SHORT FORM PATIENT EXPERIENCE SURVEY – **RESEARCH FINDINGS**

OCTOBER 2015

MHQP
MASSACHUSETTS HEALTH QUALITY PARTNERS
trusted information. quality insights.

CHPI
California Healthcare Performance Information System

Final findings report covering the bicoastal short form patient experience survey pilot conducted jointly by

Massachusetts Health Quality Partners (MHQP)

and

California Healthcare Performance Information System (CHPI)

ACKNOWLEDGEMENTS

Massachusetts Health Quality Partners (MHQP) and the California Healthcare Performance Information System (CHPI) are grateful to the Center for Healthcare Transparency for providing funding for this pilot project.

MHQP would like to thank our study collaborators at CHPI whose expertise, hard work, and commitment to implementing a dual market test to evaluate new methods of electronic surveying and a standard but shortened survey measurement tool helped bring this project to fruition.

We are very grateful to the following provider organizations that partnered with us for providing their time and data resources, without which this pilot project would not have been possible: New England Quality Care Alliance (NEQCA), Reliant Medical Group, Steward Healthcare, and UMASS Memorial Medical Group. Equally, we appreciate the efforts of our partners from the participating health plans—Blue Cross Blue Shield of Massachusetts, Fallon Health, Harvard Pilgrim Health Care, Health New England, and Tufts Health Plan – for providing data and continuing to support our efforts.

We would also like to recognize Dr. Bill Rogers for his methodological leadership and expertise; and Paul Kallaur and Jacqueline Cho at the Center for the Study of Services for their professional management of survey administration. We are indebted to MHQP staff who worked so hard to develop this project and create this report – Rose Judge, Amy Stern, Hannah Cain, Janice Singer, and Barbra Rabson.

CHPI would like to thank our physician organizations for their partnership and support of this undertaking through their participation. Participating Physician Organizations were: Brown & Toland, Facey Medical Group, NorthBay Healthcare, Palo Alto Medical Foundation, Primecare, Prospect Medical Group, Riverside Medical Center, Sharp Community Medical Group, Sutter Gould Medical Foundation, and UC San Diego Health System Medical Group. Meghan Hardin, Kai Carter, Rachel Brodie, and Maj-Britt Llewellyn led CHPI's program work and authorship of this report and we are grateful for their contributions.

We'd also like to thank our partners at MHQP for their hard work and unique perspective on patient experience. We've learned much about their environment and approaches to patient experience surveying and look forward to future collaborations.

We would also like to recognize Dr. Bill Rogers for his methodological leadership and expertise; and Jeff Burkeen and Chris Altieri at the Center for the Study of Services for their professional management of survey administration.

Finally, we'd like to acknowledge the CAHPS Reports Team, which is funded by AHRQ, and includes researchers from RAND (Steven Martino, PhD; Andrew Parker, PhD; and Melissa Finucane, PhD); Yale University (Mark Schlesinger, PhD); the University of Wisconsin-Madison (Rachel Grob, PhD); and Shaller Consulting (Dale Shaller, MPA).

TABLE OF CONTENTS

PAGE

EXECUTIVE SUMMARY

For over a decade Massachusetts Health Quality Partners (MHQP) and California Healthcare Performance Initiative (CHPI) have measured and publicly reported about patients' experiences of care in the ambulatory care setting. Our initial efforts were groundbreaking. Over the past ten years, the drive toward patient-centered models of care and value based reimbursement have made patient experience measurement even more important and valuable for patients, providers and payers. During this same time, advances in communication technology have dramatically and profoundly changed our culture and these changes now challenge our earlier protocols for collecting information from patients. In order to maintain high response rates in a cost effective manner, we believe that traditional survey methods, such as mail and computer-assisted telephone interviews, will not be sufficient.

In an effort to spur innovation in this area, the Center for Healthcare Transparency (CHT) generously supported the implementation of a pilot study in our organizations' respective markets. The objective of the pilot was to evaluate new methods of surveying that make valid and reliable information about patient experience more widely available.

PROJECT DESCRIPTION

The overarching purpose of the Short Form Patient Experience Survey Project was to develop and evaluate new methodological approaches to make high value performance information available to the public. We also hoped to find ways to make patient experience surveys less expensive and reduce the burden of response for respondents without sacrificing the scientific rigor behind reported results. In the long term, if expenses can be significantly lowered it should be more feasible for organizations like ours to collect information about individual doctors. This level of information would be most helpful for consumers trying to make choices about care and for providers who are trying to improve patient experience.

MHQP and CHPI fielded their annual statewide patient experience measurement projects using a long form survey in early 2015 and tested a short form electronic and mail survey in parallel to these efforts. We used results and analytic work from our past surveys to support the evaluation of the pilot.

PROJECT OBJECTIVES

Our test of shorter and electronic versions of a survey was designed to answer these questions:

1) Will a short form survey provide comparable answers and rank providers similarly when compared with existing long form statewide surveys?

2) Will email approaches result in sufficient response rates and rank providers in a comparable way to the mailed short form?

These objectives are slightly different from our original intent to show that email short forms would produce results similar to that of mailed long forms. Despite a robust effort, our organizations could not generate enough email addresses from participants to address this broader and more desirable objective. The lack of widespread and systematic collection of patient emails is a major barrier. In addition, provider organizations are sensitive to the need for patient privacy and expressed concerns about using collected information to contact patients outside of their offices. These facts led, in a sense, to one of our most important findings: despite significant advances in communication technology, we must still rely on adding mail and phone survey modes to achieve sufficient response because provider organizations are not yet collecting and maintaining valid email addresses in a systematic way.

STUDY DESIGN

MHQP and CHPI recruited a subset of practices (in MA) and medical groups (in CA) who were willing to provide email addresses and invite their patients to participate in the survey. In CA, medical groups had traditionally been identified as the sponsor of the survey, however, in MA, the sponsors of the statewide effort have been MHQP and MA health plans and provider organizations.

In total, 57,683 patients seen by 1,862 individual physicians from 48 physician organizations in the two states received surveys as part of the pilot. The pilot sample frame included adult patients seen by PCPs at participating organizations during each state's measurement period. In California, 4,813 patients received the short form survey via email and 16,852 by mail. In Massachusetts, a total of 12,303 received the short form by email and 23,715 by mail. Patients who received email surveys had the option of also responding to open-ended questions at the end of their survey.

Both states fielded a parallel long form survey. In California, using the long form Patient Assessment Survey (PAS), 51,173 patients were surveyed originally by email and over 167,893 by paper. In Massachusetts, 177,685 patients received the long form Patient Experience Survey (PES) by mail; no Massachusetts patients received the long form by email.

Sample design was multi-layered to maximize usefulness of the available sample. The first layer was at the provider level to answer questions about doctors because information about doctors is of most interest to consumers and providers and because we see the largest differences in performance among providers. The second layer was at the practice (MA) or medical group (CA) level, and it was designed to allow for similar comparisons at those levels, reflecting the focus of current statewide surveys.

STUDY ANALYSIS AND FINDINGS

A primary aim of the pilot study was to test a shorter survey to determine if it gives comparable answers and ranks providers in a comparable way to existing statewide surveys. The demographic characteristics of the two regions differed significantly and substantially. California's sample was more racially and ethnically diverse, less educated and had more females than the MA study population. Differences in demographics were controlled for within each state's analysis so that comparisons could be made.

SHORT FORM ANALYSES

Our hypotheses were: 1) the short form will have better response rates than long forms within each state; 2) the short form and long form will produce comparable distribution of responses in both California and Massachusetts at the provider level; and 3) the short form and long form will rank doctors (in both states), practices (in MA) and medical groups (in CA) in a similar way.

Table 1: Massachusetts and California Short Form Results

Short Forms	Massachusetts	California
Response rate	Short form higher than long form (same time in field) (26.9% v.24.5%)	Short form same response as long form (responses arrived quicker but short form was fielded a shorter time than long form (25.4% v. 26.3%)
Results	Overall findings same for short and long forms with small exceptions at the item level*	Overall findings same for short and long forms with small exceptions at the item level*
Ranking of providers	Same for short and long form at the item level*	Same for short and long form at the item level*

* This study did not test comparability of results or ranking at the composite level.

