Research Report

MACRA Palliative Care Quality Measure Development—Testing Summary Report

Measure Name: Feeling Heard and Understood

Sangeeta C. Ahluwalia, Brian G. Vegetabile, Maria Orlando Edelen, Claude Messan Setodji, Anthony Rodriguez, Adam Scherling, Jessica Phillips, Carrie M. Farmer, Jordan M. Harrison, Julia Bandini, Wenjing Huang, Lucy Schulson, Anne Walling, Sarah Dalton, Monique Martineau, Owen Hall, Danielle Schlang, Melissa A. Bradley, Katherine Ast
About This Report

The Centers for Medicare & Medicaid Services (CMS) entered into a cooperative agreement (1V1CMS331639-01-00) 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 in the domains of symptoms and communication. The performance measures are intended to assess (1) the extent to which patients who used ambulatory palliative care received the help they wanted for their pain and (2) the extent to which they felt heard and understood by their palliative care provider and team. Under this agreement, AAHPM has partnered with the RAND Corporation and the National Coalition for Hospice and Palliative Care to develop and test the proposed measures and engage patients and caregivers throughout the performance measure development process.

This research was funded by CMS and carried out within the Quality Measurement and Improvement Program in RAND Health Care.

The testing of the proposed measures was led by RAND Health Care, a division of the RAND Corporation. RAND Health Care promotes healthier societies by improving health care systems in the United States and other countries. We do this by providing health care decisionmakers, practitioners, and consumers with actionable, rigorous, objective evidence to support their most complex decisions. For more information, see www.rand.org/health-care, or contact

**RAND Health Care Communications**

1776 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
(310) 393-0411, ext. 7775
RAND_Health-Care@rand.org
## Contents

About This Report .................................................................................................................. iii  
Figures ............................................................................................................................... v  
Tables ................................................................................................................................. vi  
Summary ............................................................................................................................. vii  
Acknowledgments ............................................................................................................... xiv  
Abbreviations ..................................................................................................................... xv  
1. Background and Overview .............................................................................................. 1  
   Importance ....................................................................................................................... 1  
   Performance Measure Impact ......................................................................................... 2  
   Proposed Performance Measures ................................................................................ 2  
   Overview of Testing Process ......................................................................................... 5  
   Organization of the Report ......................................................................................... 7  
2. Patient Population and Sampling .................................................................................. 8  
   Target Measure Population and Data Sources ............................................................. 8  
   Sample Characteristics ................................................................................................. 13  
3. Critical Data Element Testing ....................................................................................... 18  
   Methods ......................................................................................................................... 18  
   Results .......................................................................................................................... 20  
4. Performance Measure Testing ....................................................................................... 23  
   Methods ......................................................................................................................... 23  
   Results .......................................................................................................................... 30  
   Threats to Validity Caused by the COVID-19 Pandemic .............................................. 42  
5. Feasibility and Implementation Considerations ............................................................. 45  
   Overview ......................................................................................................................... 45  
   Cost and Burden Assessment ...................................................................................... 45  
   Program Perspectives .................................................................................................... 47  
   Experiences of Patients from Racial and Ethnic Minorities ....................................... 50  
6. Performance Measure Implementation and Future Considerations ............................ 53  
   Considerations for Program Planning ......................................................................... 53  
   Usability of the Performance Measure ....................................................................... 56  
   Conclusion .................................................................................................................... 57  
Appendix A. National Beta Field Test Survey Instrument ................................................. 58  
Appendix B. Methods and Analytic Procedures ................................................................ 67  
References .......................................................................................................................... 70
Figures

Figure 1.1. Measures Selected for Development ................................................................. 4
Figure 1.2. Testing Phases in Measure Development ............................................................ 5
Figure 2.1. Timeline of Data Collection Procedures ............................................................ 12
Figure 2.2. National Beta Test Timeline .............................................................................. 13
Figure 2.3. Distribution of Completed Surveys Across Programs ....................................... 17
Figure 2.4. Distribution of Program-Level Nonresponse ...................................................... 17
Figure 4.1. Distribution of the Proportion of Top-Box Responses Across All Individuals ...... 30
Figure 4.2. Distribution of the Average Percentage of Top-Box Responses Within Programs .. 31
Figure 4.3. Adjusted Program Performance ......................................................................... 36
Figure 4.4. Posterior Distribution of Adjusted Program Scores .......................................... 37
Figure 4.5. Estimated Posterior Distribution of the Adjusted Intraclass Correlation

   Coefficient ...................................................................................................................... 39
Figure 4.6. Estimate Posterior Distribution of Reliability at Fixed Sample Sizes ................. 39
Figure 4.7. Distribution of Program-Specific Reliability ....................................................... 40
Figure 4.8. Unadjusted Versus Adjusted Program Scores .................................................... 41
Figure 4.9. Recalculated Program Performance ................................................................ 41
Figure 4.10. Recalculated Posterior Distribution of Reliability at Fixed Sample Sizes .......... 42
Figure 5.1. Example Data Collection and Reporting Schedule for Measure Performance

   Year 2022 ...................................................................................................................... 45
Figure 6.1. Survey Field Period ......................................................................................... 55
Tables

Table 2.1. National Beta Field Test Recruitment Targets and Final Recruitment Data ........................................... 9
Table 2.2. Final Disposition of Fielded Surveys ($n = 7,595$) ............................................................................. 14
Table 2.3. Survey Nonresponse ($n = 3,356$) .................................................................................................................. 14
Table 2.4. Excluded Surveys Due to Ineligibility ($n = 1,435$) ............................................................................. 15
Table 2.5. Patient Characteristics ............................................................................................................................... 16
Table 3.1. Data Elements Evaluated for Inclusion in the National Beta Field Test .................................................. 19
Table 3.2. Distribution of Responses to the *Feeling Heard and Understood* Performance Measure Data Elements (in percent) .................................................................................................................. 21
Table 3.3. Frequency of Number of Questions Missing Responses in the *Feeling Heard and Understood* Performance Measure Data Element Set .................................................................................................................. 22
Table 4.1. Diagnosis Groupings ............................................................................................................................... 26
Table 4.2. Association of Potential Risk Factors with *Feeling Heard and Understood* Performance Measure Score and Program (sorted by measure score p-values) .......................................................... 32
Table 4.3. *Feeling Heard and Understood* Performance Measure Score by Survey Mode ................................. 34
Table 4.4. Mean Score by Diagnosis Grouping ........................................................................................................ 34
Table 4.5. Mean *Feeling Heard and Understood* Performance Measure Score According to Proxy Group .............................................................. 35
Table 4.6. Summary of Regression Coefficients of Adjustment Model Fixed Effects ........................................... 36
Table 4.7. Change in Measure Score, Assuming 100 Ranked Programs ................................................................. 38
Table 4.8. Pre-Pandemic and Mid-Pandemic Measure Score Distributions ............................................................... 44
Table 4.9. Generalized Linear Model Results for the *Feeling Heard and Understood* Performance Measure ............................................................................................................................................................. 44
Table 5.1. Breakdown of Programs That Participated in the Program Interviews ($n = 25$) ................................. 48
Summary

Background

Palliative care has grown tremendously in the past 20 years. Once thought to be limited to delivery in hospital or hospice settings, palliative care services can be delivered to patients in ambulatory (office) clinics or as a supplement to ongoing treatment, such as for cancer. At its core, palliative care is patient-centered care, and effective palliative care focuses on patient values and care preferences, symptom management, and peace at the end of life.

Palliative care stakeholders, including patients and their caregivers, providers, and health care systems, have come to see the value in systematically measuring the quality of palliative care. Yet there are few measures endorsed by the National Quality Forum to assess provider performance, and the performance measures that are in use are limited to narrowly defined populations, such as hospice patients. Perhaps surprisingly, given the focus of palliative care on the patient, there are no patient-reported measures of palliative care quality to provide insight into patients’ experiences of care.

The Centers for Medicare & Medicaid Services (CMS) set out to address gaps in palliative care quality measurement, particularly for accountability programs (such as the Quality Payment Program), through a cooperative agreement (1V1CMS331639-01-00) with the American Academy of Hospice and Palliative Medicine (AAHPM) as part of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). The purpose of this agreement was to

- develop and test two patient-reported performance measures for patients receiving ambulatory palliative care
- ensure that the proposed measures incorporate patient voice and preferences and are broadly applicable to patients with serious illness and their families receiving palliative care services in a range of ambulatory primary and specialty care settings
- convene a technical expert panel that incorporates patient, caregiver, and family input directly into the performance measure development, specification, testing, and implementation processes
- submit the proposed palliative care measures to CMS’s 2021 Measures Under Consideration list for the Quality Payment Program and to the National Quality Forum for review and endorsement so that providers can measure and improve the quality of care in ambulatory palliative care.

AAHPM partnered with RAND Health Care and the National Coalition for Hospice and Palliative Care to develop two patient-reported performance measures of palliative care experience in the domains of symptoms and communication. The symptom measure is intended to assess the extent to which patients who used ambulatory palliative care received the help they wanted for their pain, and the communication measure is intended to measure the extent to which
they felt heard and understood by their palliative care provider and team. In this report, we describe the final measure specifications, methods, testing results, and considerations for implementation for the proposed *Feeling Heard and Understood* performance measure. A companion report covers the same information for the proposed *Receiving Desired Help for Pain* measure (Ahluwalia et al., 2021).

**Methods and Results**

The proposed *Feeling Heard and Understood* performance measure was tested rigorously in a multiphase process. We began by creating a patient experience survey instrument, then tested the wording and structure of data elements with patients who received palliative care in an ambulatory setting. We conducted an alpha test first to test and finalize the survey-based data elements and pilot the survey within five palliative care programs. Subsequently, we tested the performance measure at the national level in a beta field test, which occurred over ten rounds through 44 programs. The national beta test was paused from March to September 2020 because of interruptions in program operations during the coronavirus disease 2019 (COVID-19) public health emergency.

**National Beta Test Patient Population and Sampling**

Feedback from our technical expert clinical user and patient panel (TECUPP) and project advisory group reinforced the importance of including in the performance measure’s target population all patients ages 18 and older who had at least one ambulatory palliative care visit. Working with the Center to Advance Palliative Care and the National Coalition for Hospice and Palliative Care, we identified 395 ambulatory palliative care programs in the United States from which to develop a sampling frame.

We contacted 238 palliative care programs, 70 of which were deemed ineligible to participate in the test for such reasons as not providing ambulatory care or not having practitioners eligible for the Merit-based Incentive Payment System (MIPS). Of the remaining programs, 44 participated by sending at least one sample file during field testing, 39 participated before and after the pause in data collection, three did not continue participation in September 2020, and two programs joined at that time. Ten were hospice sites, 24 were hospitals, and ten were ambulatory or other administrative sites.

Among these programs, we aimed to field 6,000 to 7,500 surveys and receive 2,400 to 3,000 completed surveys from patients receiving ambulatory palliative care. Patient sampling was conducted each month from November 2019 to February 2021, with a pause from March 2020 to September 2020 because of the COVID-19 pandemic. Fielding was conducted in ten rounds of 56 days each, beginning with patient prenotification and ending after 56 days. Patients were first contacted by mail or email to notify them (or their caregivers, if they were too ill to participate) that the survey would arrive soon, then they were mailed or emailed the survey. Patients were
given three weeks to respond to the survey and up to eight attempts to respond to the survey through a computer-assisted telephone interview. Patients who did not respond to the survey within 56 days were considered nonrespondents.

Over the course of the national beta field test, the survey instrument changed slightly. The instrument contained 43 data elements at the start of the test. In March 2020, during the sixth round of fielding, five data elements were added regarding the COVID-19 pandemic. When we resumed fielding in September 2020, we revised the instrument to include a total of eight data elements regarding the COVID-19 pandemic.

We fielded the survey to 7,595 sampled patients across ten rounds of data collection. Of the surveys fielded, 3,356 were not returned, 1,435 were excluded from any analyses because of ineligibility for the larger study, and 2,804 were returned and included in analyses. This resulted in a 37 percent raw response rate and a 46 percent response rate when excluding ineligible patients. Responses were distributed evenly across programs. With respect to mode, the majority of returned surveys were completed by mail (as opposed to by phone or web). About 19 percent of surveys were excluded because of ineligibility, such as the patient being deceased or not being involved in completing the survey. Most respondents were female (56 percent), White (88 percent), non-Hispanic/Latino (95 percent), and highly educated. The mean age of respondents was 63 years. One-third of programs each contributed 31 to 50 completed surveys.

**Critical Data Element Testing**

The construct of feeling heard and understood previously was measured using a single question: “Over the past two days, how much have you felt heard and understood by the doctors, nurses, and hospital staff?” That question was developed for use in the inpatient palliative care setting (Gramling et al., 2016); to improve the reliability of the data element for ambulatory palliative care patients, we evaluated ten data elements for possible inclusion as a scale during national beta testing. These potential data elements addressed feelings of comfort, trust, and value alignment, as well as whether patients feel considered as a whole person in their care. Our TECUPP and project advisory group provided input on the final wording, response options, and inclusion of each data element in the survey instrument. We also evaluated the psychometric properties and the reliability (both internal consistency and test-retest) and convergent validity of the potential data elements, as follows:

- Survey participants tended to respond at the high end of the response scale.
- The mean data element–level missingness for the entire survey was low (0.8 percent).
- There were no discernable patterns among the four data elements that we ultimately selected for the scale. Correlation of data elements ranged from 0.54 to 0.69.
- Reliability of the four–data element scale was high, indicating that scores obtained are reliable and can, therefore, be used in the construction of the performance measure.
• High scores on *Feeling Heard and Understood* were associated with high Consumer Assessment of Healthcare Providers and Systems (CAHPS) communication scores, supporting convergent validity of the data elements.

**Performance Measure Testing**

We evaluated the *Feeling Heard and Understood* performance measure first by the proportion of item-level top-box scores (i.e., items for which the most favorable response option was selected) across respondents and programs; then by a set of potential risk adjusters using program information, program data, and U.S. Census data; then by three potential denominator exclusions, two approaches to scoring, statistically significant differences among programs, reliability at the program level, and validity at the program level. We found the following:

• The proportion of top-box scores across respondents had a mean of 0.71, showing that responses skewed toward the high end of the scale across the four data elements. Variability across programs, however, was high.

• With one exception (survey mode), all the potential risk adjustment variables were found to have significant or nearly significant associations with programs, which means they could confound the relationship between quality of care and measure score.

• Analysis of variance by respondent type (i.e., patient-only, proxy-assisted, and proxy-only) found no significant difference between proxy-only and patient-only groups, but there was a significant difference between these two and the proxy-assisted type. In keeping with the spirit of the performance measure, we decided to exclude proxy-only responses but retained proxy-assisted responses.

• Using survey mode and indicator of proxy assistance as risk adjusters, we found that measure performance adjusted slightly downward for phone-completed surveys and slightly upward for web-completed surveys and for proxy assistance.

• The distribution of scores showed differences from the mean across programs.

• Bayesian generalized mixed-effects models showed a reasonable level of reliability based on between-program variability and within-program variability, given our sample sizes. Recalculations according to program size showed some sensitivity to small programs.

• The *Feeling Heard and Understood* performance measure was significantly and positively associated with the CAHPS communication measure, supporting the convergent validity of the *Feeling Heard and Understood* performance measure.

**Feasibility Assessment**

To assess the feasibility of implementing *Feeling Heard and Understood* in practice, we interviewed stakeholders to understand (1) costs and burden to programs, (2) programs’ perspectives on implementation, and (3) potential variability in patients’ experiences of care by racial and ethnic minority groups. Our findings are as follows:

• Using a survey vendor to administer the survey would minimize burden to programs. Most programs have prior experience using such vendors.

• Quotes from five CMS-approved survey vendors to administer the mixed-mode patient surveys varied from approximately $2,500 to $12,500 per year.
• Program interviewees reported generally positive perceptions of the *Feeling Heard and Understood* performance measure, with some concerns, including survey fatigue and understandability of the survey for those with low literacy levels.

