Feeling Heard and Understood* data elements we performed a chi-squared test for independence between the data element and the assigned diagnosis and conducted an ANOVA F-test for the *Feeling Heard and Understood* raw score (% of top-box response) by diagnosis group.

Citations:

Anhang Price, R., Elliott, M. N., Zaslavsky, A. M., Hays, R. D., Lehrman, W. G., Rybowski, L., Edgman-Levitan, S., & Cleary, P. D. (2014). Examining the role of patient experience surveys in measuring health care quality. *Medical care research and review : MCRR*, *71*(5), 522-554.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300.

Coats, H., Downey, L., Sharma, R. K., Curtis, J. R., & Engelberg, R. A. (2018). Quality of Communication and Trust in Patients With Serious Illness: An Exploratory Study of the Relationships of Race/Ethnicity, Socioeconomic Status, and Religiosity. *J Pain Symptom Manage*, *56*(4), 530-540.e536.

Elliott, M. N., Swartz, R., Adams, J., Spritzer, K. L., & Hays, R. D. (2001). Case-mix adjustment of the National CAHPS benchmarking data 1.0: a violation of model assumptions? *Health services research*, *36*(3), 555-573.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K., & Giordano, L. (2009). Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. *Health services research*, *44*(2 Pt 1), 501-518.

Ingersoll, L. T., Saeed, F., Ladwig, S., Norton, S. A., Anderson, W., Alexander, S. C., & Gramling, R. (2018). Feeling Heard & Understood in the Hospital Environment: Benchmarking Communication Quality Among Patients with Advanced Cancer Before and After Palliative Care Consultation. *J Pain Symptom Manage*, *56*(2), 239-244.

John, D. A., Kawachi, I., Lathan, C. S., & Ayanian, J. Z. (2014). Disparities in perceived unmet need for supportive services among patients with lung cancer in the Cancer Care Outcomes Research and Surveillance Consortium. *Cancer*, *120*(20), 3178-3191.

Keating, N. L., Landrum, M. B., Huskamp, H. A., Kouri, E. M., Prigerson, H. G., Schrag, D., Maciejewski, P. K., Hornbrook, M. C., & Haggstrom, D. A. (2016). Dartmouth Atlas Area-Level Estimates of End-of-Life Expenditures:

How Well Do They Reflect Expenditures for Prospectively Identified Advanced Lung Cancer Patients? *Health services research*, *51*(4), 1584-1594.

O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., & Cleary, P. D. (2005). Case-mix adjustment of the CAHPS Hospital Survey. *Health services research*, *40*(6 Pt 2), 2162-2181.

Parast, L., Haas, A., Tolpadi, A., Elliott, M. N., Teno, J., Zaslavsky, A. M., & Price, R. A. (2018). Effects of Caregiver and Decedent Characteristics on CAHPS Hospice Survey Scores. *Journal of pain and symptom management*, *56*(4), 519-529.e511.

Wachterman, M. W., Pilver, C., Smith, D., Ersek, M., Lipsitz, S. R., & Keating, N. L. (2016). Quality of End-of-Life Care Provided to Patients With Different Serious Illnesses. *JAMA Intern Med*, *176*(8), 1095-1102.

Welch, L. C., Teno, J. M., & Mor, V. (2005). End-of-life care in black and white: race matters for medical care of dying patients and their families. *J Am Geriatr Soc*, *53*(7), 1145-1153.

Wennberg, J. E., Fisher, E. S., Stukel, T. A., Skinner, J. S., Sharp, S. M., & Bronner, K. K. (2004). Use of hospitals, physician visits, and hospice care during last six months of life among cohorts loyal to highly respected hospitals in the United States. *BMJ (Clinical research ed.)*, *328*(7440), 607.

Zaslavsky, A. M., Zaborski, L. B., Ding, L., Shaul, J. A., Cioffi, M. J., & Cleary, P. D. (2001). Adjusting Performance Measures to Ensure Equitable Plan Comparisons. *Health care financing review*, *22*(3), 109-126.

[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]
Table 9 below shows the test statistics for the association of potential risk factors with measure scores and with programs. All but one of the potential risk adjustment variables were found to have significant or nearly significant associations with programs at the a=0.05 level of significance, indicating that the distributions of these variables were not identical across programs. This implies that, if these variables affected measure scores, they could cause scores to differ across programs and thereby confound the relationship between quality of care and measure score. However, only one of the potential risk adjustment variables was significant in its relationship with the measure after adjustment for multiple comparisons. In particular, survey mode was the only variable to demonstrate statistically significant associations with both the measure (p=0.013) and with programs (p=0.001) after the multiple-testing adjustment, thus leaving survey mode as the sole remaining candidate variable for adjustment. For reference, measure score means and standard deviations by survey mode are given in Table 10. As seen in the table, scores did not differ much between mail and web responses but were slightly lower for responses by phone.