In addition to the high level results in the table above, the data provides several interesting findings:

- In Massachusetts, where the short and long forms were kept in the field for a similar amount of time, mail response rates were substantially higher for the short form. In California, short form responses were returned more quickly than the long form, however, the total response rates were the same. It should be noted that the fielding time was little more than half that of the long form, and therefore, it is possible that had the short form been in the field longer, California's response rates would have been higher for the short form than the long form.

- In general, in both states responses were comparable at the provider level regardless of survey length although there were some small differences according to the mode of response, particularly with regard to talking about stress in MA where mean scores were higher with the short form web responses.

- Doctors, practices (MA), and medical groups (CA) were almost universally ranked similarly in both states. All of the convergences (agreement of "true" answers) in both states were high, and in fact higher than in other pilots of this type.

EMAIL ANALYSES

Our hypotheses for this arm of the pilot were: 1) short form email response rates will be better than long form email response rates; 2) emailed short forms will give comparable responses to emailed long forms and mailed short forms; 3) emailed forms will rank providers in a comparable way to mailed forms that have similar content; and 4) patients in the sample frame who have emails will have different demographics than patients who do not have emails.

Table 2: Massachusetts and Californian Email Results

Email results	Massachusetts	California
Response rate (RR) emailed vs. mailed short form only	Short form email RR was not as high as short form mail (21.5% vs. 29.7%)	Short form email with mixed mode follow-up RR was higher than short form mail (33.6% vs. 24.1%)
Response rates emailed short vs. emailed long form	Not tested in MA	No significant difference
Comparability of results	Comparable to both long and short form mail survey	Comparable to both long and short form mail survey
Email population demographics vs. mailed population	Slightly older and female	Slightly older and female

Email results in the table above were interesting to us for the following reasons:

- In MA, email yield rates (usable response/surveys attempted) were higher than expected (21.5%) but not as high as mail yield rates (29.7%). They also were not as high as the total mixed mode yield in CA which was 33.6%. This is comprised of a yield of 9.8% for email responses, 21.2% for follow up with mailed paper surveys, and 2.4% for web responses using the link from the paper survey.

- There was not an appreciable increase in email yield rates associated with the short form in CA where a comparison to long form email could be made.

- Email short forms did provide comparable results to both long and short form mail surveys in both states.

- In both states, the demographics of those with email have shifted from what they have been in previous studies. Patients with email now tend to be slightly older and more female, whereas in earlier email studies they were younger and overwhelmingly male.

In addition to testing the hypotheses related to our key questions, we used the data collected through this pilot to do additional analyses and these results are included in our detailed results. Based on adjusted results, we looked at the relationship of responses by survey approach (long form vs. short form and email vs. mail) and response mode (mail vs. online). Differences found at this level were small. However, for organizations in MA and CA that are sensitive to small changes that could affect financial incentives, these pilot results could be used to adjust for trending.

We also considered different ways to summarize results of the short form survey through question groups or composites. Grouped communication questions and grouped questions about providers engaging with patients in talking about care showed high internal consistency among the items and fit together well. Two questions related to care coordination and two questions related to access did not perform as well when grouped together as composite measures. Additional questions could improve the access composite but the care coordination questions are best kept separate.

Lastly, in response to the growing interest in patient narratives about services, we included open-ended questions that were fielded in both states' electronic short form pilots to test the feasibility of eliciting narrative responses from patients. We chose to test two different sets of open-ended questions with respondents randomly assigned to receive either a three-item elicitation or a five-item elicitation. Preliminary findings indicate that response rates were similar for both sets of questions. However, those receiving the five-question protocol gave longer responses than did those receiving the three-question protocol. There were differences in response rates by patient characteristics in both MA and CA samples, but none of these varied significantly between the two sets of questions. In general, non-respondents were younger, less educated, and more often Asian. There were no differences in length of narrative (i.e., word count) across patient characteristics within the CHPI data, but within the MHQP data, younger adults and women gave longer responses. Responses were largely positive and correlated with responses to the close-ended items.

CONCLUSIONS

The results of this pilot indicate that new and more innovative approaches to surveying are evolving and promising for our efforts to reduce survey costs and burden. A move in this direction will help make high value performance information available to the public. Along with the promise of new approaches, we encountered the reality of the current state of patient information systems and the limited experience many physician organizations have in engaging their patients in new ways. We found that full implementation of the email modes of survey is hindered by the lack of systematic collection, verification, and maintenance of patient emails. Email surveying is currently not a viable survey option by itself. However, the results in both states, including a response rate of almost 40% for phone responses to the PAS long form survey in California, suggest that multi-mode surveys that reach patients in numerous ways are the best option. Indeed, multi-mode surveys must be fielded to achieve results that are reliable enough for high stakes use. We do believe that if systematic collection of patient contact information is widely adopted, and the concerns about privacy protection are addressed for patients and providers, the ability to survey through electronic modes can improve substantially.

With the constraints noted, the generally positive results we achieved in testing our hypotheses through this pilot are strengthened by the fact that overall, similar results were found in two regions with significantly different health systems and patient characteristics. Contrary to prevailing wisdom, the Massachusetts response rates were better for a short form survey than they were for the longer 61-question survey, and California might have had higher response rates had the fielding period been longer. In addition, the short form survey instrument appears to be a viable alternative to longer form surveys with the finding that relative scores for doctors, practices, and groups to the same question are generally comparable for the short form based on convergence statistics with case mix control. Further, response to the email short form survey produced results that are comparable to long form surveys and mail short form surveys.

It will be important for organizations that are considering whether to use a short form to understand that the composite scores from the current long form and the new short form may or may not rank providers, practices or groups similarly depending on the items included in the composites. In this study, we were not able to test our composites to determine if they score and rank providers the same. However, as the items that comprise the composites rank providers the same, the new short form composite rankings will be valid. If results for some composites are used for high stakes purposes, it would be prudent to do one of the following:

- Keep the composites construction stable until other changes are made to the instrument;
- Remove items that are not in the short form composites from the previous year's long form survey composites and run an analysis to see if the composite results, now having identical items, are comparable (note that these should be comparable since the items from both years are convergent);
- Or complete a small pilot project using the old and new composites so that adjustments can be made if needed.

Our detailed analysis of adjusted results found small differences in responses by survey approach (long form vs. short form and email vs. mail) and response mode (mail vs. online) that can be adjusted for in CA and MA for provider organizations that might be sensitive to small changes as they relate to financial incentives. Other markets can use the same adjustments noted in this report as they are likely to follow the same pattern as in CA and MA. However, if stakes are sufficiently high in a given market, a small test with 2500 respondents using the new survey prospectively (to compare with a large fielding of the old survey) or the old survey restrospectively (to compare with a large fielding of the new shorter survey) would be suggested so that local adjustment factors can be determined.

With our short form pilot and the release of CG-CAHPS 3.0, we are headed toward broader use of shorter surveys that will be less expensive to mail and less burdensome to patients. If these shorter surveys are fielded with a scientific sampling frame and a full multi-modal survey approach including email, they will yield comparable information to current surveys with better response rates. Altogether, these steps should increase confidence in the ability to achieve reliable results at a lower price.

Finally given the numerous avenues consumers have to comment on a variety of products and services they receive, including health care services, it is important that a systematic way to elicit and analyze such comments be designed. The development of nationally recognized protocols for collecting and reporting narratives is necessary to increase provider acceptance and the likelihood of effective quality improvement efforts. The public reporting of such comments on health care quality websites has the potential to help consumers make more informed choices and increase their interest in engaging with their providers in conversations about care. The work we have done in this pilot is a first step in the direction of systematic collection and reporting. We hope to continue this work and advance the field further in the coming months and years.