• For quality assessment and improvement efforts, program interviewees discussed examining external factors, such as scheduling and appointment delays, to further assessment.

• Most programs had little to no experience with providing services over telehealth before the pandemic. Benefits of telehealth included enhanced personal connections (including with family members) and medication reconciliation; challenges included loss of human touch with patients.

• Patients and family members who identify as Black or African American or as Asian who were interviewed for their perceptions of palliative care indicated (1) a general lack of awareness of palliative care, (2) generally feeling heard and understood by their palliative care providers, and (3) higher satisfaction with palliative care providers than providers in the general health care system.

**Implementation Considerations and Conclusions**

As part of our examination of the testing and analytic procedures used during the national beta field test, we identified several considerations for programs planning to implement the *Feeling Heard and Understood* performance measure into clinical practice, quality improvement efforts, and quality payment programs.

First, it is important to recognize that certain populations and types of visits had to be excluded from the national beta field test, but these populations and visits will be important for programs to consider as they start to use this performance measure. Pediatric populations, for instance, were not included because these patients and their families require distinct considerations for ensuring their voice is reflected in the performance measure. Telehealth visits were not included because of their low incidence before the pandemic, although the public health emergency shifted many in-person visits to a telehealth platform. Because it is likely that telehealth visits will continue in greater numbers after the pandemic ends, this performance measure should incorporate telehealth in future testing. Indeed, programs with smaller numbers of visits tended to exhibit lower reliability, which should be considered when establishing requirements for program participation.

In terms of data collection, it is important to note that scores were lower for surveys completed by telephone. Because patients (or caregivers) were only contacted as a last resort if the survey had not been completed by mail or phone, we think this effect may be due to phone respondents being disinclined to participate because of negative experiences with palliative care. Thus, future tests could explore ways to optimize data collection. The survey instrument should be used and tested in other languages as well, particularly Spanish. Although we developed a Spanish-language translation of the survey instrument used in testing, we were unable to test the reliability and validity of the performance measure using Spanish-language data collection.
While testing this performance measure, we engaged clinicians, health systems, and patients and caregivers, and we offered guidance on how to use this performance measure in practice to drive systematic quality measurement and improvement in priority areas for palliative care. Stakeholder feedback suggests that this performance measure can be successfully implemented and can provide valuable information to guide palliative care improvement in patient experience.

We also feel that future research could focus on three areas:

- a systematic examination of survey administration modes, which may affect response rates and the representativeness of respondents
- an evaluation of the implementation of this performance measure into clinical practice and regular use
- an exploration and analysis of disparities in ambulatory palliative care access, utilization, and experience, including (but not limited to) racial, gender, ethnic, cultural, and language disparities.

**Feeling Heard and Understood: Final Measure Specification**

**Performance Measure Description**

*Feeling Heard and Understood* is a multi-data element measure consisting of four data elements:

- 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.”

**Performance Measure Type**

This is a patient-reported outcome performance measure.

**Denominator**

All patients ages 18 years and older who had an ambulatory palliative care visit.

**Denominator Criteria**

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

**AND**

Ambulatory palliative care visit, 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.
An eligible provider type: physicians (including doctors of medicine, osteopathy, dental surgery, dental medicine, pediatric 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, and registered dietitians or nutrition professionals.¹

Denominator Exclusions

- Patients who do not complete at least one of the four data elements in the multi-data element measure.
- Patients who do not complete and return 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 past six months (i.e., disavow the provider or program).
- Patients who were deceased when the survey reached them.
- Patients for whom a proxy completed the entire survey on their behalf for any reason (i.e., with no patient involvement), including patients who were deceased at the time the survey reached them.

Numerator

The *Feeling Heard and Understood* performance measure is calculated using top-box scoring. The *top-box score* refers to the percentage of patient respondents who give the most positive response. For each of the four data elements in this performance measure, individuals can respond using discrete values ranging from 0 (“Not at all true”) to 4 (“Completely true”). For all four data elements, 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.

Missing data in the outcome are naturally accommodated among the four data elements by the modeling procedure for adjusted score estimation; no outcome imputation is necessary and should not be performed.

¹ This list is based on 2019 MIPS eligible clinician types.
Acknowledgments

This project is the result of a multiyear collaborative partnership with the American Academy of Hospice and Palliative Medicine (the prime recipient of this cooperative agreement), the National Coalition for Hospice and Palliative Care, the National Patient Advocate Foundation, and the Center to Advance Palliative Care. The work described in this report would not have been successful without the thoughtful input provided by the AAHPM-MACRA Technical Expert, Clinical User, and Patient Panel and the AAHPM-MACRA Project Advisory Group over the three-year project period, from measure conceptualization to final measure specifications. We are immensely grateful for their engagement in this project.

We are especially appreciative of the 44 ambulatory palliative care programs that graciously gave of their time and engaged in a complex data collection effort over a 15-month period. Even during the COVID-19 pandemic, program leaders, staff, and health care workers contributed to our performance measure testing effort in multiple ways, which allowed us to successfully complete this project on time.

Most importantly, we want to acknowledge and express our sincerest gratitude to the almost 9,000 patients and caregivers who shared their experiences of palliative care and significantly informed the performance measure development effort through their survey data and one-on-one interviews. They willingly shared their personal experiences, thoughts, feelings, and often emotionally difficult narratives about their health and health care as people living with advanced illness. Without them, we would not have the first two performance measures that are specifically designed for ambulatory palliative care and available for use in quality improvement and accountability programs.

Several research staff were integral to the overall success of this project. Monica Rico supported data collection, cleaning, and analysis across multiple tasks. Andrew Beck, Allison Kierkegaard, Aaron Lang, and Justin Lee also made important contributions to data collection and project administration. Finally, we would like to thank Paul Cleary of Yale University and Layla Parast of RAND for their thoughtful substantive feedback on this report, as well as on several other project deliverables throughout the three-year project period. Their feedback has improved our work immensely.
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAHPM</td>
<td>American Academy of Hospice and Palliative Medicine</td>
</tr>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>CAHPS</td>
<td>Consumer Assessment of Healthcare Providers and Systems</td>
</tr>
<tr>
<td>CATI</td>
<td>computer-assisted telephone interview</td>
</tr>
<tr>
<td>CMS</td>
<td>Centers for Medicare &amp; Medicaid Services</td>
</tr>
<tr>
<td>COVID-19</td>
<td>coronavirus disease 2019</td>
</tr>
<tr>
<td>ICC</td>
<td>intraclass correlation coefficient</td>
</tr>
<tr>
<td>MACRA</td>
<td>Medicare Access and CHIP Reauthorization Act of 2015</td>
</tr>
<tr>
<td>MIPS</td>
<td>Merit-based Incentive Payment System</td>
</tr>
<tr>
<td>NCHPC</td>
<td>National Coalition for Hospice and Palliative Care</td>
</tr>
<tr>
<td>SRG</td>
<td>Survey Research Group</td>
</tr>
<tr>
<td>TECUPP</td>
<td>technical expert clinical user and patient panel</td>
</tr>
</tbody>
</table>
1. Background and Overview

Importance

Palliative care has expanded rapidly in the past 20 years, and there is growing consensus within the palliative care community regarding the need to systematically measure and support the quality of palliative care. However, there is still little information collected regarding the quality of palliative care delivered, particularly among patients who receive palliative services early in their disease in ambulatory (office) settings. It is likely that the quality and experience of palliative care received in ambulatory clinics differ substantially from palliative care received in other settings, such as the hospital or hospice. For example, ambulatory palliative care often supplements, and is provided concurrently with, another service, such as oncology. Rather than emphasizing treatment per se, care might focus on establishing values and preferences for care, identifying appropriate surrogate decisionmakers, and managing symptoms that may arise from treatment. In addition, patients might see multiple members of a palliative care team over several visits and establish a longer-term relationship, or they might receive only a single consultation. This variability in the patient experience of ambulatory palliative care underscores the need to systematically measure and assess the quality of this service, but it also raises important measurement challenges.

Stakeholders such as palliative care patients and their caregivers, as well as palliative care providers and health systems, lack actionable measures to assess performance and guide improvement efforts in palliative care, as noted by the National Quality Forum and the Centers for Medicare & Medicaid Services (CMS) Measures Application Partnership, as well as the 2017 CMS Environmental Scan and Gap Analysis Report (CMS, Health Services Advisory Group, 2017a). Measures of palliative care quality are also underrepresented in the CMS Quality Payment Program, with current measures addressing narrow populations that are often limited to patients with cancer or hospice patients. Perhaps most importantly, quality measurement using patient preferences and goals for care is notably absent, despite the patient-centered nature of palliative care (Teno, Anhang Price, and Makaroun, 2017; Anhang Price, Stucky, et al., 2018; Anhang Price and Elliott, 2018; Anhang Price, Elliott, et al., 2014). Patient-reported, patient-centered measures are a critical complement to other data on quality and can provide a more complete perspective on the overall care experience.

To address these gaps, CMS entered into a cooperative agreement (1V1CMS331639-01-00) 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 performance measures of palliative care experience in the domains of symptoms and communication. The performance measures are intended to assess the extent to which patients
who used ambulatory palliative care received the help they wanted for their pain and the extent to which they felt heard and understood by their palliative care provider and team. Under this agreement, AAHPM partnered with the National Coalition for Hospice and Palliative Care (NCHPC) and RAND Health Care to develop and test the proposed measures. Specifically, AAHPM sought to

- develop and test two patient-reported performance measures for patients receiving ambulatory palliative care
- ensure that the proposed measures incorporate patient voice and preferences and are broadly applicable to patients with serious illness and their families who receive palliative care services in a range of ambulatory primary and specialty care settings
- convene a technical expert panel that incorporates patient, caregiver, and family input directly into the performance measure development, specification, testing, and implementation processes
- submit the proposed palliative care measures to CMS’s 2021 Measures Under Consideration list for the Quality Payment Program and to the National Quality Forum for review and endorsement so that providers can measure and improve the quality of care in ambulatory palliative care.

Performance Measure Impact

Various studies and experts have highlighted significant unmet physical, psychological, and existential and spiritual needs among seriously ill people and gaps in symptom management and meaningful communication measures, despite the importance of these domains to patients and their families (CMS, Health Services Advisory Group, 2017a). These gaps are 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.

Patient-reported measures of palliative care experience should improve care quality by providing

- information to providers about how patients experience their care and whether their needs are being appropriately met
- a mechanism for systematically assessing quality across palliative care programs, enabling a standard of care, and facilitating quality improvement efforts in specific palliative care domains
- a rigorous and standardized measure that can be used to incentivize high performance in palliative care while rewarding adherence to the central tenets of palliative care (patient-centered multidisciplinary care)
- a way for patients and families to compare and select palliative care providers.

Proposed Performance Measures

The two proposed measures developed under this cooperative agreement are as follows:

- **Short title: Feeling Heard and Understood**
Palliative care outpatients’ experience of feeling heard and understood. A multi-data element measure consisting of four data elements:

- 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.”

Short title: Receiving Desired Help for Pain

Palliative care outpatients’ experience of receiving desired help for pain. The percentage of patients ages 18 years and older who had an ambulatory palliative care visit and report getting the help they wanted for their pain from their palliative care provider and team within six months of the ambulatory palliative care visit.

Measure Selection

The proposed performance measures were selected for development and testing using existing literature, input from an expert panel convened for this project, and results from several consensus panels held in recent years with experts, patient advocates, researchers, payers, and other stakeholders (Dy, Kiley, et al., 2015; National Consensus Project for Quality Palliative Care, 2018; CMS, Health Services Advisory Group, 2017b). Our selection of measures was informed by the following sources:

- Literature showing that many patients with life-limiting illness experience pain and other symptoms that are not adequately managed (Mistry et al., 2015; Heyland et al., 2006; Steinhauser, Christakis, et al., 2000; Steinhauser, Clipp, et al., 2000; Grudzen et al., 2010; Connors et al., 1995; Puntillo and Naidu, 2018), experience inadequate communication about prognosis and treatment options (Covinsky et al., 2000; Teno, Claridge, et al., 2004; Casarett et al., 2008a; Casarett et al., 2008b; Dy, Shugarman, et al., 2008; Teno, Lima, and Lyons, 2009), and receive care that is not consistent with their preferences (Teno, Claridge, et al., 2004; Teno, Lima, and Lyons, 2009; Teno, Freedman, et al., 2015; Khandelwal et al., 2017).

- Clinical practice guidelines for quality palliative care developed by the National Consensus Project for Quality Palliative Care (National Consensus Project for Quality Palliative Care, 2018) that recommend appropriate attention to communication among seriously ill patients and symptom management, with a focus on the patient’s and family members’ experience of care. Practice guidelines developed by various medical professional societies for the care of individuals with serious illness echo these recommendations, e.g., the American Society of Clinical Oncology (Ferrell et al., 2017) and the Society of Critical Care Medicine (Truog et al., 2008).

- High-priority palliative care items identified by NCHPC and AAHPM, in partnership with affiliates and experts, for CMS’s Quality Payment Program. The Coalition’s Quality Workgroup selected pain and symptom management and effective communication, with caregiver involvement and support related to pain management, as potential areas for further measurement work (Chen et al., 2020).
Focus groups led by the project team with palliative care providers and interviews with palliative care patients and family caregivers, which highlighted the importance of communication and symptom management in the context of high-quality palliative care, provided direction for useful questions in these two domains.

Taken together, these inputs point to the importance of measuring communication and symptom management as core domains of palliative care and highlight patient-centeredness and the patient and caregiver experience.

In this report, we provide a summary of the development and testing of the Feeling Heard and Understood performance measure. The summary of the work related to the Receiving Desired Help for Pain measure is reported separately (Figure 1.1).

**Figure 1.1. Measures Selected for Development**

Additional Rationale for the Focus on Feeling Heard and Understood

Seriously ill people often report feeling silenced, ignored, and misunderstood in medical settings (Institute of Medicine, 2015; Frosch et al., 2012; Norton et al., 2003). 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 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 (Street et al., 2009; Medendorp et al., 2017; Murray, McDonald, and Atkin, 2015). Assessing the extent to which the patient feels heard and understood has implications for communicating prognosis and treatment options and assuring adherence to the treatment plan (Gramling et al., 2016; Chochinov et al., 2015).

**Measure Alignment**

This performance measure aligns with the “Strengthen person and family engagement as partners in their care” goal of CMS’s Meaningful Measures Initiative and with the domain of “Person and Caregiver-Centered Experience and Outcomes” within MACRA. The goal of the Feeling Heard and Understood performance measure is to facilitate and improve effective
patient-provider communication that engenders trust, acknowledgment, and a whole-person orientation to the care that is provided. The desired and measurable outcome is that the patient feels heard and understood by the ambulatory palliative care provider and team.

Overview of Testing Process

The proposed patient-reported measures of ambulatory palliative care experience were rigorously tested in a multiphase process (Figure 1.2). First, we created a patient experience survey instrument, including data elements (i.e., items) for the proposed measures of Feeling Heard and Understood and Receiving Desired Help for Pain from the palliative care provider and team. The data elements necessary for these proposed measures have been tested and used in non–palliative care populations in different formats; however, the exact wording and structure proposed for these performance measures had not been used among ambulatory palliative care populations. The survey instrument also included other patient-reported data elements to capture information on respondent health and sociodemographic information, use and details of proxy involvement in survey completion, perceived overall quality of care and communication, and related measure concepts, such as receiving desired emotional support. These additional data elements were used for measure analyses, e.g., to explore data element–level and measure-level validity and risk adjustment. Once the instrument was drafted, we tested in two phases: (1) a pre-testing phase focused on cognitively testing and finalizing survey-based data elements for the proposed measures and establishing testing parameters through a small pilot, or alpha, test among five palliative care programs (see Rodriguez et al., 2020 for alpha test results) and (2) a national beta field test, described in this report.