Table 9. Association of Potential Risk Factors with Feeling Heard and Understood Measure Score and Program (sorted by measure score p-values)

	Association with Measure Score					Association with Program			
Potential Risk Adjuster	Effect Size	Test	Statistic	p-value	Adjusted p-value	Test	Statistic	p-value	Adjusted p-value
Survey data									
Survey Mode	NA	Fisher	NA	0.000	0.013	Fisher	NA	0.000	0.001
Patient Race	NA	Fisher	NA	0.049	0.168	Fisher	NA	0.000	0.001
Proxy Assistance	NA	Fisher	NA	0.242	0.420	Fisher	NA	0.686	0.686
Patient Education	NA	Fisher	NA	0.271	0.432	Fisher	NA	0.000	0.001
Patient Hispanic	NA	Fisher	NA	0.541	0.732	Fisher	NA	0.000	0.001
Patient Language	NA	Fisher	NA	0.656	0.732	Fisher	NA	0.002	0.002

	Association with Measure Score					Association with Program			
Potential Risk Adjuster	Effect Size	Test	Statistic	p-value	Adjusted p-value	Test	Statistic	p-value	Adjusted p-value
Program data									
Patient Age	-0.005	Z	-1.754	0.080	0.207	F	6.780	0.000	0.000
Patient Female	NA	Fisher	NA	0.704	0.732	Fisher	NA	0.051	0.054
Census data									
Owner occupied housing unit	0.006	Z	2.624	0.009	0.114	F	6.437	0.000	0.000
Median household income	0.000	Z	2.429	0.015	0.132	F	17.682	0.000	0.000
Female	0.034	Z	2.202	0.028	0.144	F	4.591	0.000	0.000
Family below poverty line	-0.012	Z	-2.203	0.028	0.144	F	12.190	0.000	0.000
Married	0.007	Z	2.048	0.041	0.168	F	7.378	0.000	0.000
Health insurance public only	-0.008	Z	-1.946	0.052	0.168	F	20.076	0.000	0.000
Race AIAN	-0.018	Z	-1.853	0.064	0.185	F	3.674	0.000	0.000
Disabled	-0.013	Z	-1.632	0.103	0.243	F	12.870	0.000	0.000
Health insurance private only	0.004	Z	1.556	0.120	0.259	F	15.454	0.000	0.000
Urban population	0.001	Z	1.340	0.180	0.339	F	9.109	0.000	0.000
Unemployed	-0.017	Z	-1.333	0.183	0.339	F	21.400	0.000	0.000
Health insurance insured	0.007	Z	1.075	0.282	0.432	F	21.273	0.000	0.000
Race White	0.002	Z	0.692	0.489	0.707	F	21.027	0.000	0.000
Employed	0.002	Z	0.563	0.574	0.732	F	19.242	0.000	0.000
Labor force participation	0.002	Z	0.504	0.614	0.732	F	17.231	0.000	0.000
Race Hispanic	0.001	Z	0.393	0.694	0.732	F	71.635	0.000	0.000
Race Asian	0.004	Z	0.384	0.701	0.732	F	20.859	0.000	0.000
Race Black	-0.001	Z	-0.220	0.826	0.826	F	23.627	0.000	0.000

Table 10. Feeling Heard and Understood Measure Score by Survey Mode

Survey mode	N*	Mean	Standard Deviation
Mail	1268	0.73	0.37
Phone	979	0.67	0.37
Web	525	0.75	0.37

Survey mode	N*	Mean	Standard Deviation

* For purposes of this analysis, we included only those patients who responded to at least three out of the four data elements comprising the *Feeling Heard and Understood* measure.

A single data element ("I felt this provider and team understood what is important to me out of life") out of the four *Feeling Heard and Understood* data elements was significantly associated with diagnosis group (p<0.01), and the raw measure score (Table 11) was also significantly associated with diagnosis group. These results held after multiple comparison adjustment. Due to challenges with data quality, we were unable to conduct further analyses within the scope of this effort, but these findings provide preliminary indication that diagnosis may impact responses to the measure data elements and overall measure performance, underscoring the importance of further research in this area.