INTRODUCTION

For over a decade Massachusetts Health Quality Partners (MHQP) and California Healthcare Performance Initiative (CHPI) have measured and publicly reported about patients' experiences of care in the ambulatory care setting. Our initial efforts were groundbreaking. Over the past ten years, the drive toward patient-centered models of care and value based reimbursement have made patient experience measurement even more important and valuable for patients, providers and payers. During this same time, advances in communication technology have dramatically and profoundly changed our culture and these changes now challenge our earlier protocols for collecting information from patients. In order to maintain high response rates in a cost effective manner, we believe that traditional survey methods, such as mail and computer-assisted telephone interviews, will not be sufficient.

In an effort to spur innovation in this area, the Center for Healthcare Transparency (CHT) generously supported the implementation of a pilot study in our organizations' respective markets. The objective of the pilot was to evaluate new methods of surveying that make valid and reliable information about patient experience more widely available.

PROJECT DESCRIPTION

The overarching purpose of the Short Form Patient Experience Survey Project was to develop and evaluate new methodological approaches to make high value performance information available to the public. We also hoped to find ways to make patient experience surveys less expensive and reduce the burden of response for respondents without sacrificing the scientific rigor behind reported results. In addition, in the long term, if expenses can be significantly lowered it will possible for regional collaborative organizations like ours to collect information about individual doctors. This information resonates best with consumers and providers.

Both MHQP and CHPI planned to field their annual statewide patient experience measurement projects using a long form survey in early 2015. In response, to CHT's request for proposals for research projects to develop and evaluate new methodological approaches to make high value performance information available to the public, MHQP and CHPI proposed testing a short form electronic survey in parallel to these efforts, and using results and analytic work from our past surveys to support evaluation of the results.

PROJECT OBJECTIVES

Our test of a shorter electronic survey was designed to answer these questions:

1) Will a short form survey provide comparable answers and rank providers similarly when compared with existing long form statewide surveys?

2) Will email approaches give sufficient response rates and rank providers when compared with existing long form statewide surveys?

These objectives are slightly different from our original intent to show that email short forms would produce results similar to that of mailed long forms. Despite a robust effort, our organizations could not generate enough email addresses from the provider organizations that agreed to participate to address this broader and more desirable objective. That was, in a sense, one of our most important findings: in spite of significant technology advances in communication, we must still rely on something other than email (e.g., hard copy mail or phone) to achieve adequate response numbers because provider organizations are not yet collecting email addresses in a systematic way.

This report contains detailed analyses of both organizations' results as well as a combined entity analysis that highlights similarities and differences between the two organizations' findings. Each organization's analyses present information on sample numbers and yield rates (usable responses/ surveys attempted), followed by information on the comparability of rankings and item means, comparing short and long form as well as differences in method of survey administration and survey response mode.

In answer to growing interest in patient narratives about services, open-ended questions were fielded in both states' electronic short form pilots to test the feasibility of eliciting narrative responses from patients. Two different sets of open-ended questions were tested with respondents randomly assigned to receive either a three-item elicitation or a five-item elicitation. Preliminary findings are included here and more work is planned to conduct more in-depth qualitative analysis.

CALIFORNIA HEALTHCARE PERFORMANCE INFORMATION SYSTEM (CHPI)

Over the last three years, CHPI has seen response rates to our traditional long form (Patient Assessment Survey (PAS)) decline, and so the launch of the short form pilot comes at a very critical point for the program where we have the opportunity to evaluate our options and chart the future of our statewide survey. The pilot returned encouraging findings that confirmed our hypotheses about how a short form survey will compare to the standard long form instrument. Most notably we saw that the short form produced about the same response rates as the long form. This is important because short form surveys are less expensive to administer than long form surveys particularly when comparing costs for the paper-based, postage paid format. Response rates to the short form might have been even higher if the fielding timeframe had been longer or equivalent to that of the long form PAS. A second outcome of the pilot was that medical groups and doctors were ranked similarly in the short form pilot and the long form PAS.

The pilot also tested the utility of using email to improve response rates, provide similar responses to other modes, rank providers and medical groups the same way across modes, and to study the demographics of email responders. We found that not enough patients have good email addresses for email to be a viable option by itself, however, deploying a multi-modal approach resulted in a 33% response yield, confirming that email is a valuable contributor to higher response rates. If a larger number of email addresses had been useable (good) in our pilot, response rates would have bested previous years and approached the level of hardcopy mail. While the email mode was helpful for improving response rates, there was no appreciable difference between response rates for the emailed short form and the emailed long form survey. It did however produce comparable results to both long and short form paper surveys. Finally, we found that patients who responded to the email invitation shifted to older females, compared to previous studies where they were younger and mostly male.

CHPI SAMPLE FRAME

Participants for this study were identified through physician organizations (POs) with whom CHPI had previously or is currently engaged with to field the “Long Form” Patient Assessment Survey (PAS). There were three primary methods for reaching these groups to share information about the short form pilot. The first was via an informational webinar held on September 9th, 2014 and was open to all PAS participating medical groups. The presentation contained information about the benefits of a short form survey, answers to frequently asked questions about participation in the pilot, an example of the short form survey, and information for participants about how to get involved. A second, more formal webinar was held on September 19th, 2014 to garner interest in the program and to follow up directly with participants. Secondly, letters were sent to all PAS participating physician organizations describing the reason for the pilot, and the benefits and requirements of participation. Finally, direct phone calls and email correspondence was carried out in October and November in order to secure interested groups and to assess readiness, answer questions, and encourage participation. Twenty-one groups initially showed strong interest, and after phone and email follow-up, 11 medical groups accepted the terms to participate in the pilot.

CHPI STUDY DESIGN AND DATA COLLECTION

With 11 participating medical groups representing 997 individual practice sites and over 10,500 physicians, CHPI began the fielding process on February 25th, 2015. The 29,700 patients surveyed as a part of this study were seen by 180 unique physicians. All recipients were 18 years of age or older, commercially-insured or managed care patients (HMO and POs) enrolled in POs in California who: 1) had at least one medical encounter between January 1 and October 31, 2014 and 2) were members of the PO on October 31, 2014. The survey asked patients to evaluate the care they received during the previous 12 months.

While California's PAS does assess specialists, for the purposes of this study, only PCPs were included in the comparative analysis of both long and short forms in order to eliminate an additional variable in an already complex study.

Surveys for the pilot were administered and collected in one of three ways: 1) short form email only, 2) short form paper based mail only, and 3) short form email followed by short form paper-based mail. All mail versions also offered an option for the respondent to go online using a unique web address to complete the survey electronically. While the short form pilot surveys were fielded alongside the traditional PAS, it should be noted that the core PAS Group Survey samples were drawn before the pilot samples, so as to track the one-adult-per-household limit across all samples. Additionally, the pilot data were captured separately from those of the standard PAS surveys – the email invitations had links to pilot versions of the web survey and the printed versions had unique codes to distinguish them for data entry.

Enrollment data from each PO was collected in the form of three files: patient visits, patient demographics and active practitioners (PAS Data Specifications). An automated quality assurance system performed over 100 data quality checks on each data submission based on the PAS data quality criteria.

The survey research firm, the Center for the Study of Services (CSS), drew a sample of patients for each reporting unit and then stratified them by visits to Primary Care Physicians and Specialty Care Physicians, and within strata, patients were randomly selected. To increase the likelihood of responding, sampling was prioritized by the most recent date of visit. The unit of analysis is, in most cases, the PO.

CHPI YIELD RATE

In conducting this pilot, we wanted to better understand the impact of survey length and administration mode on yield rates (usable responses/surveys attempted). In California, two protocols were used, depending on whether the patient had an email address or not. These two protocols were used for both the standard long form PAS and the short form pilot. If a patient had an email address, they were sent an email invitation to participate in the survey, and a second email invitation was sent out 4 weeks later to all non-responders, excluding those with bad email addresses. The fielding period was kept open for an additional four weeks, after which all non-responding patients were included in the paper-based survey mailing ("Email->Paper"), along with patients who did not have an email address on file and did not receive the email invitation. These patients could respond by mail or by web using a unique web address provided on the paper-based form. If responding electronically at this point, the patient manually typed in the web address to get to the survey.