Figure 1.2. Testing Phases in Measure Development

Pre-Testing Phase

The purpose of the pre-testing phase was to finalize the data elements prior to fielding the national beta field test. The alpha pilot test, which ran from August 2019 to October 2019, focused on providing evidence for the feasibility of data collection. Results are described in a separate report (Rodriguez et al., 2020). The pre-testing phase included the following elements:

- Cognitive testing of proposed data elements through 25 interviews with ambulatory palliative care patients and their family members. We established the comprehensibility,
readability, and adaptability of survey instructions and specific data elements, including response options.

- Recruitment of palliative care programs and the establishment of optimal site recruitment and retention processes.
- Assessment of data capabilities across palliative care programs and tailoring sample file requests accordingly.
- Field testing among five palliative care programs to
  - establish average times from program recruitment to survey fielding to inform the national beta field test timeline
  - explore the feasibility of using email to administer a web-based survey in the national beta field test
  - determine response rates and explore approaches to optimizing the number of completed responses
  - assess the rates of, and reasons for, proxy response versus patient response
  - identify data elements with low reliability, potentially caused by high rates of missingness or “topping out” (in which scores are clustered at the high end of the scale, making it difficult to distinguish meaningful differences between responses).

**National Beta Field Test Phase**

The nationally representative beta field test phase was conducted from November 2019 to February 2021. Through this large-scale national test, we sought to establish (1) psychometric properties of the performance measure data elements, (2) the scientific acceptability of the performance measure specifications (their feasibility with regard to administration, mode, and calculation of the performance measures), and (3) final measure technical specifications, the construction of the rules regarding the numerator and denominator, and the reliability of the performance measures. Central to this process was establishing data element–level and measure-level reliability and validity.

To meet these goals, we fielded an enhanced mixed-mode survey to eligible patients receiving care in 44 ambulatory palliative care programs. The survey instrument, developed and refined during the pre-testing phase described earlier, included data elements that were relevant to the two proposed performance measures (*Feeling Heard and Understood* and *Receiving Desired Help for Pain*), as well as data elements that were necessary for testing and analysis. Using data collected via the survey instrument, we conducted all data-element and measure analyses and developed our final measure specifications. During the national beta field test phase, we also conducted interviews with a subsample of participating ambulatory palliative care programs to elucidate implementation considerations and to better understand the actual and anticipated impacts of the coronavirus disease 2019 (COVID-19) pandemic on service delivery and quality measurement capabilities. We also conducted interviews with patients and caregivers of patients from racial and ethnic minorities to obtain a more robust understanding of how palliative care experience might differ across diverse populations and inform future refinements to palliative care performance measurement, as necessary.
Organization of the Report

In this report, we describe the results of our national beta field test for the *Feeling Heard and Understood* performance measure. We provide a detailed overview of our full performance measure testing process for both the *Feeling Heard and Understood* data element (i.e., the patient-reported scale) and the performance measure. In Chapter 2, we describe our patient population and sampling methodology, data collection procedures, and the characteristics of our final respondent sample used for testing. In Chapter 3, we first present the methods used to evaluate the reliability and validity of the *Feeling Heard and Understood* data element and then present the results of the data element reliability and validity testing. In Chapter 4, we outline the methods used to construct the *Feeling Heard and Understood* performance measure and to test its reliability and validity. We then describe the results of the performance measure testing that led to our final measure specifications. In Chapter 5, we describe our feasibility assessment approach and key findings, and we outline considerations for measure implementation and future work in Chapter 6. The results of our national beta field test for the *Receiving Desired Help for Pain* performance measure are described in a companion report (Ahluwalia et al., 2021).
2. Patient Population and Sampling

Target Measure Population and Data Sources

The target population for testing the *Feeling Heard and Understood* performance measure included all patients ages 18 years and older who had at least one ambulatory palliative care visit. Feedback elicited from our technical expert clinical user and patient panel (TECUPP) and project advisory group underscored the relevance of the proposed measure to all adult patients receiving ambulatory palliative care from a palliative care provider and team.

To develop a sampling frame, we first identified a total of 395 ambulatory palliative care programs across the United States through two key sources:

1. The Center to Advance Palliative Care, a project partner and national organization dedicated to increasing the availability of quality health care for people living with a serious illness, collects and manages self-reported palliative care program information annually via the Mapping Community Palliative Care Project and the National Palliative Care Registry. We identified 360 programs through the Center to Advance Palliative Care.

2. NCHPC, a project partner and membership-based advocacy group, hosted an informational webinar for its 13 member organizations regarding the performance measure development effort, which then advertised the effort among their respective constituencies. Interested palliative care organizations reached through this effort were asked to submit a project interest form via the NCHPC project website. We identified 35 programs through NCHPC.

Program Recruitment

We aimed to recruit a total of 45 programs for the national beta field test, using 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) and stratifying recruitment efforts by administrative home type (i.e., hospice, hospital, ambulatory, and other administration) and geographic location to ensure representation across Census Regions (Table 2.1). Using the list of 395 ambulatory palliative care programs, we sorted programs into recruitment queues according to these criteria. The RAND Corporation’s Survey Research Group (SRG) began contacting programs to discuss participation in the national 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 Merit-based Incentive Payment System (MIPS) eligibility 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 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 and 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 an adequate sample size for testing purposes
- were a Program of All-Inclusive Care for the Elderly or Veterans Administration program and thus not eligible for the MIPS program
- 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 test period; Table 2.1).

<table>
<thead>
<tr>
<th></th>
<th>Midwest</th>
<th>Northeast</th>
<th>South</th>
<th>West</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted sites to recruit</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Sites recruited</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Hospital</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted sites to recruit</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Sites recruited</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>Ambulatory/other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted sites to recruit</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Sites recruited</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>All settings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted sites to recruit</td>
<td>11</td>
<td>11</td>
<td>15</td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td>Sites recruited</td>
<td>9</td>
<td>13</td>
<td>12</td>
<td>10</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 2.1. National Beta Field Test Recruitment Targets and Final Recruitment Data

Patient Sampling

Our target patient sample was guided by several assumptions that we based, in part, on our alpha pilot test findings, including that (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, according to existing literature (Parast, Elliott, et al., 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.

To reflect a quarterly survey fielding cycle as envisioned for future implementation (see Figure 6.1 for implementation guidance), we assigned participating programs to one of three groups according to the time of their entry into the field test, and patient sampling was conducted.
each month, rotating through the three groups of programs such that each program participated once per quarter. This approach also allowed us to better manage fielding workload and track survey return. In March 2020, following the outbreak of the COVID-19 pandemic, we decided to temporarily halt data collection. Data collection resumed in September 2020, with a planned additional four rounds of data collection. Overall, we completed ten separate rounds of sampling and data collection (rounds one through six before the COVID-19 pandemic and rounds seven through ten during the restart period).

Within each round, the sampling procedures were as follows:

1. Each program participating in the round was required to submit a file of all ambulatory palliative care visits occurring in the prior six months (from the date of the data pull) through RAND’s secure file transfer system. The files were then transferred to a secure computing environment at RAND. The file had to include visit date, visit type, provider name, provider type, patient name, patient date of birth, and patient contact information for all records so that eligibility could be determined.

2. Each program’s file was processed by RAND staff and saved as a new clean file that contained one row for each visit that each patient had with each provider. If a file was missing any of the required fields, RAND followed up with the program to request the missing information. RAND staff also created a single ID for each patient, which required us to compare each data file to previous files submitted by each program so that the same patient would not be assigned two IDs. This was accomplished using a combination of manual checks, programmatic checks, and a probabilistic record linkage algorithm.

3. Once each file was processed, the clean data across all programs participating in the round were combined into a single data set. This data set was then refined to create the sampling file for SRG to use when fielding the survey instrument.

   a. 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 (i.e., the planned date for mailing the prenotification letter to patients; see the following section for details on data collection procedures) for the round. Although we asked programs to submit six months of data, 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, even if the six months of data they each submitted did not align.

   b. We next excluded patients under 18 years of age, patients identified as deceased, patients with contact information outside the United States, and patients who had already been included in a prior sample.

   c. 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).

   d. If patients had multiple visits, we selected the most recent visit for each patient with the reference provider.
Once these sampling procedures were completed, we created a final sample file for SRG to use for survey fielding. This file included only the fields necessary for survey administration, such as patient ID, patient name, patient contact information, provider name, and program name. This file was sent to SRG in the month prior to the start of fielding to allow time to prepare data collection materials.

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 because of capacity constraints, a random sample of eligible patients up to the number SRG could field in that period was selected.

Data Collection Procedures

Patient experience data were collected via the survey instrument. We used a mixed-mode survey administration protocol (i.e., web to mail to telephone follow-up; see Figure 2.1). First, a prenotification letter was mailed to patients without email addresses in their contact information to inform patients of their eligibility for the upcoming survey. The prenotification letter informed patients that they would receive the survey within a few days. It referenced their program and provider, as well as the organization conducting the research, and provided a toll-free number that patients could call if they had any questions about the survey. In addition, because some patients would likely be too ill to complete the survey or perhaps even open an envelope themselves, the prenotification letter gave advance notice for the patient to arrange help to complete the survey and alerted caretakers to the upcoming survey mailing (Dillman, Smyth, and Christian, 2014). Returned undeliverable letters or calls from caregivers who opened patient mailings provided feedback that the patient had moved or was deceased.

If patients had an email address in their contact information, an invitation to complete a web-based survey was sent by email in lieu of the prenotification letter. During the alpha pilot test, we established the feasibility of using an email and web-based survey mode, and a web survey was developed for administration during the national beta field test. In keeping with best practices, the web survey was formatted to present one question on each screen. Up to two email reminders were sent to emailed patients one and three days after the initial email. The prenotification letter was sent to respondents without email addresses on the same day as the email invitation, so that each sampled patient received either an email invitation or a prenotification letter. All other data collection activities were the same for patients with or without email addresses.

The prenotification letter was followed within a week by a mailed survey with cover letter. Patients who received the emailed invitation letter but had not completed the web survey within one week were also sent the mailed survey with cover letter. The survey cover letter included a reprise of the information in the prenotification along with a description of the voluntary nature of the survey, privacy and confidentiality safeguards, how the data would be used, and instructions that the patient could get help or use a proxy to complete the survey. The survey
packet also included a postage-paid postcard that could be returned with updated address information for the patient or to indicate that the patient was deceased.

If the survey was not returned within three weeks of the mailing, attempts were made to contact the patient by telephone and complete the survey via computer-assisted telephone interview (CATI). Up to eight attempts were made to complete the CATI survey over several weeks before the patient was considered a “nonresponse.”

**Figure 2.1. Timeline of Data Collection Procedures**

We fielded three versions of the survey instrument. The first version of the survey instrument contained 43 data elements. For round six in March 2020, which was the last fielding round prior to pausing data collection because of the COVID-19 pandemic, the survey instrument was updated to include five additional data elements asking about the impact of the COVID-19 pandemic. Beginning with round seven, when data collection resumed in September 2020, the COVID-19 section of the survey instrument was revised to include eight total data elements (Figure 2.2).

**Impact of COVID-19 Pandemic on the National Beta Field Test**

On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a global pandemic. As a result, most nonessential health services in the United States either halted or converted to remote delivery (i.e., through telehealth). In addition, many providers and other health care staff were deployed to critical settings, such as the intensive care unit or general hospital units, to provide surge capacity in the hardest-hit areas and at the height of the pandemic. National beta field testing, which began in November 2019, was in round five of survey fielding when the pandemic was declared. In light of the anticipated impact on resources the pandemic would take on palliative care programs, providers, and patients, we paused data collection after round six, which began on March 30 and closed on May 25, 2020. We pursued this sixth round before halting data collection because most participating programs for that round
had already submitted data files in preparation for fielding. In addition, we were able to modify the survey instrument for round six to include five additional data elements pertaining to the impact of the pandemic on the respondent’s daily activities and their use of telehealth services (Appendix A). During the pause in data collection, we conducted interviews with participating programs (see the “Program Perspectives” section in Chapter 5 for details) and queried programs regarding the impact of the pandemic on their service delivery and resources and regarding when they believed they would be able to resume data collection as part of the national beta field test. Using this information, we resumed data collection with round seven, which opened on September 14. For the remainder of the beta field test data collection, we used a third version of the survey that included eight additional data elements regarding the impact of the pandemic on the respondent’s daily activities and their use of telehealth services (Appendix A).

**Figure 2.2. National Beta Test Timeline**

<table>
<thead>
<tr>
<th>2019</th>
<th>2020</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov</td>
<td>Dec</td>
<td>Jan</td>
</tr>
<tr>
<td>Rounds 1–6</td>
<td>Pause in data collection</td>
<td>Rounds 7–10</td>
</tr>
</tbody>
</table>

Of the 44 programs that participated in the national beta field test,
- 39 programs participated both before and after the pause in data collection
- three participated before the pause but did not rejoin the field when data collection resumed
- two programs joined the national beta field test after data collection resumed (one was newly recruited during our data collection pause, and the other had been recruited but was only able to send a sample file after the pause).

**Sample Characteristics**

We fielded a total of 7,595 surveys to eligible patients in the national beta field test, of which 2,804 are completed surveys, or cases, that were used for analysis (Table 2.2). *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). Completed surveys could still have some missing items.
Table 2.2. Final Disposition of Fielded Surveys \((n = 7,595)\)

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed surveys</td>
<td>2,804</td>
<td>36.92%</td>
</tr>
<tr>
<td>Mail</td>
<td>1,298</td>
<td>17.09%</td>
</tr>
<tr>
<td>Phone</td>
<td>980</td>
<td>12.90%</td>
</tr>
<tr>
<td>Web</td>
<td>526</td>
<td>6.93%</td>
</tr>
<tr>
<td>Survey nonresponse</td>
<td>3,356</td>
<td>44.19%</td>
</tr>
<tr>
<td>Excluded surveys due to ineligibility</td>
<td>1,435</td>
<td>18.89%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7,595</strong></td>
<td></td>
</tr>
</tbody>
</table>

Chapter 4 describes denominator exclusions in detail; here, we provide 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, and inability to reach patient after maximum attempts (Table 2.3).

**Table 2.3. Survey Nonresponse \((n = 3,356)\)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum calls (8) reached without response</td>
<td>2,128</td>
</tr>
<tr>
<td>Not found (i.e., bad phone number)</td>
<td>583</td>
</tr>
<tr>
<td>Refusals</td>
<td>463</td>
</tr>
<tr>
<td>Final refusal (patient reached but refused participation)</td>
<td>348</td>
</tr>
<tr>
<td>Informant refusal (someone other than patient or proxy declined)</td>
<td>87</td>
</tr>
<tr>
<td>Breakoff (respondent discontinued during CATI)</td>
<td>28</td>
</tr>
<tr>
<td>Patient unable and no available proxy</td>
<td>121</td>
</tr>
<tr>
<td>Late complete</td>
<td>61</td>
</tr>
<tr>
<td>Mail</td>
<td>58</td>
</tr>
<tr>
<td>Web</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,356</strong></td>
</tr>
</tbody>
</table>

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 might have notified us when reached by telephone), returned surveys completed solely by a proxy without patient involvement (see Chapter 4 for exclusion analysis related to proxy assistance), surveys in which the respondent disavowed the program (i.e., indicated that they had not received care in the past six months from the stated palliative care provider and team), or
patients for whom we had bad contact information, such as when a patient moved or the mailed packet was returned to sender (Table 2.4).