Table 11. Mean Score by Diagnosis Grouping

Diagnosis grouping	Mean	N
Cancer (solid and liquid)	.72	1670
End-organ disease (non-neurological)	.648	223
Other	.735	597
ANOVA p-value: p < 0.01		

Cell left intentionally blank.

[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]
The selection of the variables to be collected for consideration in the statistical risk model was be informed by statistical results presented in section 2b.24 (Table 9) on the potential risk factors. Based on these results and input from our project advisory group and TECUPP, we determined it was not appropriate to adjust this measure for social risk factors.
[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]

We used Kendall’s tau to assess the unadjusted and adjusted scores and explore how rankings among providers change after risk-adjustment. We also used statistical tests to assess the significance of covariates in the risk adjusted model and discussed results on what variables to include with our technical expert panel.

Tests using Kendall’s tau look at comparing the rank order of unadjusted scores to the order of risk-adjusted scores and assessing the proportion of cases where the order has changed. A statistic of 1 would imply that risk-adjustment has no effect on the rank order of programs and a statistic of -1 would imply that the order is completely reversed by risk adjustment (Parast et al., 2018). Values of 0.8 to 0.95 are typical of those reported in NQF documentation for the CAHPS surveys.

Description	Result
Kendall’s τ Test Statistic	0.88
Proportion Where Rank Order Changed = $(1-\tau)/2\ \%$	6.0%

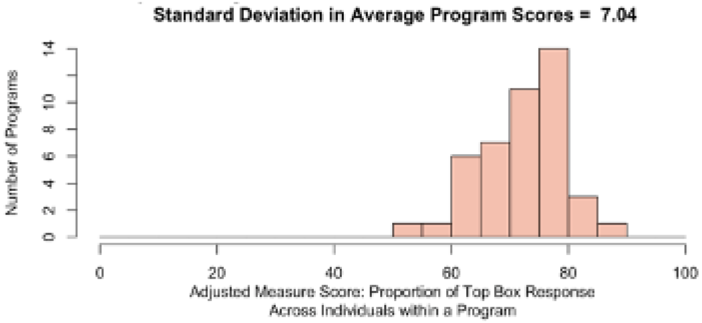
We discussed results from the risk adjustment and exclusion analyses with our project advisory group and arrived at a final model that included two risk adjusters: 1) survey mode, and 2) an indicator of proxy assistance. Table 12 below provides a summary of regression coefficients of the fixed effects of the adjustment model (i.e., ignoring the b_i terms). This indicates measure performance is adjusted slightly downward for phone-completed surveys and slightly upward for web-completed surveys and when a proxy assists the patient with survey completion.

Table 12. Summary of Regression Coefficients of Adjustment Model Fixed Effects

Parameter	Estimate	Error	Lower 95%	Upper 95%
Intercept – β_0	0.97	0.08	0.82	1.13
Survey Mode - Phone	-0.25	0.05	-0.35	-0.15
Survey Mode - Web	0.23	0.07	0.10	0.36
Proxy Status – Assisted	0.23	0.09	0.09	0.40

The model above can be used to assess the distribution for the estimated (i.e., adjusted) program scores shown in Figure 6. This results in reduced variability (SD=7.04) in program performance on adjusted average scores. However, the measure still demonstrates variability across programs.

Figure 6. Adjusted Program Performance



Citations:

Parast, L., Haas, A., Tolpadi, A., Elliott, M. N., Teno, J., Zaslavsky, A. M., & Price, R. A. (2018). Effects of Caregiver and Decedent Characteristics on CAHPS Hospice Survey Scores. *Journal of pain and symptom management*, *56*(4), 519-529.e511.

[Response Ends]

2b.27. Provide risk model discrimination statistics.

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

[Response Begins]
N/A
[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]
N/A
[Response Ends]

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

[Response Begins]
N/A
[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]
We incorporate risk adjustment variables into our models to provide fair comparisons among programs and to provide a best effort to ensure that the observed differences from programs are truly from differences in performance and not due to baseline differences in risk variables (including survey mode) that represent the programs. We also retained proxy response as a case-mix adjustment variable, consistent with existing CAHPS surveys (e.g., Medicare CAHPS, Prescription Drug Plan CAHPS), to alleviate ongoing concerns about the potential for impact.