Table 1: Yield Rate and Yield Demographics for California Protocol

Survey Protocol	N	Response Type	Yield %	Average Age	Male	College Ed.	White	Hispanic
PAS Email*	51,173	Email (Web)	14.1	58.0	34.5	49.7	65.2	13.5
PAS **Email->Paper**	44,158	Mail	18.9	58.6	34.2	45.3	61.0	15.3
PAS **Email->Paper**	44,158	Web	2.6	49.9	39.6	58.4	54.9	16.4
PAS **Email->Paper**	44,158	No Response		46.8	34.5			
Total PAS Email			32.7					
PAS Paper	167,893	Mail	23.9	59.5	38.2	36.8	54.0	22.3
PAS Paper	167,893	Web	2.4	50.1	48.5	54.5	55.7	18.1
PAS Paper	167,893	No Response		48.0	41.3			
Total PAS Paper			26.3					
Short Form Email*	4,813	Email (Web)	9.8	56.1	37.8	45.1	59.7	18.0
Short Form **Email->Paper**	4,342	Mail	23.5	60.0	35.2	36.5	59.3	13.3
Short Form **Email->Paper**	4,342	Web	2.9	48.8	45.6	55.3	44.6	18.9
Short Form **Email->Paper**	4,342	No Response		44.9	36.6			
Total Short Form Email			33.6					
Short Form Paper	16,852	Mail	21.8	58.2	38.7	39.0	57.8	20.8
Short Form Paper	16,852	Web	2.3	49.7	50.0	60.0	57.2	16.7
Short Form Paper	16,852	No Response		47.1	42.1			
Total Short Form Paper			24.1					

**** For the purpose of this analysis, there were some patients who only received emails and were not forwarded into the mail protocol for email non-respondents.***

The short-form yield rate for email was lowered by the relatively high occurrence of bad email addresses for the "patient with email address" sample (33.7%). The short form email response yield could have been as high as 21% if the short form email addresses had been as good as for PAS. The short form mail response rates were also lowered due to a short fielding period (8 weeks) which was deemed necessary at the time of fielding due to time constraints around the completion of the pilot. Had the short form been kept in the field longer, the overall number of usable responses might have been substantially higher, possibly as much as 4% higher. In summary, even with the large quantity of bad email addresses and the shorter fielding time for the short form survey, the overall short form yield was about the same as the long form PAS. If the projected adjustments noted above were included, the short form yield rates could have been several percentage points higher than PAS.

Aside from showing the possible benefits of a short form survey, the table also clearly illustrates the value of an email component. About one fourth of all patients surveyed have email addresses, and about one fifth of those patients who received an email invitation responded. The percentage of paper survey respondents is not much reduced by the prior email, and the net gain in yield rates is between 4% and 5%.

The table also shows that demographic differences are almost all statistically meaningful given the large sample sizes. The short forms appear to widen the appeal of the survey to less educated, minority respondents. Non-respondents are typically much younger than respondents.

COST CONSIDERATIONS

In addition to the benefit of producing potentially better response rates, a shorter survey also costs less. This is important because short form surveys are less expensive to administer than long form surveys when comparing costs for the paper-based, postage paid format. The table below shows the base costs for each mode.

Table 2: Base Cost by Survey Mode

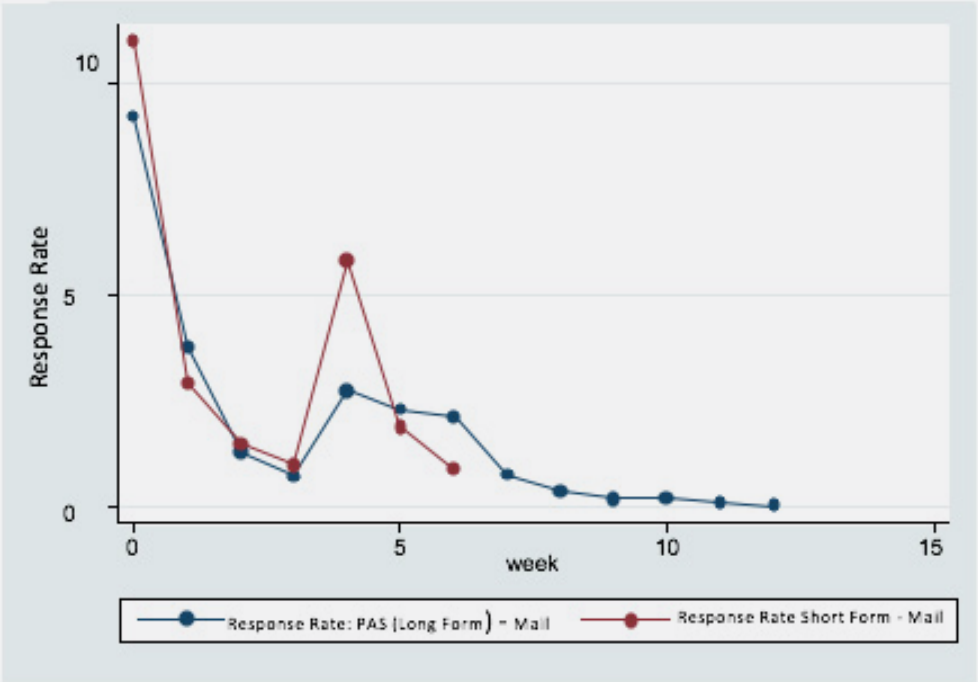
Mode	Short Form	Long Form
Paper*	$8.95/complete	$10.24/complete
Email	$0.25/complete	$0.25/complete
Phone Follow up	$14.25/complete	$20.37/complete

** Base cost per sample member, discounted to $0.25 if the patient responds by email/web prior to the first paper mailing.*

SHORT FORM RESPONSE RATES

Comparing response rates for the short form mailed survey against those for the PAS long form survey, we found that the short form produced higher response rates in weeks 1 and 4, but that at other time points, rates were variable. The chart below illustrates this point, and more importantly shows that the fielding time for the short form ended at 8 weeks, after which it is possible that more surveys could have arrived, as they did for the long form. Therefore we can conclude that mailed short form survey response rates were not worse, but they may or may not have been better, depending on what was lost due to early survey closure. The graph also conveys the conclusion that responses to the short form were faster (higher in week 0 but lower in week 1, and similarly for week 4 vs. 5-6), however, because the surveys were launched at different times with different conditions at the post office (post office behavior is a big factor in the timing of receipts), we cannot completely understand whether the phenomenon is user based or post office based.

Figure 1: CA Adjusted Response Rate (%) for the Two Mail Response Samples by Week



PSYCHOMETRICS OF RESPONSES

A key objective of this pilot study was to determine if the rankings of doctors will be comparable 1) between the short form and the long form and 2) by survey mode (email versus paper) for short form only.

There are essential psychometric issues that are considered when evaluating results for the short form as compared to the long form: 1) reliability and 2) concordance or conceptual agreement with the current long form survey and composites measures. As noted in the MHQP section below, the analyses conducted for California and Massachusetts were not identical due to differences in the survey administration and/or response rates. In Massachusetts, there were sufficient email responses to perform email and web response psychometric analyses as well as short versus long form analyses. In California, there were not sufficient email responses to evaluate email versus web responses. Therefore, in the following two tables, the results from email and web are combined whereas for Massachusetts these are broken out into more granular analyses.

Reliability is a statistical measure that indicates how accurately a measure captures information by measuring the consistency of the information provided by respondents. Reliability can be expressed in a number of ways. It depends on survey context (i.e., the entities being compared); for this study, the context examined is a particular set of doctors or groups that have been sampled with a given instrument. Because more responses improve our ability to calculate consistency, reliability is also dependent upon the sample size and completed responses received, noted in results analysis as "N". In our analysis, we evaluated how many responses are needed to achieve reliability levels that are acceptable for high stakes use of results. A reliability level of 70% is generally accepted as appropriate for high stakes use.