Table 2.4. Excluded Surveys Due to Ineligibility (n = 1,435)

<table>
<thead>
<tr>
<th>Ineligibility Category</th>
<th>Count (% of total excludes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient deceased</td>
<td>748 (52%)</td>
</tr>
<tr>
<td>Proxy-only response</td>
<td>435 (30%)</td>
</tr>
<tr>
<td>Disavowed program/provider</td>
<td>146 (10%)</td>
</tr>
<tr>
<td>Bad contact information; moved</td>
<td>35 (2%)</td>
</tr>
<tr>
<td>Multiple ineligibilities (e.g., patient deceased and proxy-only response)</td>
<td>71 (5%)</td>
</tr>
<tr>
<td>Total</td>
<td>1,435</td>
</tr>
</tbody>
</table>

Survey respondents, i.e., those who comprised our analytic sample (n = 2,804), generally reflected the larger patient sample (n = 7,595). Respondents were predominantly female (56 percent, compared with 55.3 percent in the full sample), with a mean age of 63.4 years (compared with 62 years in the full sample). According to their submitted survey data, respondents were largely White (88 percent) and non-Hispanic/Latino (95 percent) and very educated, with 66 percent having some college or more (Table 2.5).

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).
### Table 2.5. Patient Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Summary</th>
<th>Standard Deviation</th>
<th>Percentage Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>$n = 2,804$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (mean)</td>
<td>63.36</td>
<td>13.32</td>
<td>0.04%</td>
</tr>
<tr>
<td>Gender (male %)</td>
<td>43.81%</td>
<td></td>
<td>0.04%</td>
</tr>
<tr>
<td>Race</td>
<td>$n = 2,753$</td>
<td></td>
<td>1.82%</td>
</tr>
<tr>
<td>White</td>
<td>87.61%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>5.88%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>0.91%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiracial</td>
<td>2.76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>0.44%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native Hawaiian or other Pacific Islander</td>
<td>0.25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2.14%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>$n = 2,782$</td>
<td></td>
<td>0.78%</td>
</tr>
<tr>
<td>More than 4-year college degree</td>
<td>15.74%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-year college graduate</td>
<td>15.46%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some college or 2-year degree</td>
<td>34.94%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school graduate or GED</td>
<td>25.63%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some high school but did not graduate</td>
<td>6.40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8th grade or less</td>
<td>1.83%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>$n = 2,743$</td>
<td></td>
<td>2.18%</td>
</tr>
<tr>
<td>Yes, Hispanic or Latino</td>
<td>4.67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No, not Hispanic or Latino</td>
<td>95.33%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Distribution of Completed Surveys Across Participating Programs**

The number of completed surveys varied across programs. For example, although approximately one-third of participating programs each contributed 31 to 50 completed surveys across the national beta field test period, one program contributed more than 500 completed surveys (Figure 2.3).
Program-Level Nonresponse

The rate of nonresponse to fielded surveys (after removing exclusions) was approximately 54.4 percent (i.e., 3,356 / (2,804 + 3,356)). Figure 2.4 demonstrates that there are no clear outliers in terms of program nonresponse. The distribution of nonresponse is clustered around the mean, suggesting similarity in nonresponse levels across programs.
3. Critical Data Element Testing

In this chapter, we present an overview of the methods used to develop and test the reliability and validity of the *Feeling Heard and Understood* data element. We then describe the results of our data element testing. All analyses described here pertain to the individual respondent level.

Methods

*Scale Development*

The construct of feeling heard and understood previously was measured using a single data element (“Over the past two days, how much have you felt heard and understood by the doctors, nurses, and hospital staff?”) that was developed for the inpatient palliative care setting (Gramling et al., 2016). To improve the reliability of the data element and, ultimately, the performance measure, we sought to develop a scale comprising multiple data elements that reflect the construct of feeling heard and understood. Through an extensive literature review and information-gathering effort (Chen et al., 2020), we selected ten data elements to be evaluated for inclusion as part of the national beta field test (Table 3.1). Where appropriate, selected data elements were modified to align with our testing parameters and palliative care context and further refined through cognitive interviews with patients and caregivers. Prior to testing, our TECUPP and project advisory group provided input on the final wording, response options, and inclusion of each data element in the survey instrument. We evaluated the psychometric properties of the ten data elements using classical test statistics (e.g., Cronbach’s alpha coefficient and item-total correlations), as well as factor analysis and item response theory analysis.
Table 3.1. Data Elements Evaluated for Inclusion in the National Beta Field Test

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Construct Related to Feeling Heard and Understood</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt heard and understood by this provider and team.</td>
<td>Feeling heard and understood</td>
<td>Gramling et al., 2016</td>
</tr>
<tr>
<td>I felt comfortable asking this provider and team questions.</td>
<td>Therapeutic alliance</td>
<td>Mack et al., 2009</td>
</tr>
<tr>
<td>I trusted this provider and team.</td>
<td>Trust</td>
<td>Safran, Kosinski, et al., 1998</td>
</tr>
<tr>
<td>I could tell this provider and team anything, even things I might not tell anyone else.</td>
<td>Trust</td>
<td>Safran, Karp, et al., 2006</td>
</tr>
<tr>
<td>I felt this provider and team put my best interests first when making recommendations about my care.</td>
<td>Trust</td>
<td>Safran, Karp, et al., 2006</td>
</tr>
<tr>
<td>I felt this provider and team always told me the truth about my health, even if there was bad news.</td>
<td>Trust</td>
<td>Safran, Kosinski, et al., 1998</td>
</tr>
<tr>
<td>I felt this provider and team saw me as a person, not just someone with a medical problem.</td>
<td>Whole-person orientation; therapeutic alliance</td>
<td>Shi, Starfield, and Xu, 2001; Mack et al., 2009</td>
</tr>
<tr>
<td>I felt this provider and team knew what worried me most about my health.</td>
<td>Whole-person orientation</td>
<td>Safran, Kosinski, et al., 1998</td>
</tr>
<tr>
<td>I felt this provider and team understood what is important to me in my life.</td>
<td>Whole-person orientation</td>
<td>New; Shi, Starfield, and Xu, 2001</td>
</tr>
<tr>
<td>I felt this provider and team would know what I would want done if I was unconscious or in a coma.</td>
<td>Values alignment</td>
<td>New</td>
</tr>
</tbody>
</table>

Data Element Reliability

We evaluated the reliability of the Feeling Heard and Understood multi-data element scale 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 times. We invited patients who completed the survey by phone (i.e., the CATI survey) to be re-administered a shortened survey, including the Feeling Heard and Understood data elements, at a second point 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 235 of these also providing responses at Time 2. 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 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), in which 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 bivariate correlation, with a correlation of at least 0.70 demonstrating acceptable reliability.

Data Element Validity

We also assessed the convergent validity of the proposed Feeling Heard and Understood scale. As part of the study design, we included additional survey data elements from other instruments in use (e.g., the Consumer Assessment of Healthcare Providers and Systems [CAHPS] Hospice Survey) that we expected would be related to Feeling Heard and Understood. Our selection of additional data elements was based on theory, prior literature, and clinical practice and/or expert feedback. Specifically, we hypothesized that Feeling Heard and Understood would be associated with better patient-provider communication and thus included the four-item CAHPS Communication composite measure (e.g., “In the last three months, how often did this provider and team listen carefully to you?”), which assesses the quality of key doctor-patient communication behaviors. We also hypothesized that Feeling Heard and Understood would be positively associated with the companion measure under testing, Receiving Desired Help for Pain (included data element: “In the last six months, did you get as much help as you wanted for your pain from this provider and team?” [Yes, definitely; Yes, somewhat; No]). Both measures assess patient-centered care and the extent to which the provider understood what was important to the patient. Associations between the proposed data elements and validity items were evaluated using bivariate correlations. Our interpretation of correlations followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50; Rosnow and Rosenthal, 1989).

Results

Scale

Corrected classical item-total correlation test statistics and exploratory factor analyses supported a single underlying factor for the ten Feeling Heard and Understood data elements; however, two of the ten data elements were not strong indicators of the factor and were removed (detailed information on the psychometric analyses will be reported in a separate publication). According to an item response theory model, the data element properties showed that the eight remaining data elements were strongly related to the underlying construct. Informed by the results of the item response theory analyses combined with TECUPP ratings of the importance of each data element, we selected a four–data element subset to comprise the Feeling Heard and Understood scale:

• 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.”

Individuals can respond to each data element using discrete values ranging from 0 (“Not at all true”) to 4 (“Completely true”). A response of 4 (“Completely true”) is considered a top-box response.

The distribution of responders (in percent) who selected each response option for each Feeling Heard and Understood performance measure data element is shown in Table 3.2 and demonstrates that, on average, individuals are responding at the high end of the response scale, with low levels of missingness.

### Table 3.2. Distribution of Responses to the Feeling Heard and Understood Performance Measure Data Elements (in percent)

<table>
<thead>
<tr>
<th>Response</th>
<th>Value</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completely true</td>
<td>4</td>
<td>69.41</td>
<td>74.81</td>
<td>77.9</td>
<td>63.64</td>
</tr>
<tr>
<td>Very true</td>
<td>3</td>
<td>22.54</td>
<td>18.03</td>
<td>15.65</td>
<td>25.14</td>
</tr>
<tr>
<td>Somewhat true</td>
<td>2</td>
<td>4.66</td>
<td>4.70</td>
<td>4.00</td>
<td>8.22</td>
</tr>
<tr>
<td>A little bit true</td>
<td>1</td>
<td>1.55</td>
<td>1.23</td>
<td>1.51</td>
<td>1.41</td>
</tr>
<tr>
<td>Not at all true</td>
<td>0</td>
<td>1.84</td>
<td>1.23</td>
<td>0.94</td>
<td>1.59</td>
</tr>
<tr>
<td>Missing</td>
<td>—</td>
<td>1.25</td>
<td>1.32</td>
<td>1.07</td>
<td>1.53</td>
</tr>
</tbody>
</table>

**Critical Data Element–Level Missingness**

Among the 2,804 completed surveys, the mean data element–level missingness is 0.8 percent for the entire survey. Appropriately skipped survey data elements are not counted as missing. Among the four data elements that comprise the Feeling Heard and Understood performance measure, the data element–level missingness ranged from 1.07 percent (Q3: “I felt this provider and team saw me as a person, not just someone with a medical problem”) to 1.53 percent (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 scale data elements. Among responders who have any missing data across the four data elements ($n = 60$), approximately 43 percent ($n = 26$) skipped all of the performance measure data elements, whereas 47 percent ($n = 28$) skipped only individual questions (Table 3.3). This suggests there are generally low levels of missingness and no discernible patterns.
Table 3.3. Frequency of Number of Questions Missing Responses in the *Feeling Heard and Understood* Performance Measure Data Element Set

<table>
<thead>
<tr>
<th>Data Elements with Missing Response</th>
<th>Number of Respondents with Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data elements</td>
<td>26</td>
</tr>
<tr>
<td>Three data elements</td>
<td>1</td>
</tr>
<tr>
<td>Two data elements</td>
<td>5</td>
</tr>
<tr>
<td>One data element</td>
<td>28</td>
</tr>
<tr>
<td>Total missing responses</td>
<td>60</td>
</tr>
</tbody>
</table>

**Critical Data Element Correlations**

The correlation among the performance measure data elements ranged from 0.54 to 0.69. The highest agreement among the data elements is between Q2 (“I felt this provider and team put my best interests first when making recommendations about my care”) and Q3 (“I felt this provider and team saw me as a person, not just someone with a medical problem”).

**Critical Data Element Reliability**

Results provide support for the reliability of the *Feeling Heard and Understood* scale. There was high internal consistency (Cronbach’s alpha = 0.90), and 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 reliable and can, therefore, be used in the construction of the performance measure.

**Critical Data Element Validity**

Results of validity testing also support the use of the *Feeling Heard and Understood* score. As hypothesized, higher scores on the *Feeling Heard and Understood* scale were associated with higher CAHPS communication scores ($r = 0.54, p < 0.001$) and *Receiving Desired Help for Pain* ($r = 0.48, p < 0.001$). Taken together, these results support the convergent validity of the *Feeling Heard and Understood* scale.
4. Performance Measure Testing

In this chapter, we first present an overview of the methods used to develop and test the *Feeling Heard and Understood* performance measure and then describe the results of our performance measure testing. All analyses described here pertain to the program level.

**Methods**

*Performance Measure Characteristics*

We first describe the distribution of unadjusted measure scores by presenting two visual summaries of the distributions of scores: (1) the distribution of the proportion of top-box responses across all respondents and (2) the average proportion of top-box responses across programs.

*Risk Adjustment*

As suggested by our 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 it might be appropriate to adjust the performance measure calculation to account for differences in performance caused by extrinsic factors beyond the control of the palliative care program or provider. The relevant factors are those that systematically differ across programs and are related to the performance 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 affect 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 affect experiences of care and/or response tendencies (Ingersoll et al., 2018; Elliott, Swartz, et al., 2001; Zaslavsky et al., 2001; Elliott, Zaslavsky, et al., 2009). For example, Ingersoll et al., 2018, finds that 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 among individuals hospitalized with metastatic cancer. 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 in CAHPS surveys (Anhang Price, Elliott, et al., 2014). In addition, primary diagnosis is closely related to disease burden and likely affects ratings of feeling heard and understood independent of a provider’s or practice’s performance.

- **Patient race and ethnicity.** The relationship between race and ethnicity and experiences of patient-provider communication is mixed. For example, some studies (Welch, Teno,
and Mor, 2005; Smith, Davis, and Krakauer, 2007) 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,
whereas 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 nonproxy respondents (O’Malley et al.,
2005).

Using our literature review and guided by TECUPP and project advisory group feedback, we
selected a set of potential risk factors to be evaluated in our final models. We selected data from
three sources: patient information provided by programs in their submitted data files, responses
from completed surveys (e.g., proxy respondent characteristics), and U.S. Census data matched
to the ZIP code of a patient’s residence. We assessed variables 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. We tested
binary or categorical variables for association with the *Feeling Heard and Understood* measure
using Fisher’s exact test. Fisher’s exact test was also used to assess whether there were
differences in these variables across programs. For continuous variables, we used a \( Z \)-test from a
generalized linear model to test for an association with the *Feeling Heard and Understood*
measure, and we used an analysis of variance (ANOVA) \( F \)-test to test for differences among
programs. We adjusted \( p \)-values using the Benjamini-Hochberg correction for multiple
comparisons to control the false discovery rate (Benjamini and Hochberg, 1995).

Our selected set of potential risk adjusters are as follows:

- **Survey data or program information**
  - patient education
  - patient Hispanic
  - patient language
  - patient race
  - proxy level
  - survey mode

- **program data**
  - patient age
  - patient female

- **Census data**
  - female residents (%)
  - marriage status (% of residents ages 15 and older)
  - disabled residents (%)
  - labor force participation (% of residents ages 16 and older)
  - employed residents (% of residents ages 16 and older)
  - unemployed residents (% of residents ages 16 and older)
- median household income ($)
- families below the poverty line (%)
- urban residents (%)
- owner-occupied housing units (%)
- residents with health insurance (%)
- private health insurance only (% of noninstitutionalized residents)
- public health insurance only (% of noninstitutionalized residents)
- race = American Indian or Alaska Native (%)
- race = Asian (%)
- race = Black (%)
- race = White (%)
- ethnicity = Hispanic (%).

Diagnosis Analysis

We also considered the potential for patient diagnoses to vary across programs and affect measure scores. In the development of the Hospice CAHPS measures, investigators found that primary diagnosis varied across hospice programs and was significantly and strongly associated with assessments of experience and therefore was included as a case-mix adjustment variable (Parast, Haas, et al., 2018). Although the target respondent population and setting were different from ours (bereaved caregiver versus patient, hospice versus 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, as follows:

- Not all programs consistently provided diagnostic information across the ten rounds of fielding.
- Programs that provided 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 International Classification of Diseases, Tenth Revision (ICD-10) codes, Current Procedural Terminology (CPT) codes, problem codes, reasons for visit codes, program- or software-specific codes, and free text.