Based on input from our project advisory group and TECUPP, we determined it was not appropriate to adjust this measure for social risk factors (e.g., race/ethnicity, urbanicity, median household income, gender, marital status, public insurance use, unemployment, and families below poverty line). After adjustment for multiple comparisons, none of these variables were significant in their relationship with the measure.

We did not have clinical data to evaluate risk adjustment for disease type or severity/acuity and note this as an important area for future research. Exploratory descriptive analyses based on a rough grouping of diagnostic categories showed the *Feeling Heard and Understood* raw measure score was significantly associated with diagnosis group, i.e., patients with end-organ disease such as heart failure and kidney disease had slightly lower scores than patients with any cancer and those in the “other” category. Due to challenges with data quality, we were unable to conduct further analyses within the scope of this effort, but these findings provide preliminary indication that diagnosis may impact responses to the measure data elements and overall measure performance, underscoring the importance of further research in this area. However, we hypothesize that any differences in measure performance based on disease type (e.g., cancer versus heart failure) may be a proxy for other variables such as where a patient was receiving care. We also hypothesize that any differences in the measure based on disease severity/acuity are likely due to differences in care processes that should and could be targeted for quality improvement and therefore from a conceptual standpoint, would not be a good candidate for inclusion in risk adjustment models.

[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]
N/A
[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]
Other (Please describe)
Patient-reported data is collected via survey instrument. The instrument was developed for this measure and can be completed via web survey, on paper or over telephone in English. Patient eligibility is determined based on coded visit information in the electronic health record.
[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]
Patient/family reported information (may be electronic or paper)
[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]
Rationale for using other than electronic sources: This is a patient-reported measure of experience that is not currently collected in structured electronic fields or EMR-based clinical data fields. Future work should explore options for embedding these measures in the EMR. These data should optimally be collected post-visit, away from the point of care, via survey. Findings from the alpha pilot test and beta field test indicate the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. Interviews with programs who participated in the alpha pilot test and beta field test also support the perceived feasibility of the measure in clinical practice across providers and administrators.
[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]
This is not an eMeasure.
[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]
Information for the measure calculation is collected via a survey data collection instrument, which will be provided to CMS, to be made available to CMS-approved survey vendors and palliative care programs. Palliative care programs will contract with a survey vendor to administer the survey to eligible patients. To minimize bias and reduce workload burden on programs, the survey vendor will be responsible for identifying eligible cases using electronic/automated queries, fielding the survey in the appropriate timeframes, receiving, cleaning, and summarizing survey data for program-level

quality improvement (if requested by the program), and submitting a final program-level data set to CMS for measure scoring. This last step may include the submission of both program-level data as well as unadjusted program scores to CMS, for risk-adjustment once data are aggregated across programs.

Findings from the alpha pilot test and beta field test support the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. Data collection for the beta field test occurred between November 2019 and February 2021, with a pause between March and September 2020 due to the COVID-19 pandemic. The survey response rate (37% raw response rate; 46% response rate excluding ineligible patients) achieved during the beta field test supports ease of use for patients responding to the survey. Interviews with programs that participated in the alpha pilot test and beta field test support the perceived feasibility of the measure in clinical practice across providers and administrators.

The majority of respondents to the 2021 public comment period supported feasibility of the proposed measures. When asked “How feasible would it be to implement these measures (e.g., contracting with a survey vendor, identifying eligible patients through administrative or medical record data, submitting scores to CMS, etc.)?” 21.8% of respondents said “very feasible” and 42.7% said “somewhat feasible.” The majority of comments indicated support for feasibility of the proposed measure, although some commenters raised concerns about the cost of hiring a survey vendor and implementation burden (e.g., staffing and support limitations).
[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]
No fees or licensing requirements will be necessary for users to implement the proposed measure. However, implementation costs include the cost of hiring an authorized survey vendor to field surveys and process data.
[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]
Not in use
Newly developed measure
[Response Ends]

4a.02. Check all planned uses.

[Response Begins]
Payment Program
[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]
This measure is newly developed and not currently in use. The Centers for Medicare & Medicaid Services (CMS) entered a cooperative agreement with the American Academy of Hospice and Palliative Medicine (AAHPM) as part of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) to develop two patient-reported measures of palliative care experience, broadly in the domains of symptoms and communication. The measures are intended to assess the extent to which patients receiving ambulatory palliative care received the help that they wanted for their pain, and that they were heard and understood by their palliative care provider and team. AAHPM partnered with the National Coalition for Hospice and Palliative Care and RAND Health Care to develop the proposed measures for use in CMS's Quality Payment Program.