Observed results for doctors or medical groups vary for two reasons. One of these reasons is sampling error, which in statistics is captured by the amount that patients within the doctor or practice vary, and the size of the sample. Variation within a doctor or practice is measured by the within-entity standard deviation (SD). The other reason observed results vary is that the doctors or practices in the particular context are truly different from each other. The true differences are captured through a statistical estimation process known as variance components analysis, and it is expressed as a standard deviation which we label the context SD. The formula used to evaluate reliability within context is the following:

R = (context SD)2 / ((context SD)2 + ((within-entity SD)2/sample size))

The square of the standard deviation is given the technical name "variance" in statistics; in these terms the reliability is the proportion of observed variance that can be credited to true differences between doctors or practices. The within-entity (e.g., within doctor) variance depends on the survey question, the available responses, and the amount of agreement among the responses for given entities.

To compare differences in long form versus short form results, our evaluation matches results for both versions at the doctor level. The matched dataset only uses doctors who are in both versions. The values in the following table were the values observed in the matched sample from the long form statewide survey and the short form pilot for mail responses and web responses.

Table 3: CA Reliability Details at the Doctor Level

	Within-Doctor Standard Dev.		Doctor-level Standard Dev.	
Item	PAS (Long Form)	Short Form	PAS (Long Form)	Short Form
Getting care right away	28.1	28.7	8.6	8.3
Doctor gives clear explanations	18.8	19.4	4.3	5.1
Doctor listens carefully	19.5	20.2	4.4	5.8
Doctor knows medical history	22.8	21.8	5.3	5.8
Doctor spends enough time	21.2	22.4	5.2	6.4
Doctor is informed about specialty care	28.5	28.9	6.1	6.5
Doctor follows up after tests	33.4	30.2	6.5	7.3
Doctor talks about stress	49.2	49.6	10.2	9.0
Doctor rating	16.8	17.7	4.3	5.5
Rating of care	15.9	16.2	3.0	3.1

In most cases the within-doctor SD went up slightly (lowering reliability), but the doctor-level SD went up significantly, raising reliability. This can be understood in terms of the N size required for 70% reliability, as shown in the following table.

Table 4: CA Doctor Psychometric Properties of Short-Form vs. Adjusted PAS Items

	Full Datasets		Matched Datasets				
Item	Mean PAS (Long Form)	Mean (Short Form)	Mean PAS (Long Form)	Mean (Short Form)	N PAS (Long Form)*	N (Short Form)*	Converg ence
Getting care right away	79.0	77.6	78.8	78.0	25	28	1.000
Doctor gives clear explanations	91.7	91.6	92.6	91.3	46	34	1.000
Doctor listens carefully	91.6	91.5	92.6	91.2	46	29	1.000
Doctor knows medical history	87.0	88.7	88.7	88.2	42	34	1.000
Doctor spends enough time	89.4	88.8	90.7	88.5	39	29	0.988
Doctor is informed about specialty care	80.3	81.2	81.9	80.4	50	47	1.000
Doctor follows up after tests	78.0	82.3	79.6	81.7	61	41	0.935
Doctor talks about stress	43.0	46.9	44.8	46.8	55	72	1.000
Doctor rating	89.1	89.4	90.3	88.9	36	24	1.000
Rating of care	88.0	87.3	89.0	87.0	68	64	1.000

** Note: N PAS (Long Form) and N (Short Form) are the sample sizes required for 70% reliability for the regular PAS and short forms, respectively.*

The convergence statistic seen in Table 4 is a measurement of how well two concepts being measured would correlate if there were no sampling variaton. In this table, the two concepts come from the two groups being compared, i.e., the mean values for short form vs. long form. The correlation is across the results for doctors (or medical groups in Table 5 below). Since conversion is a correlation coefficient, +1 represents perfect agreement, 0 is no agreement, and -1 is perfect inverse agreement.

As in reliability, convergence depends on the concept of a "true" doctor or group mean, which provides the "context" of the study. To calculate this, we first compute the adjusted doctor or medical group mean in each setting using the available data. We then estimate the accuracy of each mean based on (variable within-entity variance)/(entity sample size), as we would for a reliability calculation. Any entities that do not have data in one sample or the other are discarded. The rest of the calculation is based on a maximum likelihood estimate where the parameters are the two global averages, the two entity level standard deviations, and the correlation (convergence) between the two scores. The two global averages are shown as the "matched averages" and the convergence is the correlation. All of the convergence scores betwen the short form and PAS are very high. This suggests that the long and short form provide similar informaton about providers.

Except for the item "Doctor Talks about Stress", the sample sizes improved or stayed within reasonable bounds. A similar table is constructed for medical group data, with similar results:

Table 5: CA Group Psychometric Properties of Short Form vs. Adjusted PAS Items

	Full Datasets		Matched Datasets				
Item	Mean PAS (Long Form)	Mean (Short Form)	Mean PAS (Long Form)	Mean (Short Form)	N PAS (Long Form)*	N (Short Form)*	Conver-gence
Getting care right away	79.0	77.6	79.3	79.0	140	90	1.000
Doctor gives clear explanations	91.7	91.7	92.7	91.3	433	112	1.000
Doctor listens carefully	91.6	91.5	92.6	91.5	362	119	1.000
Doctor knows medical history	87.0	88.8	88.6	88.4	215	100	1.000
Doctor spends enough time	89.4	88.8	90.7	88.7	300	102	1.000
Doctor is informed about specialty care	80.3	81.3	82.2	80.6	162	128	1.000
Doctor follows up after tests	78.0	82.3	79.6	81.9	138	73	1.000
Doctor talks about stress	43.0	46.9	45.5	47.8	202	369	1.000
Doctor rating	89.1	89.5	90.3	89.0	134	63	1.000
Rating of care	88.0	87.4	89.0	87.1	157	103	1.000

** Note: N PAS (Long Form) and N (Short Form) are the sample sizes required for 70% reliability for the regular PAS and short forms, respectively.*

COMPOSITES

While the short form does not allow all of the existing PAS composites to be scored, there are some potential composites in the short form. The following MAP (Multitrait Analysis Program) results describe the 5 composites that were initially hypothesized:

Table 6: CA Composite Items and Scaling Success

Composite	Item	Alpha	% Scaling Success
Access and Information about Access	• Getting care right away • Information about getting care after hours	0.33	63
Provider Communication	• Doctor gives clear explanations • Doctor listens carefully • Doctor knows medical history • Doctor spends enough time	0.92	100
Care Coordination	• Doctor is Informed about specialty care • Doctor follows up after tests	0.64	88
Patient Engagement	• Doctor talks about goals • Doctor talks about things that make care hard • Doctor talks about stress	0.67	100
Ratings of Care	• Doctor rating • Rating of care • Willingness to recommend doctor	0.87	100

In this table, Alpha refers to Cronbach's alpha which is a measure of the internal consistency of the items. Scaling success is the percentage of possible item reassignments where the hypothesized (listed) composite assignment is better than an alternative assignment. The composite analysis considers how well aligned each item is to its own composite and all of the other composites. With respect to a particular alternative composite, if the item is closer to its own composite, that is considered a scaling success; if closer to the alternative composite it is a scaling failure. The percent scaling success is the percentage of successes across all of the items and all of the other possible composites these items might have been assigned to.

Three of the five composites hypothesized worked extremely well. The items in the communication composite are highly correlated in the short form similarly to what we find in the in the long form. Creating a new "engagement" composite based on measures that ask about whether a doctor (or doctor's office) talked with a patient about care is a logical first step toward measuring whether doctors and patients interact. All three items involve a two-way communication between the provider and the patient and elicit information from the patient. Further work needs to be done to better understand the quality of the interaction.

The access composite for the short form fails, however, and there are no other possible items in the survey to group with "getting care right away". "Information about getting care after hours" is an item that predicts the summary scores, but has no appropriate companion in this structure. An access composite that has a higher percent scaling success could be created, using additional different questions about access. More analysis is needed to determine what items would need to be included in the survey to form a strong access composite. Similarly, the items in the care coordination composition did not align. Consideration should be given to reporting these two items separately and new items that might better align to assess care coordination should be considered.

While we have seen that scores on individual items are comparable in the short form and the long form and between email and mailed surveys, the composite scores from the long form and the short form may not rank providers or groups similarly if the items that comprise the composites are not exactly the same. For example, the access composite in the short form is missing items from the access composite on the long form and would not necessarily rank providers or groups similarly. When any composite is revised whether in the current long form or in a new, short form, this potential impact on ranking must be acknowledged; where possible, a pilot study should be conducted to determine how to adjust for the change in composite composition.