Instead, we undertook an exploratory descriptive analysis to identify any indication 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) we used the primary diagnosis if it was indicated in the program file, and (2) if the primary diagnosis was not indicated, but usable diagnosis data were provided, we assigned primary diagnosis by applying the following condition hierarchy according to prevalence in our sample, prior research (Wennberg et al., 2004; Keating et al., 2016; Wachterman et al., 2016), and our research team physicians’ expert opinion:
1. cancer (both solid and liquid)
2. non-neurological end-organ disease (e.g., heart failure, end-stage liver disease, renal failure, and chronic obstructive pulmonary disease)
3. dementia (e.g., Alzheimer’s, vascular, and others)
4. movement disorders (e.g., stroke, Parkinson’s, multiple sclerosis, and amyotrophic lateral sclerosis [ALS])
5. other diagnosis (e.g., fibromyalgia, sequelae of diabetes, acquired immunodeficiency syndrome [AIDS], and symptom-only diagnoses, such as dyspnea).

All other data were categorized as missing (e.g., noninterpretable diagnostic information). A physician then reviewed these groups to ensure clinical accuracy. See Table 4.1 for counts of each assigned diagnosis group.

**Table 4.1. Diagnosis Groupings**

<table>
<thead>
<tr>
<th>Diagnosis Grouping</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer (solid and liquid)</td>
<td>1,685</td>
<td>60</td>
</tr>
<tr>
<td>End-organ disease (non-neurological)</td>
<td>225</td>
<td>8</td>
</tr>
<tr>
<td>Dementia</td>
<td>14</td>
<td>0.5</td>
</tr>
<tr>
<td>Movement disorders (e.g., Parkinson’s, multiple sclerosis)</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>Other (e.g., Parkinson’s, multiple sclerosis)</td>
<td>561</td>
<td>20</td>
</tr>
<tr>
<td>Missing</td>
<td>290</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,804</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

NOTE: Percentages do not add to 100 because of rounding.

Because of the low numbers of respondents with an assigned primary diagnosis of dementia or movement disorders, we focused on comparing differences among 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 *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 (percentage of top-box responses) by diagnosis group.

**Performance Measure Exclusions**

With guidance from our project advisory group and the TECUPP, we considered five exclusions from the proposed denominator of all adult patients with an ambulatory palliative care visit:

1. patients who did not complete at least one of the four data elements in the multi–data element measure ($n = 26$)
2. patients who did not complete and return the patient experience survey within six months of the eligible ambulatory palliative care visit ($n = 3,356$)
3. patients who responded on the patient experience survey that they did not receive care by the listed ambulatory palliative care provider in the past 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 at least one of the four data elements in the multi–data element measure, who did not return the survey at all, who disavowed the program, or who died before the survey could be completed were necessary exclusions because no survey data for the performance measure would be available. We further excluded the small number of 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 \( (n = 61\) out of 3,356 nonrespondents, or 1.8 percent). 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 (see the “Sample Characteristics” section in Chapter 2).

We also considered the impact of proxy assistance on measure responses to determine whether to apply exclusions based on proxy assistance. 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, they retain this variable to alleviate ongoing concerns about the potential for impact (CMS, 2020; National Cancer Institute Division of Cancer Control & Population Sciences, 2020).

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), and proxy only (proxy answered all questions, and patient was not involved). We compared descriptive statistics for the performance measure components for each of these three groups to inform the impact of proxy assistance and to determine whether to include or exclude proxy responses.

**Performance Measure Scoring**

To identify the optimal approach to scoring the *Feeling Heard and Understood* performance measure, we considered and compared reliability estimates for the following approaches:

1. Setting a threshold on a summary of data element–level scores. For example, we could sum the four data elements into a single scale score and then set a threshold to define “feeling heard and understood,” i.e., passing the performance measure. Another example of this approach would be setting a threshold on the number of top-box (i.e., 4: “Completely true”) responses across all four data elements in the scale. This approach consolidates all responses into a single binary outcome for each individual and thus risks reducing signal on the concept.
2. Treating each of the four data elements of the *Feeling Heard and Understood* scale as reflecting the construct of feeling heard and understood. Under this idea, the four
individual responses could be aggregated to estimate the program-specific proportion of individuals who feel heard and understood by their provider. This approach greatly increases the reliability estimate because there are four responses rather than one response for each individual; therefore, overall sample size is also increased. In our approach, we define each data element as passing if an individual “top-box” responds, and thus each individual contributes at most four binary outcomes to the overall score. The overall measure for a program summarizes the overall proportion of top-box responses across all individuals and data-elements to which individuals within a program respond.

We therefore pursued the second approach for consolidating measure scores at the program level (rather than thresholding individual composite scores) because of the increased measure reliability.

To estimate risk-adjusted program-level measure scores, we use hierarchical generalized linear models. The hierarchy of the data is patient observations within programs. The model is adjusted for survey mode and proxy assistance; covariates were set to the mail survey mode and “no proxy assistance” values. In all analyses, we only include data from surveys with at least one (of four) data elements completed for the Feeling Heard and Understood performance measure (i.e., those with no outcome responses were removed from the sample). See Appendix B for technical details.

Statistically Significant and Meaningful Differences in Performance

We guided our analyses using 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 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, our null and alternative hypotheses can be tested using a likelihood ratio statistic. See Appendix B for technical details.

Because the concept of clinically meaningful or practical significance difference is a notion without perfect agreement among researchers on how it should be defined (Quigley et al., 2018), 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 the equivalence of a difference in measure score to a 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.

Performance Measure Score Reliability

To assess the reliability of the performance measure score at the program level, we used a traditional “signal-to-noise” analysis that decomposes variability in the performance measure score into (1) between-subject variability and (2) 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, for which the hierarchy of data is patient observations within the program. The variance of the model can be decomposed using the adjusted intraclass correlation coefficient (ICC), which provides a summary of the reliability of the performance measure as tested, with higher values implying more variability between programs. Additionally, 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 caused by baseline differences in risk variables (including survey mode) that represent the programs. The reliability from the performance measure test is then projected out according to 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 performance measure. Reliability values of approximately 0.7 are our target for an acceptable level of reliability and help determine required within-provider sample sizes. See Appendix B for technical details.

**Performance Measure Score Validity**

To evaluate the validity of the performance measure, we examined the association of the *Feeling Heard and Understood* performance measure score with the *Receiving Desired Help for Pain* measure score, the CAHPS communication measure score, and individuals’ overall rating of their palliative care provider and team at the program level, with the hypothesis that scores would be positively associated. Associations between the proposed performance 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; Rosnow and Rosenthal, 1989).

**Face Validity**

Face validity of the performance measure score was determined through a systematic and transparent process by convening experts who explicitly addressed whether scores resulting from the performance measure, as specified, can be used to distinguish good from poor quality. In May 2021, following completion of testing, a panel of seven advisers 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 performance measure score. The expert panel consisted of six palliative care physicians and a researcher with expertise in palliative care communication. Six of the seven advisers also had expertise in clinical quality measurement. Advisers were asked to consider how well the performance measure scoring approach distinguishes between programs with high, medium, and low performance and how useful the
measure is to quality improvement efforts. Advisers 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).

Results

Performance Measure Characteristics

Distribution of Top-Box Responses

The *Feeling Heard and Understood* performance measure score, defined as the proportion of top-box responses to the four–data element scale (0 to 1), generates the score distribution depicted in Figure 4.1.

![Figure 4.1. Distribution of the Proportion of Top-Box Responses Across All Individuals](image)

As shown by this distribution, which has a mean of 0.71 and a large proportion of respondents in the highest score category, responders skew toward top-box responses across the four data elements, indicating that they trend toward feeling completely heard and understood by their providers. Although this distribution suggests low variability in individual measure responses, the distribution of the average percentage of top-box responses at the program level demonstrates a much higher degree of variability (Figure 4.2).
Figure 4.2. Distribution of the Average Percentage of Top-Box Responses Within Programs

Program Performance: Heard & Understood Measure
Standard Deviation in Average Program Scores = 11.47

Number of Programs

0 2 4 6 8 10

Unadjusted Measure Score: Proportion of Top Box Response Across Individuals within a Program

Risk Adjustment

Table 4.2 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 $\alpha = 0.05$ 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 performance measure after adjustment for multiple comparisons. In particular, survey mode was the only variable to demonstrate statistically significant associations with both the performance 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 4.3. As seen in Table 4.3, scores did not differ much between mail and web responses but were slightly lower for responses by phone. We provide further interpretation of this finding and describe its implications for future implementation in Chapter 6.
Table 4.2. Association of Potential Risk Factors with *Feeling Heard and Understood* Performance Measure Score and Program (sorted by measure score p-values)

<table>
<thead>
<tr>
<th>Potential Risk Adjuster</th>
<th>Effect Size</th>
<th>Test</th>
<th>Statistic</th>
<th>p-value</th>
<th>Adjusted p-value</th>
<th>Test</th>
<th>Statistic</th>
<th>p-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Survey data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey mode</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.000</td>
<td>0.013</td>
<td>Fisher</td>
<td>NA</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Patient race</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.049</td>
<td>0.168</td>
<td>Fisher</td>
<td>NA</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Proxy assistance</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.242</td>
<td>0.420</td>
<td>Fisher</td>
<td>NA</td>
<td>0.686</td>
<td>0.686</td>
</tr>
<tr>
<td>Patient education</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.271</td>
<td>0.432</td>
<td>Fisher</td>
<td>NA</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Patient Hispanic</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.541</td>
<td>0.732</td>
<td>Fisher</td>
<td>NA</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Patient language</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.656</td>
<td>0.732</td>
<td>Fisher</td>
<td>NA</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Program data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient age</td>
<td>−0.005</td>
<td>Z</td>
<td>−1.754</td>
<td>0.080</td>
<td>0.207</td>
<td>F</td>
<td>6.780</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Patient female</td>
<td>NA</td>
<td>Fisher</td>
<td>NA</td>
<td>0.704</td>
<td>0.732</td>
<td>Fisher</td>
<td>NA</td>
<td>0.051</td>
<td>0.054</td>
</tr>
<tr>
<td><strong>Census data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owner occupied housing unit</td>
<td>0.006</td>
<td>Z</td>
<td>2.624</td>
<td>0.009</td>
<td>0.114</td>
<td>F</td>
<td>6.437</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Median household income</td>
<td>0.000</td>
<td>Z</td>
<td>2.429</td>
<td>0.015</td>
<td>0.132</td>
<td>F</td>
<td>17.682</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Female</td>
<td>0.034</td>
<td>Z</td>
<td>2.202</td>
<td>0.028</td>
<td>0.144</td>
<td>F</td>
<td>4.591</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Family below poverty line</td>
<td>−0.012</td>
<td>Z</td>
<td>−2.203</td>
<td>0.028</td>
<td>0.144</td>
<td>F</td>
<td>12.190</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Married</td>
<td>0.007</td>
<td>Z</td>
<td>2.048</td>
<td>0.041</td>
<td>0.168</td>
<td>F</td>
<td>7.378</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Health insurance public only</td>
<td>−0.008</td>
<td>Z</td>
<td>−1.946</td>
<td>0.052</td>
<td>0.168</td>
<td>F</td>
<td>20.076</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Race American Indian or Alaska Native</td>
<td>−0.018</td>
<td>Z</td>
<td>−1.853</td>
<td>0.064</td>
<td>0.185</td>
<td>F</td>
<td>3.674</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Disabled</td>
<td>−0.013</td>
<td>Z</td>
<td>−1.632</td>
<td>0.103</td>
<td>0.243</td>
<td>F</td>
<td>12.870</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Health insurance private only</td>
<td>0.004</td>
<td>Z</td>
<td>1.556</td>
<td>0.120</td>
<td>0.259</td>
<td>F</td>
<td>15.454</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Urban population</td>
<td>0.001</td>
<td>Z</td>
<td>1.340</td>
<td>0.180</td>
<td>0.339</td>
<td>F</td>
<td>9.109</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Potential Risk Adjuster</td>
<td>Effect Size</td>
<td>Test</td>
<td>Statistic</td>
<td>p-value</td>
<td>Adjusted p-value</td>
<td>Test</td>
<td>Statistic</td>
<td>p-value</td>
<td>Adjusted p-value</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------</td>
<td>------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------------</td>
<td>------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Unemployed</td>
<td>−0.017</td>
<td>Z</td>
<td>−1.333</td>
<td>0.183</td>
<td>0.339</td>
<td>F</td>
<td>21.400</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Health insurance insured</td>
<td>0.007</td>
<td>Z</td>
<td>1.075</td>
<td>0.282</td>
<td>0.432</td>
<td>F</td>
<td>21.273</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Race White</td>
<td>0.002</td>
<td>Z</td>
<td>0.692</td>
<td>0.489</td>
<td>0.707</td>
<td>F</td>
<td>21.027</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Employed</td>
<td>0.002</td>
<td>Z</td>
<td>0.563</td>
<td>0.574</td>
<td>0.732</td>
<td>F</td>
<td>19.242</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Labor force participation</td>
<td>0.002</td>
<td>Z</td>
<td>0.504</td>
<td>0.614</td>
<td>0.732</td>
<td>F</td>
<td>17.231</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Race Hispanic</td>
<td>0.001</td>
<td>Z</td>
<td>0.393</td>
<td>0.694</td>
<td>0.732</td>
<td>F</td>
<td>71.635</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Race Asian</td>
<td>0.004</td>
<td>Z</td>
<td>0.384</td>
<td>0.701</td>
<td>0.732</td>
<td>F</td>
<td>20.859</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Race Black</td>
<td>−0.001</td>
<td>Z</td>
<td>−0.220</td>
<td>0.826</td>
<td>0.826</td>
<td>F</td>
<td>23.627</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Table 4.3. Feeling Heard and Understood Performance Measure Score by Survey Mode

<table>
<thead>
<tr>
<th>Survey Mode</th>
<th>n(^a)</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mail</td>
<td>1,268</td>
<td>0.73</td>
<td>0.37</td>
</tr>
<tr>
<td>Phone</td>
<td>979</td>
<td>0.67</td>
<td>0.37</td>
</tr>
<tr>
<td>Web</td>
<td>525</td>
<td>0.75</td>
<td>0.37</td>
</tr>
</tbody>
</table>

\(^a\) For the 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 performance measure.

Diagnosis Analysis

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 (Table 4.4) 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, underscoring the importance of further research in this area before the measure is used for high-stakes decisions.

Table 4.4. Mean Score by Diagnosis Grouping

<table>
<thead>
<tr>
<th>Diagnosis Grouping</th>
<th>Mean</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer (solid and liquid)</td>
<td>0.720</td>
<td>1,670</td>
</tr>
<tr>
<td>End-organ disease (non-neurological)</td>
<td>0.648</td>
<td>223</td>
</tr>
<tr>
<td>Other</td>
<td>0.735</td>
<td>597</td>
</tr>
</tbody>
</table>

ANOVA p-value: \(p < 0.01\)

Performance Measure Exclusions

As noted earlier in this chapter, we did not have survey data to analyze the impact on measure outcomes of necessarily excluding from the denominator (1) respondents who did not answer any of the data elements comprising the multi–data element measure, (2) nonrespondents, (3) program disavowals, and (4) decedents. We did compare demographic characteristics of nonrespondents to respondents (see Chapter 2) and found that respondents were older and consisted of more White and fewer Black patients when compared with nonrespondents (though race data in our sample were limited). Because age and race may differentially affect patient experiences of care, it should be noted that this is not reflected in measure performance because
of nonresponse. 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 the scope of this testing effort.

Our exclusion analysis primarily focused on exploring the impact of proxy-involved survey data. We had a total of 2,548 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 4.5 shows the mean (standard deviation) *Feeling Heard and Understood* performance measure score according to these three groups.