[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]
The goal of this project is to produce quality measures that can be used by providers eligible for CMS' Merit-Based Incentive Payment System (MIPS) who provide palliative care services to their patients, so that the patient experience of core components of high-quality palliative care can be attributed to their providers and used to incentivize quality improvement. Medicare providers now choose one of two payment tracks – alternative payment models (APMs) and MIPS – which offer different combinations of incentives and requirements to encourage high-quality, low-cost care. Although MIPS applies to all Medicare patients, with no limit or focus on patients with serious illness, a strong portfolio of MIPS quality measures helps ensure measurement is meaningful and relevant to providers and their patients. The two palliative care measures were submitted to the 2021 MUC list for inclusion into CMS' Quality Payment Programs, including MIPS and APMs.
[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]
Performance data from our test has been provided to all participating programs (i.e., those who were being measured) as well as key stakeholder groups, e.g., the technical expert clinical user and patient panel (TECUPP) and project advisory panel.

For the testing sample, we identified approximately 360 palliative care programs in the United States that reported providing ambulatory palliative care to the Mapping Community Palliative Care Project and The National Palliative Care Registry. An additional 35 programs were added to the recruitment list because they submitted a project interest form via email, or attended an informational webinar hosted by the National Coalition for Hospice and Palliative Care (NCHPC) in June 2019. We sought national representation by oversampling larger programs (i.e., those with more patients) and stratifying recruitment efforts by administrative home type (i.e., hospice, hospital, ambulatory, and other administration) and by geographic location to ensure representation across U.S. Census Regions. A detailed description of the program sample is included in section 2a.05.

Five ambulatory palliative care programs participated in the alpha pilot test. Programs were included in the alpha pilot test based on their ability to sign-up and complete the required participation agreement processes in time for the start of the test. We included only five programs because our goal was primarily to evaluate and refine our fielding procedures in advance of the beta field test, and to identify any critically important changes to the data collection parameters and processes. At the end of the alpha pilot test, each program was provided a summary of their own data, including frequency and percent of responses to each survey item, including the items proposed to comprise the *Feeling Heard and Understood* measure. We connected each program with a contact from the project for assistance with interpretation. Because this was a pilot test, the conclusions that can be drawn from the data were limited but did offer each program a preliminary assessment of their patients' experience with the program.

A total of 44 programs ever participated in the beta field test, defined as providing at least one sample file during the test (see description of programs in section 2a.05). All programs that participated in the beta field test will also receive a summary report describing their performance on each survey item as well as their performance on the *Feeling Heard and*

Understood measure. Based on feedback from alpha pilot test programs, the summary reports were refined to better suit the needs of programs that participated in the beta field test.
[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]
Reports of performance data from testing were provided to all participating alpha palliative care programs (n=5) and all beta programs (n=44 programs). For the alpha programs, data from patient experience surveys was aggregated by program and sent only to that program via emailed report. Data were sent once, upon the completion of the test. Each report provided a summary of frequency and percent of responses to each survey item along with brief narrative descriptions of the results.

All programs that participated in the beta field test have also received a summary report describing their performance on each survey item as well as their performance on the *Feeling Heard and Understood* measure. For the beta field test, we provided comparative data so each program could understand their performance against the performance of comparable programs. Data were sent once, upon the completion of the test.
[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]
AAHPM conducted interviews with 25 palliative care programs that participated in the alpha pilot test and beta field test to better understand potential implementation challenges, resource requirements for measure implementation, how the proposed measure may be used to facilitate quality improvement, and the perceived financial and administrative burden of measure implementation and associated quality improvement activities. Because of the different topics discussed, interviews included a palliative care provider and/or a data specialist or program manager. We encouraged the point of contact for each palliative care program (typically a Program Manager or Medical Director) to invite a data specialist or provider to participate with them in the interview depending on their role.