COMPARISON OF RESPONSE

When a survey is changed, it can affect both the response rates (whether sample members choose to respond at all) and the responses (what they say if they do respond). This section looks at what the respondents said in the short form survey as compared to the long form, in email, web, paper, and phone formats. Specifically, it is about their relationship to external factors, the relationship they have to each other, and the statistical behaviors that were described in the section on psychometrics. The most important question that we evaluate in this section is whether the responses were more favorable for the short form or the long form PAS survey.

Within the scope of the long form PAS and the short form pilot survey, we want to be able to figure out whether and how to adjust for email and phone sampling and the amount of time delay. When we compare the PAS to the short form, or to surveys collected in another year, we must be cautious since these comparisons are confounded by differences in protocol, primarily the shorter fielding time for the short form survey.

The following table contains mean scores for items (scored on a scale of 0-100) following various protocols. The response group is labeled by the type of survey, the method of approach, and the mode of response. The approach method noted "Forward" means respondents were sent a paper survey after not responding to the email. In this table, phone responses are treated as nonresponses. The items that were not in the regular PAS are blank, and the bolded cells show where scores are different from PAS Paper Mail (P<.05), or from the SF Paper Mail if not in the PAS. The entries in the table are adjusted using a regression model with available adjusters (age, gender, education, health status, and race). In most instances, the short form means are lower than the long form means, but not always.

Table 7: CA Sample Means of Items by Response Group

Item	PAS Paper Mail	PAS Paper Web	PAS Forward Mail	PAS Forward Web	PAS Email Web	SF Paper Mail	SF Paper Web	SF Forward Mail	SF Forward Web	SF Email Web
N	40096	4018	7826	1107	6486	4163	436	531	79	471
Getting care right away	78.2	79.4	78.6	82.4	78.5	77.9	80.2	78.4	81.2	76.0
Information about getting care after hours						69.2	74.3	73.4	61.7	68.1
Doctor gives clear explanations	90.5	91.6	91.6	92.1	90.9	89.5	90.4	90.2	91.0	89.7
Doctor listens carefully	90.5	90.9	91.5	91.7	90.8	89.8	89.6	89.7	91.3	89.5
Doctor knows medical history	85.8	87.0	86.4	87.1	85.4	85.4	86.4	85.7	89.2	86.2
Doctor spends enough time	88.3	88.9	89.9	90.2	89.0	86.9	87.2	87.1	88.6	86.8
Doctor is informed about specialty care	75.6	76.6	76.2	78.5	74.6	75.2	75.3	74.4	78.5	72.5
Doctor follows up after tests	77.1	76.1	76.0	75.9	73.8	78.8	83.7	77.1	77.7	77.1
Doctor talks about goals						68.5	69.2	69.2	63.0	71.5
Doctor talks about things that make care hard						36.5	41.6	37.0	35.5	35.0
Doctor talks about stress	35.9	36.5	37.7	41.6	36.6	38.7	44.4	37.0	39.9	42.3
Doctor rating	88.2	88.5	89.1	89.1	88.4	87.5	88.2	87.8	89.0	87.1
Willingness to recommend						86.9	87.7	88.2	88.4	88.8

In a model where we control for the timing of the response , there are some additional patterns that show how respondents scored doctors (on a scale of 0-100) using the following response modes. The table below shows regression coefficients for the survey items and it is derived from a regression model that includes indicator variables for short form, being in the email response group (early response to email) and having a good email, plus the adjustment variables and when the response was received. The numbers in the table reflect the difference between having the characteristic and not having it. The bolded numbers are statistically significant at the .05 level.

Table 8: CA Response Patterns of Short-Form vs. PAS Items in Relation to Fielding Data

Item	Short Form vs Long	Email Available vs Not Available	*Good Email Available vs. Bad Email &/or Not Available
Getting care right away	-0.54	-2.06	-0.60
Doctor gives clear explanations	-1.36	-1.99	1.13
Doctor listens carefully	-1.42	-1.64	0.71
Doctor knows medical history	-0.61	-2.19	1.34
Doctor spends enough time	-2.00	-1.34	0.96
Doctor is informed about specialty care	-1.20	-2.97	0.05
Doctor follows up after tests	1.85	-2.21	-0.97
Doctor talks about stress	2.35	-0.54	1.44
Doctor rating	-1.26	-1.23	0.32
Rating of care	-1.98	-1.21	0.52

**"Good Email" = response received from patient via email (no bounce back)*
"Bad Email" = response from patient via paper or web whose email bounced back

In summary, to address our guiding question about responses, the short form did have a small but significant effect on the general level of the responses, most often resulting in lower values. For example, the item "rating of care", when scored 0 to 100 was estimated to be 1.98 points lower on the short form mean than the long form mean, when demographics and doctor are constant. The differences are relatively small, but are consistent and most are statistically significant (bolded) with p values of less than or equal to 0.05 level. In general, we can conclude from this that short form and early email responders tend to be slightly more critical, and patients who maintain a good email with their doctors tend to be slightly more positive. There is more significance seen than one would expect from random sampling. The implication is that if CHPI were to switch from long to short forms, there would need to be small adjustments to account for the form effect and other sampling variables.

[1] Control was necessary because the short form and the regular PAS were kept open for substantially different lengths of time.

MULTIMODAL RESPONSE RESULTS

Although the short form pilot study did not explicitly involve phone, we include an analysis of phone here because it is related to the larger question of what can be done to improve response rates. While this issue deserves more study, the current data can provide some guidance. It is important to understand that the data are not just related to the phone, but rather about the phone as one component in a multimodal strategy.

Analyzing the multi-modal results requires that we make the following substantial assumptions:

- The modes fit together in the way we have described and assumed, namely, that having a good email does not depend on the patient experience and that there are no implications of dropping the phone responses to the PAS.

- The adjustment model is accurate and properly specified in terms of main effects. For example, if respondents 55 and over tend to give more favorable ratings, they do so equally for web, mail, and phone. Without this assumption, comparisons between providers might differ based on the ages of the patients, which would be difficult to account for.

- The answers to the survey reflect the views of the respondent and are not altered by the presentation or the email. The differences in response values across mode represent unspecified differences in the populations responding. This is known to be false for phone but thought to be true for web. That is, phone responses are known to answer the same questions more positively on the phone than by mail or on the web. On the other hand, several studies suggest that respondents to the same question answer the same way by web and by mail because the visual presentation is the same and the anonymity of the response is similar.

- Respondents represent non-respondents, conditional on the variables included in the model. Comparison of response values (Tables 1 and 7) by type of sample warn us that this might not be the case, although other studies of mail non-response suggest that non-response comes from the way that mail is processed in the household, which is age-gender dependent but largely random.

Ultimately, these assumptions are tenuous and we should not expect the emailed or mailed short form to rank and score doctors in the same way that long form phone responses do. We cannot compare short-form phone responses with the long form because the short-form did not use phone responses.

The most important question is whether the phone and mail/email surveys rank doctors in a consistent way, and whether they are more or less efficient. This analysis is conducted within the PAS data only and controls for age, gender, race, general health perceptions, and education, and is similar to the psychometrics. The following analysis is based on a mean for each doctor using the PAS mail/web responses and the phone responses separately, without short-form data.