### Table 4.5. Mean *Feeling Heard and Understood* Performance Measure Score According to Proxy Group

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Mean (Standard Deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient only</td>
<td>2,548</td>
<td>0.71 (0.37)</td>
</tr>
<tr>
<td>Proxy assisted</td>
<td>224</td>
<td>0.77 (0.34)</td>
</tr>
<tr>
<td>Proxy only</td>
<td>430</td>
<td>0.69 (0.37)</td>
</tr>
</tbody>
</table>

A one-way ANOVA test for differences among these three means was significant ($F(2, 3199) = 3.80, p = 0.023$), 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* performance measure score means between the proxy-only and patient-only groups, we decided for conceptual reasons to exclude surveys that were completed solely by a proxy with no patient involvement after discussing these results with our project advisory group. Because this is 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 project advisory group. 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 caused by 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.

**Performance Measure Scoring**

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
an indicator of proxy assistance. Table 4.6 provides a summary of regression coefficients of the fixed effects of the adjustment model (i.e., ignoring the $b_i$ terms). This table indicates that 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>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept: $\beta_0$</td>
<td>0.97</td>
<td>0.08</td>
<td>0.82</td>
<td>1.13</td>
</tr>
<tr>
<td>Survey mode: phone</td>
<td>-0.25</td>
<td>0.05</td>
<td>-0.35</td>
<td>-0.15</td>
</tr>
<tr>
<td>Survey mode: web</td>
<td>0.23</td>
<td>0.07</td>
<td>0.10</td>
<td>0.36</td>
</tr>
<tr>
<td>Proxy status: assisted</td>
<td>0.23</td>
<td>0.09</td>
<td>0.09</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Our final model can be used to assess the distribution for the estimated (i.e., adjusted) program scores shown in Figure 4.3. This results in reduced variability (standard deviation = 7.04) in program performance of adjusted average scores. However, the performance measure still demonstrates variability across programs. The next sections discuss our ability to discern among programs, given this level of observed variability in scores.

Figure 4.3. Adjusted Program Performance

Statistically Significant and Meaningful Differences in Performance

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 4.4 shows the adjusted measure scores for each program, with the dot indicating the mean score and the lines extending to the 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 preliminary evidence that there are significant differences in program performance on the measure.

Figure 4.4. Posterior Distribution of Adjusted Program Scores

Another way to evaluate the performance measure’s ability to distinguish among programs is to compare the fit of nested models for which the difference in models is the inclusion of a program-level effect. Results of this analysis showed that the model that included a program-level indicator was significantly different from the model without one ($\chi^2_{(43)} = 60.04, p < 0.05$), demonstrating that there are differences among program scores from the “grand” mean score.

Considering the score and rank order together, Table 4.7 displays the magnitude of change in the Feeling Heard and Understood program-level measure scores associated with selected program rank differences. The program scores in our sample range from 54.05 to 85.18, and the practical interpretation of differences between program score and/or rank order varies depending on the location in the distribution and the distance between programs of interest. For example,
assuming 100 programs participated in the scoring, the difference in measure scores between programs that are ranked fifth and tenth overall is a relatively small value of 2.05 in the Feeling Heard and Understood performance measure score. The difference in measure scores between programs that are ranked tenth and 20th is 2.83 points. On the other hand, a program at the median (in the middle of the ranking) will need a large increase in measure score of 4.19 points to improve to the 20th ranking. Similarly, a program that is very low in the ranking (at the 20th or tenth ranking from the bottom) will need a large 7.28 points or 10.90 points increase, respectively, to improve at least to the middle of the ranking (median). These changes in score by program rank also suggest that measure performance is actionable (i.e., there is room for programs to improve their score).

Table 4.7. Change in Measure Score, Assuming 100 Ranked Programs

<table>
<thead>
<tr>
<th>Program Rank Difference</th>
<th>Change in Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th top to 10th top</td>
<td>2.05 points drop</td>
</tr>
<tr>
<td>10th top to 20th top</td>
<td>2.83 points drop</td>
</tr>
<tr>
<td>Median to 20th top</td>
<td>4.19 points increase</td>
</tr>
<tr>
<td>20th lowest to median</td>
<td>7.28 points increase</td>
</tr>
<tr>
<td>10th lowest to median</td>
<td>10.90 points increase</td>
</tr>
</tbody>
</table>

Performance Measure Score Reliability

As described earlier in this chapter, 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 percent confidence interval: 0.027 to 0.089; Figure 4.5). This suggests there is a reasonable level of reliability based on the observed between-program variability and the within-program variability, given our sample sizes.
We then extend our reliability results to future samples using the Spearman-Brown prophecy formula, which estimates the average sample size requirement (i.e., number of patient respondents) within programs to achieve a desired reliability for a given ICC. This is visualized in Figure 4.6 for our posterior distribution of the ICC estimates in Figure 4.5. In Figure 4.6, we estimate that in order to obtain a nominal reliability of 0.7, an average sample size of 45 would be required (orange horizontal line in the plot). However, because each participant is responding to four questions, and those questions are highly correlated (estimated design effect of 3.25 from repeated measures), an average sample size of 37 participants responding to the four data elements would be required when adjusted for this design effect.

NOTE: CI = confidence interval.
Additionally, we compute estimates of individual program-specific reliability using a method similar to the approach used in Adams, 2009. To gain consistency between the approach in Adams and our models, we use our models to estimate a posterior distribution for the overall variability of the risk-adjusted program scores (i.e., the variance of the distribution of \( \logit^{-1}(\beta_0 + b_i) \) pooling across all \( i \)) and estimate a posterior distribution of the variance of each within-program score (i.e., \( \hat{\rho}_i(1 - \hat{\rho}_i)/n_i \), as specified in Adams, 2009). This is presented in Figure 4.7, and the average reliability across programs was approximately \( r = 0.752 \).

**Figure 4.7. Distribution of Program-Specific Reliability**

![Histogram of Reliability](image)

Average Reliability = 0.752

NOTE: The outlying program contained only a small number of responses.

**Impact of Small Programs**

Although Figure 4.7 suggests that a low number of responses within programs would be necessary to estimate a reliable *Feeling Heard and Understood* performance measure score, we sought to assess sensitivity to small programs (e.g., those with low patient volume) in these estimates. Figure 4.8 shows a scatterplot of the unadjusted (x-axis) against the adjusted (y-axis) program scores. The impact of within-program sample size on the risk-adjusted program score is conveyed in this figure by the program’s proximity to the diagonal—i.e., programs far from the diagonal are affected more by our mixed-effects model. Similarly, the right panel is meant to convey the impact of smaller programs; the small programs \( (n = 13) \) with fewer than 15 responses are highlighted. Notice that smaller programs represent a high number of programs in the tails of the distribution. This suggests that a large portion of the observed variability potentially is being driven by small programs, and therefore our high reliability estimates could primarily be driven by smaller programs.
Thus, we removed all programs with less than 15 responses and recalculated the posterior distribution of the ICC (posterior mean of 0.035, 95 percent confidence interval [0.016, 0.066]) and the corresponding sample size for a nominal reliability of 0.7 (Figures 4.9 and 4.10).
We see that there is some sensitivity to small programs. A more conservative estimate of the ICC might be closer to 0.035, meaning that the performance measure would require a higher average sample size from programs of approximately 70; after the design effect adjustment, this results in an average sample size of 57 participants responding to the four data elements.

**Performance Measure Score Validity**

Results of validity testing at the program level provide evidence supporting the use of the *Feeling Heard and Understood* performance measure as constructed. As hypothesized, the *Feeling Heard and Understood* performance measure was significantly and positively associated with the CAHPS communication measure ($r = 0.635, p = 0.011$), the *Receiving Desired Help for Pain* performance measure ($r = 0.496, p < 0.001$), and the overall rating of the palliative care provider and team ($r = 0.768, p < 0.001$). Taken together, these results provide support for the convergent validity of the *Feeling Heard and Understood* performance measure.

**Face Validity**

Seven expert advisers rated face validity of the *Feeling Heard and Understood* performance measure score. On average, advisers rated face validity of the performance measure score as 8.3 on a scale of 1 to 9, corresponding with an average rating of “high.” These ratings reflect strong support for face validity of the proposed performance measure from experts in palliative care and quality measurement.

**Threats to Validity Caused by the COVID-19 Pandemic**

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 test field 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 ($\alpha = 0.05$ 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 (Table 4.8). 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 ($p = 0.21$; Table 4.9).
Table 4.8. Pre-Pandemic and Mid-Pandemic Measure Score Distributions

<table>
<thead>
<tr>
<th>Score</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-pandemic (rounds 1–4)</td>
<td>102 (13%)</td>
<td>67 (9%)</td>
<td>77 (10%)</td>
<td>107 (14%)</td>
<td>424 (55%)</td>
</tr>
<tr>
<td>Mid-pandemic (rounds 7–10)</td>
<td>198 (15%)</td>
<td>111 (8%)</td>
<td>113 (9%)</td>
<td>235 (18%)</td>
<td>660 (50%)</td>
</tr>
</tbody>
</table>

NOTE: For simplicity, measure scores reported in this table are rounded to the nearest quarter.

Table 4.9. Generalized Linear Model Results for the Feeling Heard and Understood Performance Measure

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.94</td>
<td>0.07</td>
<td>14.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Flag: mid-pandemic (rounds 7–10)</td>
<td>−0.10</td>
<td>0.08</td>
<td>−1.27</td>
<td>0.21</td>
</tr>
</tbody>
</table>

NOTE: SE = standard error.

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.
5. Feasibility and Implementation Considerations

Overview

We assessed the initial feasibility of implementing these measures into practice. During the national beta field test, we (1) interviewed multiple survey vendors to understand costs and burden to programs; (2) interviewed participating palliative care programs to understand their perceptions of value and utility of the proposed measures, whether and how they would use these measures in practice, and anticipated barriers and facilitators to implementation; and (3) interviewed patients from racial and ethnic minority groups to better understand potential variability in their experiences of palliative care that may not be fully reflected in the survey data. Findings from each of these sources support the value of these measures and the feasibility of implementation into practice. Next, we describe each of these data collection activities in further detail.

Cost and Burden Assessment

We conducted telephone interviews with five CMS-approved survey vendors to better understand the financial burden to programs to hire a vendor to administer the patient surveys and the analytic capabilities of vendors to potentially calculate measure scores. We focused on the process used during testing, involving the identification of all eligible visits during a three-month or quarterly time frame and a subsequent three-month survey administration and data collection time frame, with data from all participating programs aggregated over a 12-month, or calendar year, reporting period (Figure 5.1).

Figure 5.1. Example Data Collection and Reporting Schedule for Measure Performance Year 2022
In addition, we explored the costs of the enhanced mixed-mode survey administration protocol used during testing: web to mail to telephone over an eight-week active field period. According to quotes from five CMS-approved survey vendors, the estimated costs to administer the mixed-mode patient surveys varied from approximately $2,500 to $12,500 per year. We requested quotes from each vendor for the estimated cost to administer surveys on a monthly rolling basis to 200 eligible patients per month using enhanced mixed-mode survey administration (web to mail with telephone follow-up, up to three phone attempts to nonresponders) and to report benchmarking data to the client on a quarterly basis. Larger vendors reported higher cost estimates, including higher operational setup and maintenance costs. Vendors noted that costs vary depending on the volume of eligible patients and the data analytics requested. Vendors also noted that telephone is the most expensive survey mode, and costs are dependent on the number of follow-up calls made. In our testing protocol, up to eight follow-up telephone calls were budgeted, though the average number of calls made was five. Among the patients with complete responses by CATI \((n = 980)\), the median number of calls made was two, and the mean was 2.74.

To put the above cost estimates in context, the cost to programs for administering the 47-item CAHPS Hospice Survey through a vendor, in a mail-only administration, is estimated to be approximately $4,000 per year (information drawn from Office of Management and Budget submissions). Although we anticipate only 12 to 16 survey data elements to be necessary for data collection on the *Feeling Heard and Understood* performance measure, the mixed-mode protocol will likely contribute to the overall cost.

Most programs that participated in the national beta field test had experience working with a survey vendor to administer patient experience surveys outside of this project. RAND SRG served as the survey vendor for the national beta field test. Important factors cited in the decision to invest in support from a survey vendor included cost efficiency and tracking issues (e.g., vendors being better at tracking cases in real time and not sending surveys to deceased patients) and a vendor’s ability to compare measure performance with other programs and update patient contact information.

The national beta testing protocol demonstrated that a survey vendor has a clear role in minimizing burden to programs, minimizing bias and conflict in the data collection process, and implementing analytic steps that programs may not have the capacity to conduct. For example, the survey vendor would apply eligibility criteria and algorithms to identify the appropriate patient sample and field and track all surveys, which would minimize gaming and social desirability bias. The vendor could also directly extract patient visit data from programs with electronic health records, especially in systems that already have a vendor engaged for other patient experience surveys (e.g., CAHPS surveys), minimizing staff burden and capacity. Finally, the survey vendor would take on the responsibility of cleaning all data and submitting files to a third party for analysis, scoring, and submission to CMS. These benefits should be weighed against the likely costs to programs. In cases in which a vendor is already engaged by a
program or health system, that survey vendor is likely to be the most cost-effective approach to measure implementation.

**Program Perspectives**

We conducted interviews with individuals from 25 palliative care programs that participated in the national beta field test to understand (1) their perceptions of the value and use of the proposed measure and how the performance measure might be used to facilitate quality improvement, (2) the perceived burden of measure implementation and anticipated challenges or unintended consequences, and (3) the impact of the pandemic on the program’s use of the performance measure and overall care delivery.

**Approach**

Between May and July 2020, we contacted 34 programs for an interview, seeking a diverse range of geographic region, type, and approximate size to elicit a variety of perspectives. An initial invitation email was followed by up to two additional emails and one phone call. Two programs declined participation, and six programs did not respond to our attempts to contact them. One program was scheduled to participate but canceled.

We interviewed a total of 47 individuals across 25 programs (see Table 5.1 for a breakdown of programs by region, type, and approximate size). We asked programs to include a palliative care provider (e.g., a medical director or program manager) in the interview and encouraged attendance by a data specialist so that we could explore the impact of measure implementation on clinical processes and the program’s data needs. Interviewees included such providers as medical directors of palliative care programs, nurse practitioners, registered nurses, and social workers who were involved in leadership in the palliative care programs. Interviewees also included nonclinicians, such as palliative care program managers, quality and compliance directors, data managers, and data analysts.
Table 5.1. Breakdown of Programs That Participated in the Program Interviews (n = 25)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number of Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>6</td>
</tr>
<tr>
<td>South</td>
<td>8</td>
</tr>
<tr>
<td>Midwest</td>
<td>5</td>
</tr>
<tr>
<td>Northeast</td>
<td>6</td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Hospital</td>
<td>12</td>
</tr>
<tr>
<td>Hospice</td>
<td>7</td>
</tr>
<tr>
<td>Outpatient</td>
<td>6</td>
</tr>
<tr>
<td>Approximate Size</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>6</td>
</tr>
<tr>
<td>Medium</td>
<td>10</td>
</tr>
<tr>
<td>Large</td>
<td>9</td>
</tr>
</tbody>
</table>

Interviews were guided by a semistructured protocol focused on three broad areas: (1) potential implementation of the proposed performance measure, including barriers and facilitators; (2) value, usability, utility, and potential unintended consequences of the proposed performance measure; and (3) implications of the COVID-19 pandemic on the use of the performance measure and participation in the test and on programs’ outpatient palliative care practices in general. Interviews were conducted through Microsoft Teams. Most interviews were recorded and transcribed, and the length of interviews ranged from 24 to 63 minutes. One respondent declined recording; a trained notetaker took notes instead. Interviews were analyzed using a rapid thematic analysis to facilitate efficient and timely delivery of findings to the project team (Taylor et al., 2018; Gale et al., 2019).