AAHPM also obtained feedback on potential implementation challenges and usefulness of the proposed measure for quality improvement during the 2021 public comment period. Respondents included patients, family caregivers, and advocates living with serious illness; providers/clinicians caring for those living with serious illness; representatives from national organizations; and other professionals.
[Response Ends]

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

[Response Begins]
Factors Influencing Measure Implementation

In interviews with palliative care providers and administrators, AAHPM inquired about potential challenges related to implementation of the proposed measures. Some providers raised practical issues including “survey fatigue,” especially for patients who are seriously ill, as well as the need for question phrasing to be understandable or “resonate” for a broad range of patient populations, including those with low literacy levels. Providers expressed concern that allowing proxy responses might introduce bias, particularly if family member perceptions were not aligned with patient perceptions (e.g., thinking that pain was undertreated). Palliative care providers also expressed concerns about attribution given that patients see multiple providers. Providers also raised concerns about selection of survey modalities (i.e., email, mail, in-person) that will yield high response rates and thoughtful responses (i.e., after patients have had a chance to think about their experience). In addition, the financial and administrative burden for programs to implement the measure was an important consideration, often dependent on program size, organizational type, and existing resources.

Resources Required for Measure Implementation

In interviews with palliative care providers and administrators, including program managers and data specialists, AAHPM inquired about resources that will be necessary for successful measure implementation and use. Resources required to implement the measure would likely include IT staff hours to extract patient visit data from the electronic health record and the cost of hiring a survey vendor to administer the survey to eligible patients. Most programs had previously worked with a vendor to administer patient surveys. Important factors cited in the decision to invest in support from a survey vendor included cost, sensitivity, and tracking issues (i.e., concerns about sending surveys to deceased patients), patient survey fatigue, ability to compare measure performance with other programs, and unstable patient mailing addresses (although, in light of COVID-19, one program noted that they now consistently collect patient emails for telehealth). Finally, another concern for implementation was the cost of quality improvement associated with the measure. Anticipated quality improvement activities related to measure implementation included provider communication training; encouraging providers to establish expectations with patients and set realistic goals; and root cause analysis to identify the sources of patient dissatisfaction, including external factors (i.e., experience in clinic or delays, long wait times to get an appointment).
[Response Ends]

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

[Response Begins]
In addition to palliative care providers, we sought feedback from administrators, including program managers and data specialists, regarding measure implementation. Please see 4a.08 for details.

Public comment respondents also provided feedback on feasibility of measure implementation. When asked, “How feasible would it be to implement these measures (e.g., contracting with a survey vendor, identifying eligible patients through administrative or medical record data, submitting scores to CMS, etc.)?” 22% of respondents said “very feasible,” 43% said “somewhat feasible,” 8% said “not feasible,” and 27% said “I don’t know.” The majority of comments indicated support for feasibility of the proposed measures. Some respondents raised concerns about implementation burden related to staffing and support limitations, as well as the cost of hiring a survey vendor. See 4b.01 for details of feedback from public comment on usefulness of the proposed measure for performance improvement.
[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]
Findings from palliative care program interviews support the perceived feasibility of the measures in clinical practice across providers and administrators, including the feasibility of data collection via survey vendor. This is the proposed mechanism for measure implementation. The survey response rate (37% raw response rate; 46% response rate excluding ineligible patients) achieved during the beta field test also supports ease of use for patients responding to the survey. To minimize patient burden, the final patient experience survey was reduced to 11 items (see Appendix for survey instrument). To prevent survey fatigue, surveys will be fielded to eligible patients no more often than once per year. Patients who have already completed the patient experience survey in a given 12-month reporting period will be excluded from measurement to avoid response bias due to priming effects and minimize patient burden. To address concerns about possible challenges with attribution given that patients see multiple providers, we referenced a specific provider and team in the patient survey. To address concerns about sending surveys to deceased patients, our recommended data collection approach is to first send eligible patients a letter notifying them of the upcoming survey with a stamped postcard that can be returned in the event of death or a move/new address.

To address potential concerns around bias from proxy responses, measure testing data were used to establish exclusion criteria related to proxy response. Respondents were categorized into three distinct groups based on proxy assistance as follows: respondent only (no proxy assistance at all), proxy assisted (proxy helped patient complete the survey but patient supplied answers e.g., proxy read questions and wrote down answers), proxy only (proxy answered all questions and patient was not involved). We compared descriptive statistics for the measure components for each of these three groups