Table 9: CA Doctor-Level Psychometric Properties of Phone vs. Mail/Web

Item	Overall Mean Mail/Web	Overall Mean Phone	Matched Mean Mail/Web	Matched Mean Phone	N Mail/Web	N Phone	Conver-gence
Getting care right away	78.2	79.4	77.7	79.0	21	24	1.000
Doctor gives clear explanations	90.6	89.7	90.3	89.9	29	53	1.000
Doctor listens carefully	90.5	90.1	90.0	90.2	26	63	1.000
Doctor knows medical history	85.5	85.3	85.0	85.6	27	57	1.000
Doctor spends enough time	88.4	88.0	88.0	88.1	24	51	1.000
Doctor is informed about specialty care	75.4	73.4	74.5	73.3	24	44	1.000
Doctor talks about stress	35.7	32.0	35.1	31.7	32	36	1.000
Doctor follows up after tests	76.1	77.8	75.5	78.3	28	45	1.000
Dating of care	87.4	86.4	87.0	86.5	26	54	1.000
Doctor rating	88.3	87.3	88.0	87.3	19	35	1.000

Table 9 shows that results for the phone surveys rank doctors the same way as results obtained by mail and web (i.e., all of the convergence values are 1. Unfortunately, as seen in Table 10, the reliabilities are lower; the within-doctor SD's are higher and the cross-doctor SDs are lower, so larger samples by phone are required). A second point to note is that the phone interviews were done with live agents, and higher scores were achieved due to the socially desirable response set introduced by the live agents (whom respondents want to please). Interviews conducted strictly by computer, with voice recognition, would be cheaper and might not suffer from this problem. Further experimentation with this mode should be done.

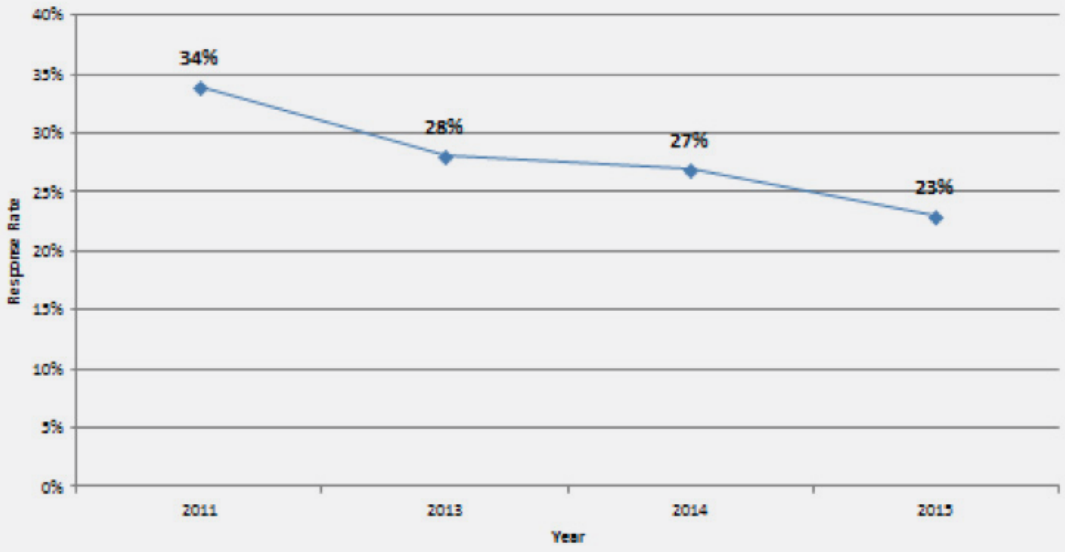
Table 10: CA Reliability Details at the Doctor Level (Phone vs. Mail/Web)

Item	Within Doctor Mail/Web SD	Within Doctor Phone SD	Across Doctor Mail/Web SD	Across Doctors Phone SD
Getting care right away	29.4	30.1	9.9	9.4
Doctor gives clear explanations	20.4	23.4	5.8	4.9
Doctor listens carefully	20.9	23.3	6.4	4.5
Doctor knows medical history	24.6	27.5	7.3	5.6
Doctor spends enough time	22.5	24.9	7.1	5.4
Doctor is informed about specialty care	32.4	37.1	10.2	8.6
Doctor talks about stress	47.6	46.8	12.9	12.1
Doctor follows up after tests	35.1	35.9	10.3	8.2
Rating of Care	16.4	15.4	4.9	3.2
Doctor rating	17.8	18.1	6.3	4.7

MASSACHUSETTS HEALTH QUALITY PARTNERS (MHQP)

Since 2005 MHQP has been fielding a commercial statewide primary care patient experience survey which has become an integral part of the state's healthcare landscape. The survey was first fielded on an every other year basis. However, as health plans and provider organizations began to use survey results for quality improvement, value based reimbursement programs, and to achieve recognition as patient centered medical homes, the survey has been fielded on an annual basis. As previously noted, measuring and understanding patient experience has grown in importance as our health care system adopts patient-centered models of care. Concurrently, we continue to observe a dramatic drop in response to surveys reflecting a trend seen in survey efforts across the country.

Figure 1: MHQP Response Rates: Over Time 2011-2015



Decreasing response rates mean that the costs of administering MHQP's mail based survey have increased as more patients need to be sampled to achieve results that are statistically reliable and appropriate for value based programs and MHQP's public reporting. Our collaborative effort to fund the statewide survey means costs of the survey are covered by health plans and providers. Economies of scale can be realized through this approach and MHQP is committed to keeping system costs as low as possible to reduce the financial burden on survey supporters.

Through this pilot, we learned that patients will respond more frequently to shorter surveys and that the results of a shorter survey are comparable to longer versions. We also found that patients will respond to shorter email surveys in ways that are similar to both long and short form mail surveys.

With regard to email surveys, we found responses rates were higher than expected based on a previous pilot study but somewhat lower than a mail only response in MA or mixed mode in CA.

MHQP SAMPLE FRAME

MHQP employed a number of strategies to recruit provider organizations to participate in the pilot project. We provided details of the project to the MHQP Physician Council, which includes major provider networks in Massachusetts as well as representatives from the Massachusetts Medical Society and Massachusetts Hospital Association. We held meetings with provider organizations that participated in our 2014 statewide PES program and issued a broad release in our PES newsletter that highlighted the opportunity to participate in the pilot. The organizations we targeted were provided with background information and guidelines about privacy, as well as a detailed Q&A document on how the pilot would function in relation to the MHQP statewide PES program.

Four provider organizations elected to participate. Each provided emails for their patients who had primaray care visits between January 1, 2014 and December 31, 2014. The participating provider organizations represent a total population of 1,435 primary care physicians and 248,968 patients with visits during the measurement period. This base was used as the pilot sample frame and it represented approximately 27% of commercial patients with visits from plans submitting data for the statewide project. The five health plans participating in the statewide survey represent approximately 87% of commercial health plan membership in Massachusetts.

To maintain consistency in sampling patients for both the long form and the short form arms of the pilot, the health plan sample frame was used as the basis for pilot sampling. The sample frame included HMO, POS or PPO adult members (18 and older as of December 31, 2014) of participating health plans with visits during the measurement period (January 1 – December 31, 2014). Plan members were attributed to providers by the plans submitting data, based on health plan claims visit data and member information. MHQP's survey research firm, the Center for the Study of Services (CSS), aggregated health plan supplied data and matched provider organization supplied emails to health plan data.

To facilitate data matching, provider organizations supplied CSS with limited additional patient information needed to accurately match patients with emails with the plan provided data. Records were matched using provider NPI, patient zip code, patient date of birth or year of birth, patient name using either of two constructed name fields to match. Valid email addresses were then standardized and patients associated with more than one email or more than one provider and multiple patients associated with the same email address were taken out of the sample frame.

MHQP STUDY DESIGN AND DATA COLLECTION

As noted, the short form pilot was fielded in parallel to MHQP's statewide survey. Trended statewide survey results are used for high stakes reimbursement programs and public reporting. Therefore, to maintain the integrity of those high profile programs, sampling was prioritized and samples for the pilot sample were drawn after the statewide long form survey sample was drawn using a stratified sampling plan. Three pilot test samples were drawn:

1. MHQP's statewide long form survey is designed for practice level results, therefore to obtain comparable short form results a sample was drawn at the practice level.

2. MHQP has historical data to support analysis at the provider level. Therefore, we drew additional provider level samples to expand our analyses.

3. In order to learn more about email response, supplemental email samples were drawn from the practices included in the practice level-sample, and from the providers included in the provider level-sample.

With four participating provider organizations representing 106 individual practice sites MHQP began the fielding process for the pilot on May 13, 2015. The total number of patients surveyed as a part of this study was 36,018. These patients were seen by 348 unique physicians. The survey asked patients to evaluate the care they received during the previous 12 months.