**Key Findings**

**Value of the Feeling Heard and Understood Performance Measure**

Overall, interviewees from programs that participated in the interviews during the national beta field test provided positive feedback about the *Feeling Heard and Understood* performance measure, noting its central importance to palliative care and the need to measure this concept. Interviewees mentioned the value of this performance measure to capture patient experience, stating that this performance measure encapsulates “the philosophy and soul of palliative care” and that “it measures something that is essential to the service we provide.” They stated that it is critical for patients to feel heard and understood in palliative care in order to establish relationships and build the trust necessary to have important and challenging conversations. Interviewees also talked about the usefulness and appropriateness of this type of 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.

Concerns about the performance measure raised by interviewees were related to patient experience surveys in general, including the issue of patient recall bias and the inability of patients to distinguish palliative care providers from their other providers. Specific to palliative care, there were some concerns that the act of delivering bad news—although a common task of palliative care—may unduly and negatively affect a patient’s experience of communication with their provider and team. Some interviewees also noted practical issues, including “survey fatigue,” especially for patients who are seriously ill, as well as the need for survey phrasing to be understandable or “resonate” for a broad range of patient populations, including those with low literacy levels.

Perceptions of Performance Measurement and Performance Improvement

Many interviewees described quality assessment and improvement efforts that they would engage in to better understand potential performance gaps in their palliative care program, including taking a close look at what aspects of their care might affect patients’ experiences of feeling heard and understood. Some interviewees also discussed examining potential external factors (e.g., scheduling, delays in appointments, miscommunication around referrals) aside from the clinical encounter that might also play a role in patients’ experiences of feeling heard and understood.

Impact of COVID-19 Pandemic on Performance Measure Testing

All programs that were interviewed had to convert to telehealth because of the pandemic, which, for most, shifted their typical operations and processes. All but three programs had no or very little prior experience using telehealth for their outpatient palliative care practice prior to the pandemic. Some participants noted that they wanted to include telehealth as part of their outpatient palliative care services prior to the pandemic. The programs with no or little experience with telehealth rapidly implemented telehealth services during the pandemic, some over the course of a weekend or a few days to be able to offer services to patients quickly. A couple of programs briefly suspended their outpatient clinic to plan for telehealth.

Interviewees described benefits and challenges with telehealth use during the pandemic. Benefits included the opportunity for providers to see patients through video in their homes, improved medication reconciliation because patients had all medications immediately at hand, and better insight into the patients’ home environment and needs. Telehealth also facilitated the opportunity to integrate family members into important conversations about goals and values for care. Importantly, access and the ability to see patients who are frail, symptomatic, live in rural areas, or are socially isolated were often raised as benefits of telehealth. Challenges mentioned included a loss of intimacy or human touch and meaningful connection, though respondents also shared that they overcame these challenges through various communication strategies, such as
emphasizing their body language to validate patients’ concerns. Other challenges were the inability to conduct a full physical exam, technical difficulties, and challenges in communicating bad news to patients. A few programs faced challenges related to electronically prescribing opioids and concerns over the inability to do drug tests or pill counts.

**Experiences of Patients from Racial and Ethnic Minorities**

The survey data collected during the national beta field test largely reflect the experiences of a non-Hispanic White patient population (Table 2.4), despite our geographically diverse sample of palliative care programs. 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). To better understand these factors and to capture the experiences of racial and ethnic minority patients with ambulatory palliative care and with feeling heard and understood, 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.

**Approach**

We recruited patient and family caregivers for interviews through our project newsletter (sent every two months) to programs in the national beta field test, referrals from TECUPP members, and collaboration with national organizations, including the National Hospice and Palliative Care Organization, the National Patient Advocate Foundation, and the Cancer Support Community, which advertised our effort through their internal channels. We specifically sought the participation of patients or family caregivers of those who identified as Black, Latinx, Hispanic, Asian, or another racial or ethnic minority group who had a serious medical illness and who had experienced palliative care (either from a palliative care provider or from another provider) or hospice care within the past year.

Interviews were guided by a semistructured protocol, with questions about the type of care a respondent or their family member had received for a serious illness, who provided that care, and the experience they (or their loved one) had with this care. One question inquired about whether respondents felt heard and understood, and respondents were asked to comment on how their experience might differ from patients of a different race or ethnicity and whether provider and patient racial concordance affected their care. Interviews were conducted primarily by phone from October 2020 to January 2021. One interview was conducted by video. Interviews were conducted by two White-identifying researchers and one Black-identifying researcher. It is
possible that patient and family caregiver responses may have differed depending on racial concordance with the interviewer. All interviews were audio recorded and transcribed by a member of the research team. Interviews ranged from 35 to 75 minutes. Interviews were analyzed using a thematic analysis approach (Ryan and Bernard, 2003).

**Key Findings**

**Participating Program Patient Population Information**

Most programs that participated in the program interviews reported that the majority of patients in their outpatient 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 the 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.

**Patient and Family Caregiver Interviews**

Thirty individuals were referred or directly reached out to the research team. Of these, seven individuals were unable to be screened for an interview because we were unable to contact them for a screening call to determine whether they were eligible to participate. Ten individuals were not eligible to participate in the interview because they (or a loved one) did not receive palliative, supportive, or hospice care for a serious illness, despite responding to our recruitment call, because of a lack of awareness around palliative, supportive, and hospice care. We conducted interviews with nine patients and four family members who reported that they (or a family member) had received palliative, supportive, or hospice care in the past year. Twelve of the 13 interviewees identified themselves (or the patient) as Black or African American, and one identified herself as Asian.

**Experiences of Access and Referral to Care**

One of the most significant challenges we encountered in the screening process to confirm eligibility for participation in the interview was a lack of awareness of palliative care. We learned that several individuals who expressed interest in participating in the interview had never heard of palliative care despite living with a serious medical condition (i.e., conditions with symptoms that are often treated by palliative care providers). This lack of awareness of palliative
care was also a common theme in our interviews with patients and family caregivers; many interviewees described not being familiar with palliative care until their referral or shared that they understood palliative care to be only for those at the end of life.

Experiences of Feeling Heard and Understood

Patients and family caregivers were generally positive about their experiences of feeling heard and understood by their palliative care provider or their provider for supportive or hospice care. During the interviews, we asked patients and family members the same *Feeling Heard and Understood* data element included in the national beta field test survey instrument: “Thinking about your overall experience with this provider and team in the last six months, how true is the following statement: I felt heard and understood by this provider and team.” Most patients and family caregivers responded with the “very true” or “completely true” response option to this question. Patients said they felt that they were listened to, not rushed in their appointments or interactions with their providers, and able to trust their provider and ask questions.

Perceptions of Differences in Care Because of Race

Patients and family caregivers provided mixed responses about their perceptions of the impact of race on the care they received. Although some respondents said that they experienced discrimination or racism in the health care system in general, they felt that they were treated the same as other patients by their palliative care provider. Patients and family caregivers also provided mixed responses as to whether they felt the racial and ethnic concordance of their provider was important to them (i.e., whether their provider was of the same race and ethnicity). They reported high satisfaction with palliative care and therefore felt that provider-patient racial concordance could only minimally improve their palliative care experience. Others felt that their educational and social status was a more significant factor in the care they received than their race.
6. Performance Measure Implementation and Future Considerations

After establishing the reliability and validity of the *Feeling Heard and Understood* performance measure, a key next step is to consider the implementation of this performance measure into clinical practice, quality improvement efforts, and quality payment programs. As part of our examination of the testing and analytic procedures used during the national beta field test, we identified several considerations for programs planning to implement this measure for these purposes.

**Considerations for Program Planning**

*Pediatric Ambulatory Palliative Care*

We developed and tested the *Feeling Heard and Understood* performance measure using an adult (i.e., 18 years of age or older) ambulatory population receiving palliative care services. This performance measure has not been tested in pediatric populations. We weighed the input and concerns of key pediatric palliative care stakeholders at the outset of our performance measure development effort and acknowledge that systematic measurement of quality, and patient and family experience, is essential to improving pediatric palliative care and supporting greater use of these services. Pediatric palliative care performance measurement entails distinct considerations that we were unable to address within the scope of this effort, including identifying appropriate self-reported measures of experience for pediatric patients, evaluating the experiences of parents caring for pediatric patients and ensuring their voice is reflected in performance measurement efforts, and testing the performance measures specifically among a pediatric population. We strongly agree that as palliative services become more accessible to children with life-limiting illnesses and their family members, quality measurement is critical and should be included in CMS’s measurement development priorities.

*Telehealth Visits*

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. 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.

**Small Programs**

Performance measure reliability tends to be lower in smaller programs. The estimate of the ICC may be closer to 0.035 rather than 0.052, and thus a reliable performance measure would require a higher average sample size of approximately 70 from programs (i.e., at least 18 responders each answering the four data elements). This may mean that some smaller ambulatory palliative care programs, which may not have enough volume for a reliable measure to justify the cost of fielding the survey and participating in the measure program, will opt not to report this performance measure (and should not be required to).

**Data Collection Procedures**

Through the process of fielding our survey during the national beta test, we identified various considerations for measure data collection. Although additional detail will be provided in a forthcoming Measure Implementation Guide, we note a few important points in the following sections.

**Survey Mode**

First, as our results show, there were differences in performance measure outcomes by survey mode. Specifically, scores for *Feeling Heard and Understood* were lower when collected by telephone compared with either web or mail surveys. Although we typically would consider the potential for an interviewer effect, we would expect such an effect to be in the positive direction: i.e., we would expect scores to be higher when collected by telephone because of a reluctance by the respondent to provide negative feedback when speaking directly to someone (Bowling, 2005; de Vries et al., 2005). However, some work in CAHPS populations, particularly the CAHPS Hospice survey, has demonstrated that bereaved primary caregiver respondents tend to respond more negatively by telephone than other modes (Parast, Elliott, et al., 2018). We found similar effects in our palliative care patient population, and we consider the lower phone results to be a timing effect. Telephone interviews were conducted only if a survey had not been returned by web (the first option available to respondents) or by mail (the second option); thus, telephone responders represent patients who may have been particularly disinclined to respond to the survey—e.g., because of negative experiences. Although it was beyond the scope of this
effort, a future mode experiment would further elucidate survey mode effects and provide guidance for optimal survey modes for performance measure data collection.

Survey Field Period

For the visit and experience to remain salient to the patient and to ensure the successful implementation of the performance measure, patients needed to receive and return the survey within six months of their visit (i.e., no later than six months after their eligible visit). Because programs sent data files at varying times and that included visits over several months, we needed to determine a method to ensure consistency in survey fielding, such that all patients had the same amount of time to complete and return the survey and no patient would report on a visit that occurred more than six months ago.

From the data files that programs submitted, we identified eligible visits that occurred within three months of the anticipated survey fielding start date (Figure 6.1). For example, if we scheduled survey fielding to begin November 1, we included all eligible visits from August 1 through October 30. Active data collection occurred over a period of eight weeks, from the date the prenotification letter was mailed (survey fielding start date) through the maximum number of attempts to reach the patient by telephone. We allowed a maximum of three months from the survey fielding start date for surveys to be returned (e.g., with a survey fielding start date of November 1, we would accept returned surveys through February 1).

Figure 6.1. Survey Field Period

Although we allowed a six-month window from earliest eligible visit to latest survey return, 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. However, programs (as well as survey vendors contracted to support survey fielding) may weigh the diminishing yield at five to six months after the earliest eligible visit against the costs required to keep a survey field open for the whole six-month window.
Survey Language

Although we developed a Spanish-language survey instrument as part of this performance measure development effort, we tested only the English-language version during the beta field test. All eligible patients should be sent the survey. In the event that the patient respondent needs proxy assistance with completing the survey because of language difficulties, the survey data can still be retained. Future work to evaluate the reliability and validity of the performance measure using Spanish-language survey data is warranted and probably would increase the use of the performance measure in clinical practice.

Data Aggregation

To evaluate performance measure reliability and validity, we aggregated data from all 44 participating programs, collected across the field test period. Patients who received the survey once during the field test period were considered ineligible for subsequent fielding within the field test period. This approach highlights several considerations for implementing the performance measure in practice. First, palliative care providers and provider groups should aggregate survey data across a reporting period (e.g., the 12-month MIPS reporting period). To do so, they should also consider all adult ambulatory palliative care patients who received care during that reporting period as initially eligible to be invited to complete the patient experience survey related to their ambulatory palliative care visit unless they meet further exclusion criteria. Patients should be invited to complete the patient experience survey only once per reporting period.

Usability of the Performance Measure

Findings from the national beta field test provide support for the usability of this performance measure for quality improvement and in quality payment programs. The proposed performance measure is being developed under MACRA legislation for use in CMS’s MIPS program and thus could be implemented into at least one accountability application and to inform performance-based payment decisions. The performance measure might also be used for public reporting programs, improved transparency, and consumer choice. To facilitate performance measure adoption and use, the measure might be combined with similar patient experience measures and survey efforts, such as CAHPS or Press-Ganey surveys, though we did not test this application. In combination with the ongoing robust stakeholder engagement efforts led by the NCHPC and supported by the AAHPM, these likely applications for the performance measure ensure routine use and usability once implemented.

To support its use in practice and for the purpose of quality improvement among palliative care programs, we have continuously engaged clinicians, health systems, patients, and caregivers in the development of this performance measure and offered guidance on how to use this performance measure in practice to drive systematic quality measurement and improvement in
priority areas for palliative care. Further, performance results and interpretive guidance were provided to all sites that participated in the national beta field test, giving them a first look at how the performance measure might be useful to their quality improvement efforts. According to the multidisciplinary stakeholder feedback we received throughout this effort, including formal public comment, we believe this performance measure can be successfully implemented and can provide valuable information to guide palliative care improvement in patient experience.

Conclusion

Ambulatory palliative care provided in clinical settings represents a growing proportion of all palliative care services today. It facilitates the integration of treatment-focused care with palliative and supportive interventions, as well as access to palliative care early in the disease trajectory. Systematically evaluating the quality and experience of ambulatory palliative care and incentivizing access and delivery are critical to ensuring high-quality, timely care for patients with serious and advanced illness and their family members.

Feeling heard and understood is a core outcome of palliative care and reflects patient-centered processes of communication, trust, interpersonal interaction, and holistic care, all of which are essential tenets of palliative care. Findings from our study demonstrate the reliability and validity of a Feeling Heard and Understood scale that can be used to measure the quality of ambulatory palliative care, guide improvement efforts, and distinguish providers on the basis of their performance. We also highlight three important areas for future research: (1) systematic examination of survey administration modes, which may have an impact on response rates and the representativeness of respondents; (2) evaluation of the implementation of this performance measure into clinical practice and regular use; and (3) exploration and analysis of disparities in ambulatory palliative care access, utilization, and experience, including (but not limited to) racial, gender, ethnic, cultural, and language disparities.

The implementation of this performance measure into system-level quality improvement programs, as well as into quality payment programs (such as MIPS), could further incentivize and reward high-quality ambulatory palliative care, facilitating its continued growth and use by patients facing serious and advanced illness.
Appendix A. National Beta Field Test Survey Instrument

**PATIENT EXPERIENCE SURVEY**

**SURVEY INSTRUCTIONS**

- This survey should be completed by the patient indicated on the survey cover letter.
- You can ask a family member or friend for help with this survey or ask them to complete the survey for you.
- If you are a family member or friend helping with this survey or completing this survey for the patient indicated on the survey cover letter, please remember that all survey questions ask about the patient’s experiences. Unless a question says otherwise, please do not consider your own experiences or information in the answers you provide.
- Use a dark colored pen to fill out the survey.
- Place an X directly inside the square indicating a response, like in the sample below.

  ![Yes](image)

  ![No](image)

- This survey uses the word “provider” throughout. When we say “provider,” we mean a medical provider like a doctor or a nurse.

---

Please return the completed survey in the provided pre-paid envelope to:

RAND Corporation  
Attn: Ryan McKay – M2IV  
1776 Main Street  
Santa Monica, CA  90401

---

58
YOUR PROVIDER AND TEAM

1. Our records show that you got care from the provider and team named below in the last 6 months.
   [Provider] and team
   Is that right?
   □ Yes
   □ No → If No, go to Question 23

The questions in this survey will refer to the provider named in Question 1 as “this provider and team.” Please think of this provider and team as you answer the survey.