to inform the impact of proxy assistance and to determine whether to include/exclude proxy responses. A one-way ANOVA test for differences among these three means was significant (F(2, 3199)=3.80, p=.023), and follow-up pairwise mean comparisons revealed no difference between patient only and proxy only (t(581)=1.22, p=0.22). However, proxy assisted was significantly different from both patient only (t(271)=-2.48, p=0.01) and proxy only (t(487)=-2.86, p=0.004). Despite the lack of a significant difference in *Feeling Heard and Understood* measure score means between the proxy only and patient only groups, after discussing these results with our advisory board, we decided to exclude surveys that were completed solely by a proxy with no patient involvement for conceptual reasons. As a patient-reported measure of palliative care experience, we wanted to ensure that at least some direct patient report was reflected in the measure response, a rationale for excluding proxy-only responses that was endorsed by the advisory board. Further the absence of a significant difference in responses by proxy-involvement suggests minimal to no impact of this decision on measure outcomes. We elected to include proxy-assisted surveys and to add an adjustment for proxy assistance to account for small differences in measure components due to the proxy involvement. This allowed us to retain as much patient-reported data as possible, while acknowledging that patients in this population will likely need some assistance with survey completion.
[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]
The measure is newly developed and not currently in use.

Respondents to the 2021 public comment period were largely supportive of the proposed *Feeling Heard and Understood* measure and its usefulness for performance improvement in palliative care settings. We received 236 responses to public comment, including responses from patients, family caregivers, and advocates living with serious illness (38%); providers/clinicians caring for those living with serious illness (42%); representatives from national organizations (5%); and other professionals (15%). Overall, 78.9% of respondents agreed that the *Feeling Heard and Understood* measure captures important information. When asked "How likely is it that ambulatory (e.g., clinic-based) palliative care providers or programs would choose to report on these measures?" 37.6% of respondents said "very likely" and 40.8% said "somewhat likely." When asked "How likely are you to use the *Feeling Heard and Understood* measure to improve your practice and/or the care you provide?" 60.0% of clinician respondents said "very likely" and 22.4% said "somewhat likely," demonstrating support from palliative care clinicians for usefulness of the proposed measure for performance improvement.
[Response Ends]

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

[Response Begins]
To date, we have not encountered any unintended adverse consequences from measuring the extent to which patients feel heard and understood by providers. In qualitative interviews with palliative care programs that participated in the alpha pilot test and beta field test, providers were asked about potential unintended consequences of the *Feeling Heard and Understood* measure. Providers noted that comparison across palliative care programs may be challenging if patient populations have differences in disease trajectories that impact communication. Another potential concern reported by providers was repercussions of negative feedback. There were concerns that some patients may have unrealistic expectations for palliative care, and patients whose expectations are not met may identify as not being heard and understood. Palliative care providers often have to deliver bad news to patients, which may negatively impact patient perceptions of the palliative care team. Providers recommended strategies to prevent some of these potential unintended consequences, including encouraging providers to establish expectations with patients up front and set realistic goals for palliative care. Providers also recommended framing the questions to help patients understand that the measure is useful for the program and ultimately for other patients.

In addition, it is possible that patients who have died may be contacted to complete the survey, potentially causing distress for families. Our recommended data collection approach is to first send eligible patients a letter notifying them of the upcoming survey with a stamped postcard that can be returned in the event of death or a move/new address.
[Response Ends]

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

[Response Begins]
In qualitative interviews with palliative care programs that participated in the alpha pilot test and beta field test, providers were asked about perceived benefits of the *Feeling Heard and Understood* quality measure. Overall, providers were very positive about the *Feeling Heard and Understood* quality measure, noting its central importance to palliative care and the need to measure it. Providers described the value of this quality measure to capture patient experience, stating that this measure encapsulates "the philosophy and soul of palliative care," the "essence," or "mission" of palliative care. Providers also talked about the usefulness and appropriateness of this type of quality measure for palliative care because it does not focus on the patient's medical status or other quality metrics that are not expected in palliative care. They noted that the measure would inform quality improvement efforts to better understand potential gaps in their program and aspects of their care that may impact patients' experiences of feeling heard and understood.
[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).

(Can search and select measures.)
[Response Begins]
2651: CAHPS® Hospice Survey (experience with care)
[Response Ends]

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

(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]
N/A
[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]
Yes
[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]
N/A
[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]
N/A - We did not identify any competing measures.
[Response Ends]

Appendix

Supplemental materials may be provided in an appendix.: Available in attached file
Attachment: Patient Experience Survey - Feeling Heard and Understood.pdf

Contact Information

Measure Steward (Intellectual Property Owner) : American Academy of Hospice and Palliative Medicine
Measure Steward Point of Contact: Ast, Katherine, kast@aahpm.org
Measure Developer if different from Measure Steward: American Academy of Hospice and Palliative Medicine
Measure Developer Point(s) of Contact: Ast, Katherine, kast@aahpm.org

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 in attached file
[Response Ends]

Attachment: Patient Experience Survey - Feeling Heard and Understood.pdf

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

Describe the members' role in measure development.