YOUR CARE FROM THIS PROVIDER AND TEAM IN THE LAST 6 MONTHS

The following questions ask about the care you have received from this provider and team in the last 6 months.

2. In the last 6 months, how often have you seen this provider and team for your care?
   □ I have seen this provider and team many times for my care
   □ I have seen this provider and team a few times for my care
   □ I have seen this provider and team only one time for my care

3. In the last 6 months, how often did this provider and team explain things in a way that was easy to understand?
   □ Never
   □ Sometimes
   □ Usually
   □ Always

4. In the last 6 months, how often did this provider and team listen carefully to you?
   □ Never
   □ Sometimes
   □ Usually
   □ Always

5. In the last 6 months, how often did this provider and team show respect for what you had to say?
   □ Never
   □ Sometimes
   □ Usually
   □ Always

6. In the last 6 months, how often did this provider and team spend enough time with you?
   □ Never
   □ Sometimes
   □ Usually
   □ Always

7. In the last 6 months, have you ever had pain?
   □ Yes
   □ No → If No, go to Question 10

8. In the last 6 months, did you want help from this provider and team for this pain?
   □ Yes
   □ No → If No, go to Question 10

9. 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
10. In the last 6 months, did you want emotional support from this provider and team?
   □ Yes
   □ No → If No, go to Question 12

11. In the last 6 months, did you get as much emotional support as you wanted from this provider and team?
   □ Yes, definitely
   □ Yes, somewhat
   □ No

YOUR OVERALL EXPERIENCE WITH THIS PROVIDER AND TEAM

12. Thinking about your overall experience with this provider and team in the last 6 months, how true are the following statements?

   I felt heard and understood by this provider and team.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

13. I trusted this provider and team.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

14. I felt comfortable asking this provider and team questions.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

15. I could tell this provider and team anything, even things I might not tell anyone else.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

16. I felt this provider and team put my best interests first when making recommendations about my care.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

17. I felt this provider and team always told me the truth about my health, even if there was bad news.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true
18. I felt this provider and team saw me as a person, not just someone with a medical problem.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

19. I felt this provider and team knew what worried me most about my health.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

20. I felt this provider and team understood what is important to me in my life.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

21. I felt this provider and team would know what I would want done if I was unconscious or in a coma.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true

22. Using any number from 0 to 10, where 0 is the worst provider and team possible and 10 is the best provider and team possible, what number would you use to rate this provider and team? Mark one number.
   □ 0 Worst provider and team possible
   □ 1
   □ 2
   □ 3
   □ 4
   □ 5
   □ 6
   □ 7
   □ 8
   □ 9
   □ 10 Best provider and team possible

23. In general, how would you rate your physical health?
   □ Excellent
   □ Very good
   □ Good
   □ Fair
   □ Poor

24. In general, how would you rate your mental or emotional health?
   □ Excellent
   □ Very good
   □ Good
   □ Fair
   □ Poor
### YOUR HEALTH OVER THE LAST 2 WEEKS

25. Compared to 3 months ago, is your overall health now:
   - [ ] Much better
   - [ ] Somewhat better
   - [ ] About the same
   - [ ] Somewhat worse
   - [ ] Much worse

26. In general, how much does pain interfere with your day to day activities?
   - [ ] Very much
   - [ ] Quite a bit
   - [ ] Somewhat
   - [ ] A little bit
   - [ ] Not at all

27. Over the last 2 weeks, how often have you been bothered by having little interest or pleasure in doing things?
   - [ ] Not at all
   - [ ] Several days
   - [ ] More than half the days
   - [ ] Nearly every day

28. Over the last 2 weeks, how often have you been bothered by feeling down, depressed, or hopeless?
   - [ ] Not at all
   - [ ] Several days
   - [ ] More than half the days
   - [ ] Nearly every day

29. Over the last 2 weeks, how often has your thinking been slow?
   - [ ] Never
   - [ ] Rarely
   - [ ] Sometimes
   - [ ] Often
   - [ ] Very often

30. Over the last 2 weeks, how often has it seemed like your brain was not working as well as usual?
   - [ ] Never
   - [ ] Rarely
   - [ ] Sometimes
   - [ ] Often
   - [ ] Very often

31. Over the last 2 weeks, how often have you had to work harder than usual to keep track of what you were doing?
   - [ ] Never
   - [ ] Rarely
   - [ ] Sometimes
   - [ ] Often
   - [ ] Very often

32. Over the last 2 weeks, how often have you had trouble shifting back and forth between different activities that require thinking?
   - [ ] Never
   - [ ] Rarely
   - [ ] Sometimes
   - [ ] Often
   - [ ] Very often
33. **What is the highest grade or level of school that you have completed?**
   - □ 8th grade or less
   - □ Some high school but did not graduate
   - □ High school graduate or GED
   - □ Some college or 2-year degree
   - □ 4-year college graduate
   - □ More than 4-year college degree

34. **Are you of Hispanic or Latino origin or descent?**
   - □ Yes, Hispanic or Latino
   - □ No, not Hispanic or Latino

35. **What is your race? Please choose one or more**
   - □ White
   - □ Black or African American
   - □ Asian
   - □ Native Hawaiian or other Pacific Islander
   - □ American Indian or Alaska Native
   - □ Other

36. **What language do you *mainly* speak at home?**
   - □ English
   - □ Spanish
   - □ Some other language (please print):

37. **Did someone help you with this survey?**
   - □ Yes
   - □ No → If No, please return the completed survey in the pre-paid envelope.

38. **Who helped you complete this survey?**
   - □ Spouse or partner
   - □ Child
   - □ Sibling
   - □ Parent
   - □ Other family member or friend
   - □ Paid caregiver
   - □ Someone else (please print):

39. **How did that person help you complete the survey? Check all that apply.**
   - □ Read the questions to me
   - □ Wrote down the answers I gave
   - □ Answered the questions for me
   - □ Translated the questions for me
   - □ Helped in some other way
40. For what reason did someone help you complete this survey? Check all that apply.
   - Too sick
   - Trouble with memory
   - Trouble seeing or reading
   - Hard of hearing
   - Do not understand English
   - Wanted someone else to complete the survey
   - Patient is in a coma
   - Patient has passed away
   - Some other reason (please print):

41. How familiar is the person who helped you complete the survey with your condition and care?
   - Completely
   - Very
   - Somewhat
   - A little
   - Not at all

42. How often does the person who helped you complete the survey also help you make decisions about your medical treatment?
   - Never
   - Sometimes
   - Usually
   - Always

43. Does the person who helped you complete this survey also provide or help you with any of the following? Check all that apply.
   - Companionship (talking, reading, keeping company) or supervision
   - Transportation (driving to doctor appointment, driving for errands)
   - Homemaking (shopping, cleaning, preparing meals)
   - Personal care assistance (feeding, bathing, toileting, dressing, grooming)
   - Healthcare assistance (help with medications, wound care)
   - Financial assistance (paying bills, managing budget)
   - Other activities (please print):

Thank you for completing this survey.

Please return the completed survey in the provided pre-paid envelope.

RAND Corporation
Attn: Ryan McKay
1776 Main Street
Santa Monica, CA 90401
COVID-19 and Your Care Section (Used in Round 6 Only)

33. Have you heard of the coronavirus, or COVID-19 pandemic?
   □ Yes
   □ No → If No, go to Question 36

34. How much have you had to change your daily activities because of the coronavirus or COVID-19 pandemic?
   □ Very much
   □ Quite a bit
   □ Somewhat
   □ A little bit
   □ Not at all

35. How much have you been worried that the coronavirus or COVID-19 pandemic would affect your care from this provider and team?
   □ Very much
   □ Quite a bit
   □ Somewhat
   □ A little bit
   □ Not at all

36. With virtual visits or virtual care, you can use a computer or telephone to talk with your doctor. With a telephone visit, you can get advice about a health issue by telephone. With a video visit, you see the doctor on your computer or telephone screen when you are talking about a health issue.

   In the past 6 months, have you had a telephone or video visit with this provider or team?
   □ Yes, telephone visit
   □ Yes, video visit
   □ Yes, both a telephone visit and a video visit
   □ None of these → If None, go to Question 38

37. Thinking about your telephone or video visit with this provider and team, how true is the following statement?

   During my telephone or video visit, I felt heard and understood by this provider and team.
   □ Completely true
   □ Very true
   □ Somewhat true
   □ A little bit true
   □ Not at all true
COVID-19 and Your Care Section (Used in Rounds 7–10 Only)

<table>
<thead>
<tr>
<th>COVID-19 AND YOUR CARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>33. How much have you had to change your daily activities because of the coronavirus or COVID-19 pandemic?</td>
</tr>
<tr>
<td>□ Very much</td>
</tr>
<tr>
<td>□ Quite a bit</td>
</tr>
<tr>
<td>□ Somewhat</td>
</tr>
<tr>
<td>□ A little bit</td>
</tr>
<tr>
<td>□ Not at all</td>
</tr>
</tbody>
</table>

| 34. How much have you been worried that the coronavirus or COVID-19 pandemic would affect your care from this provider and team? |
| □ Very much |
| □ Quite a bit |
| □ Somewhat |
| □ A little bit |
| □ Not at all |

| 35. In the last 6 months, have you had telephone or video visits with this provider and team? |
| □ Yes, telephone visits |
| □ Yes, video visits |
| □ Yes, both telephone and video visits |
| □ None of these → If None, go to Question 41 |

| 36. In the last 6 months, how many times have you had telephone or video visits with this provider and team? |
| □ 1 or 2 times |
| □ 3 or 4 times |
| □ 5 or 6 times |
| □ 7 or more times |

| 37. When you were answering the questions in this survey about the care you received from this provider and team, which visits were you thinking about? |
| □ I was thinking mainly about my in-person office visits |
| □ I was thinking mainly about my telephone or video visits |
| □ I was thinking about both my in-person and telephone or video visits |

| 38. Thinking only about your telephone or video visits with this provider and team in the last 6 months, please answer the following questions: |
| Was it easy to use the telephone or video to get care from this provider and team? |
| □ Yes, definitely |
| □ Yes, somewhat |
| □ No |

| 39. Did you get the care you wanted during your telephone or video visits? |
| □ Yes, definitely |
| □ Yes, somewhat |
| □ No |

| 40. How satisfied were you with the care you received from this provider and team during your telephone or video visits? |
| □ Completely satisfied |
| □ Very satisfied |
| □ Somewhat satisfied |
| □ A little bit satisfied |
| □ Not at all satisfied |
Appendix B. Methods and Analytic Procedures

Performance Measure Scoring

To estimate risk-adjusted program-level performance measure scores, we use 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). To calculate performance measure scores at the program level, it is necessary to perform hierarchical regressions with binomial outcomes and a provider-level random effect that will best estimate a scaled adjusted score. In particular, the random effects model for analysis of individual $j$ within program $i$ is

$$Y_{ij} \sim \text{Binomial}(p_j, n_j),$$

where $n_j \equiv$ as the number of items ind. $j$ responds to

$$p_j \equiv \logit^{-1}[\beta_0 + b_i P_i + X_{ij}^T \alpha],$$

where there is an assumption that $b_i \sim N(0, \sigma_b^2)$ and where $\beta_0$ represents overall average program performance, each term $P_i$ is an indicator that patient $j$ experienced care from 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). Because a logit transformation is used in the estimation and in order to obtain a program score in terms of the proportion of top-boxes, retransformation is conducted in which the risk-adjusted score is estimated as

$$\text{Score}_{i,adj} = \frac{\exp(b_i + \beta_0)}{1 + \exp(b_i + \beta_0)}.$$  

Therefore, this formula states that the model will be assessed at all baseline covariate values (i.e., assuming programs all responded by mail and did not require proxy assistance).

In all analyses, we only include data from surveys with at least one (of four) data elements completed for the Feeling Heard and Understood performance measure: i.e., those with no outcome responses were removed from the sample. The model was fit using the brms library in R that fits Bayesian multilevel models (a technical point is that noninformative priors were used for all model parameters—i.e., the software defaults). Bayesian modeling was chosen because it could be used to easily provide distributional estimates for both the ICC and performance measure score reliability.

Statistically Significant and Meaningful Differences in Performance

One way that we tested for statistically significant differences among programs was to use test techniques that are similar to ANOVA, which aims to compare a “full model” with a
“reduced model” (or “nested model”). In particular, we assume that there are \(N_p\) programs and assume the following binomial model for the performance measure score:

\[
\text{logit}(E[Y|P, X, \beta, \alpha]) = \left( \beta_0 + \sum_{p=1}^{N_p} \beta_p P_p \right) + X^T \alpha,
\]

where \(Y\) is defined as the proportion of top-box performance measure values out of the data elements answered (note that this model is a fixed-effects model and not a random effects model).

The reduced (or nested) model is an assumption that there are no differences among the programs, and the full model is an assumption that there is at least one difference. Statistically, the reduced model is a null hypothesis that each of the parameters related to programs is zero, i.e., \(H_0: \beta_p = 0\) for all \(p\). The alternative hypothesis for the full model is that one of these parameters is not equal to zero, i.e., \(H_A: \beta_p \neq 0\) for some \(p\). Under the generalized linear model described in the previous section, these null and alternative hypotheses can be tested using a likelihood ratio statistic. We implemented this analytic method and estimated the adjusted performance measure scores, controlling for survey mode as a risk adjuster.

**Performance Measure Score Reliability**

We conducted a signal-to-noise analysis to quantify performance measure–level reliability. We used hierarchical generalized linear models to relate our outcome measures to our providers and their covariates, in which the hierarchy of data is patient observations within the designated accountable health care entity (i.e., the program). We perform hierarchical regressions with binomial outcomes to decompose the variability. The random effects model for analysis across providers is

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

where we assume that \(b_i \sim N(0, \sigma_b^2)\). In this model, the term \(\beta_0\) 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 this model can be decomposed using the (adjusted) ICC (Rodriguez and Elo, 2003; Wu, Crespi, and Wong, 2012), as follows:

\[
\text{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi}{3}}.
\]

The reliability from the performance measure test is then extended to estimates for future samples by using a function of the observed variances and the sample size from each program using the Spearman-Brown prophecy formula. The Spearman-Brown prophecy formula using an estimated \(\hat{\text{ICC}}\) to relate the reliability \(\rho(n)\) to the sample size \(n\) is given by

\[
\rho(n) = \frac{\hat{\text{ICC}} \times n}{1 + (n - 1) \times \hat{\text{ICC}}}
\]
and allows us to project a required within-program sample size to achieve a desired reliability. Our target value for an acceptable level of reliability is approximately 0.7 and helps determine the required within-provider sample sizes.
References


Casarett, David, Amy Pickard, F. Amos Bailey, Christine Seel Ritchie, Christian Davis Furman, Ken Rosenfeld, Scott Shreve, Zhen Chen, and Judy Shea, “Do Palliative Consultations


CMS—See Centers for Medicare & Medicaid Services.


National Consensus Project for Quality Palliative Care, Clinical Practice Guidelines for Quality Palliative Care, 4th ed., Richmond, Va.: National Coalition for Hospice and Palliative Care, 2018. As of April 18, 2021: https://www.nationalcoalitionhpc.org/ncp/


https://www.rand.org/pubs/working_papers/WRA400-1.html


Smith, Alexander K., Roger B. Davis, and Eric L. Krakauer, “Differences in the Quality of the Patient-Physician Relationship Among Terminally Ill African-American and White Patients:

Steinhauser, Karen E., Nicholas A. Christakis, Elizabeth C. Clipp, Maya McNeilly, Lauren McIntyre, and James A. Tulsky, “Factors Considered Important at the End of Life by Patients, Family, Physicians, and Other Care Providers,” *JAMA*, Vol. 284, No. 19, November 2000, pp. 2476–2482.