[Response Begins]
The Advisory Panel provided input on measure specification and value and field-testing decisions.

Robert Gramling, MD, DSc, University of Vermont Medical Center

Laura Hanson, MD, MPH, University of North Carolina School of Medicine

Amy Kelley, MD, MSHS, Mount Sinai Health System

Karl Lorenz, MD, MSHS, Palo Alto Veterans Affairs Medical Center

Joanna Paladino, MD, Dana-Farber Cancer Institute and Ariadne Labs

VJ Periyakoil, MD, Stanford University School of Medicine

Christine Ritchie, MD, MSPH, University of California San Francisco School of Medicine

Richard Street, PhD, Texas A&M University

Joan Teno, MD, MS, Oregon Health and Science University School of Medicine

Neil Wenger, MD, MPH, University of California Los Angeles School of Medicine

The Technical Expert Clinical User Patient Panel (TECUPP) provided expertise and feedback on the proposed quality measures for patients with serious illness throughout the measure development lifecycle. The Measure Specifications Panel (MSP), a small subgroup of experts with highly technical measure development and specification expertise, were selected to evaluate the proposed measures for initial feasibility and review later testing results to guide decision-making regarding the measures.

**Indicates Measure Specification Panelists (MSP)*

Sydney M. Dy, MD (Co-Chair)*, Johns Hopkins Bloomberg School of Public Health

Mary T. Ersek, PhD, RN, FPCN (Co-Chair)*, University of Pennsylvania School of Nursing, Philadelphia Veterans Affairs Medical Center

Steven M. Asch, MD, MPH*, VA Palo Alto Healthcare System

Kathleen Bickel, MD, MPhil, MS*, University of Colorado, Denver School of Medicine, Veterans Affairs Eastern Colorado Healthcare System

Lori Bishop, MHA, BSN, RN, CHPN*, National Hospice and Palliative Care Organization

Brenda Blunt, DHA, MSN, RN, CVP Corp

Amy Ciancarelli, BS, CPXP, Care Dimensions

Amy L. Davis, DO, MS, FACP, FAAHPM, Drexel University School of Medicine, Main Line Health System

Sa'Brina Davis, AmerisourceBergen

Torrie Fields, MPH*, Blue Shield of California

Elizabeth L. Fricklas, PA-C, Duke Palliative Care

Joy Goebel, RN, PhD, FPCN, California State University Long Beach School of Nursing

Matthew J. Gonzales, MD, FAAHPM, Institute for Human Caring

Anna Gosline, SM, Blue Cross Blue Shield of Massachusetts

Marian Grant, DNP, CRNP, ACHPN, FPCN, RN, Marian Grant Consulting

Rev. George F.Handzo, MA, Mdiv, BCC, CSSBB, HealthCare Chaplaincy Network

Denise Hess, MDiv, BCC-PCHAC, LMFT, Supportive Care Coalition

Sarah E. Hetue Hill, PhD*, Ascension Medical Group

Faye Hollowell, National Patient Advocate Foundation

Arif Kamal MD, MBA, MHS, FACP, FAAHPM, FASCO*, Duke University School of Medicine

Rebecca Kirch, JD, National Patient Advocacy Foundation

Cari Levy, MD, PhD, CMD, University of Colorado, Denver School of Medicine, Veterans Affairs Eastern Colorado Health Care System

Phillip M. Rodgers, MD, FAAHPM*, University of Michigan Medical School

Benjamin D. Schalet, PhD*, Northwestern University

Tracy A. Schroepfer, PhD, MSW, MA, University of Wisconsin-Madison School of Social Work

Cardinale B. Smith, MD, PhD*, Mount Sinai Health System

Hannah Luetke-Stahlman, MPA, Cerner Corporation

Paul E. Tatum, III, MD, MSPH, CMD, FAAHPM, AGSF, Dell Seaton Medical Center at the University of Texas, Austin

Martha L. Twaddle, MD, FACP, FAAHPM, HMDC, Northwestern Medicine, Lake Forest Hospital, Northwestern University

Kathyrn A. Walker, PharmD, BCPS, CPE, MedStar North, University of Maryland School of Pharmacy
[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]
N/A
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

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

[Response Begins]
Unknown
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

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

[Response Begins]
N/A
[Response Ends]

Attachment: Patient Experience Survey - Feeling Heard and Understood.pdf