Surveys were administered in one of three ways:

1) short form email only,
2) short form mail only, and
3) long form mail only.

As previously stated, the long form statewide practice site survey samples were drawn before the pilot samples and short form surveys were fielded shortly after the traditional statewide survey. Patients invited to respond through short form email were provided with a link and taken directly to the online survey in either a desktop or mobile format as appropriate. For both the long form and short form mail samples, patients were offered the option of completing the survey online and given an individual pass code to access the survey. The short form email and mail surveys were sent under the name of the provider organization while the matched samples from the long form survey received mail invitations from their health plan.

For the MA short form email pilot, a three wave email protocol was used. Sample members were sent second and third waves of email in one week intervals if they did not respond to the former waves.
Both the long form and the short form mail protocol employed the same two wave fielding protocol, with the second wave occurring three weeks after the first. The analysis of results combines the long form mail and web results to compare them to the short form mail, web and email results.

MHQP YIELD RATE

In conducting this pilot we wanted to better understand the impact of survey length and administration mode on yield rates (usable responses/ surveys attempted). Unlike CHPI, MHQP has not yet introduced email surveys for its statewide survey. However, completing a survey online via the web has been an option and offered through the cover letter included with mailed materials. Respondents are asked to use the web address supplied in the survey cover letter and given an individual pass code to access the survey.

The proportion of adult long form survey responses completed using the web option has risen steadily but very slowly over time, from 10.0% in 2013 to 11.3% in 2014 and 11.9% in 2015, and has not offset the more dramatic decline in mail response. The overall yield rate for adult long form surveys was 24.45% in 2015. While 11.9% of completed surveys in 2015 were done via the web, usable responses from the web as a proportion of all surveys sent, i.e., the yield rate, for web responses was only 2.9%.

Short form mail sample members also had the option of completing the survey through the web and the yield rate was 26.7% by mail and 3.0% by web, for a total yield rate of 29.7%, which is better than the long form PES (P<.05) fielded under the same mail protocol. For the MA short form email survey, the yield rate for response was 21.5%.

Results in Table 1 illustrate differences in yield rates by survey length and mode. Overall yield rates for email would likely improve in MA if non-respondents had a follow-up mail protocol, as they did in CA.

Table 1: MA Yield Rate and Demographics by Survey Mode

Survey Protocol	N	Response Type	% Yield	Average Age	% Male	% Female	% College	% White	% Hispanic
Long Form Mail	177685	Mail	21.6	54.7	41.9	58.1	51.4	87.2	3.3
Long Form Mail	177685	Web	2.9	48.5	52.7	47.3	68.7	85.2	3.1
Long Form Mail	177685	No Response		45.6	46.8	53.2			
Long Form Total			24.5						
Short Form Email	12303	Web	21.5	53.9	36.9	63.1	51.7	88.0	1.9
Short Form Email	12303	No Response		48.4	43.6	56.4			
Short Form Email Total			21.5						
Short Form Mail	23715	Mail	26.7	54.4	45.9	54.1	45.8	89.7	2.3
Short Form Mail	23715	Web	3.0	50.2	55.9	44.1	62.2	89.7	1.6
Short Form Mail	23715	No response		46.2	50.0	50.0			
Short Form Mail Total			29.7						

PSYCHOMETRICS OF RESPONSES

SURVEY MODE

A key objective of this pilot study was to determine if the rankings of doctors will be comparable 1) between the short form and the long form and 2) by survey mode (email versus paper) for short form only.

There are essential psychometric issues that are considered when evaluating results for the short form as compared to the long form: 1) reliability and 2) concordance or conceptual agreement with the current long form survey and composite measures. As noted in the CHPI section, the analyses conducted for CA and MA were not identical due to differences in the survey administration and/or response rates. In Massachusetts, we had sufficient email responses so that we could do email and web response psychometric analyses as well as short versus long form analyses. The first two analyses that follow relate to the variation in results when responses are received via web versus mail and variation in results when surveys are sent by email versus paper delivery.

Reliability is a statistical measure that indicates how accurately a measure captures information by measuring the consistency of the information provided by respondents. Reliability can be expressed in a number of ways. It depends on survey context (i.e., the entities being compared); for this study, the context examined is a particular set of doctors or practices that have been sampled with a given instrument. Because more responses improve our ability to calculate consistency, reliability also relies upon the sample size and completed responses received, noted in results analysis as "N". In our analysis, we evaluated how many responses are needed to achieve reliability levels that are acceptable for high stakes use of results. A reliability level of 70% is generally accepted as appropriate for high stakes use.

Observed results for doctors or practices vary for two reasons. One of these is sampling error, which in statistics is captured by the amount that patients within the doctor or practice vary, and the size of the sample. Variation within a doctor or practice is measured by the within-entity standard deviation (SD). The other reason that observed results vary is because the doctors or practices in the particular context are truly different from each other. The true differences are captured through a statistical estimation process known as variance components analysis, and it is expressed as a standard deviation which we label the context SD. The formula used to evaluate reliability within context is:

R = (context SD)2 / ((context SD)2 + ((within-entity SD)2/sample size)).

The square of the standard deviation is given the technical name "variance" in statistics; in these terms the reliability is the proportion of observed variance that can be credited to true differences between doctors or practices. The within-entity (e.g., within doctor) variance depends on the survey question, the available responses, and the amount of agreement among the responses for given entities.

The values in the following table were the values observed in the matched sample from the long form statewide survey and the short form pilot for mail responses and web responses.

Table 2: MA Reliability Details at the Doctor Level

	Within Doctor Standard Deviation		Doctor Level Standard Deviation	
Item	Mail	Web	Mail	Web
Getting care right away	21.8	22.0	5.5	5.9
Information about care after hours	41.5	42.9	8.2	8.5
Doctor gives clear explanations	14.6	14.0	2.8	3.3
Doctor listens carefully	15.6	15.9	3.3	3.7
Doctor knows medical history	18.5	18.1	4.6	4.2
Doctor spends enough time	17.1	17.5	3.6	3.9
Doctor is informed about specialty care	25.8	25.6	5.6	5.9
Doctor follows up after tests	25.9	26.5	6.9	7.8
Doctor talks about goals	46.4	47.0	9.0	9.9
Doctor talks about things that make care hard	49.0	49.4	8.9	9.4
Doctor talks about stress	48.3	48.1	10.5	10.5
Doctor rating	13.7	13.8	4.0	4.3
Willingness to recommend	19.5	18.7	4.8	5.2

Typically, as seen in Table 2 above, the within-doctor SD went up slightly (which would lower reliability), but the cross-doctor SD went up significantly, raising reliability. This finding is most easily understood in terms of the N size required for 70% reliability, as shown in Table 3 below.

Table 3: MA Doctor-Level Psychometric Properties of Response by Web vs. Mail

	Long Form Mean		Short Form Mean		Required N		
Item	Mail	Web	Mail	Web	Mail	Web	Convergence
Getting care right away	87.8	88.2	87.7	87.8	38	33	1.000
Information about care after hours	77.7	75.8	77.5	75.5	60	59	0.840
Doctor gives clear explanations	94.7	95.3	94.6	95.0	62	44	0.957
Doctor listens carefully	94.3	94.3	94.3	94.0	53	44	0.944
Doctor knows medical history	91.5	91.8	91.2	91.3	39	44	0.918
Doctor spends enough time	92.9	93.0	92.8	92.8	53	47	0.982
Doctor is Informed about specialty care	83.9	84.0	83.8	83.3	50	45	1.000
Doctor follows up after tests	87.5	87.1	87.3	86.7	34	28	0.900
Doctor talks about goals	67.7	65.4	67.4	65.0	62	53	1.000
Doctor talks about things that make care hard	41.7	42.9	41.7	43.4	72	65	0.943
Doctor talks about stress	62.0	63.4	62.2	63.8	49	49	1.000
Doctor rating	91.5	91.1	91.3	90.5	28	25	1.000
Willingness to recommend	91.1	91.4	90.8	90.7	38	30	1.000

Table 3 also compares Massachusetts mail responses with web responses, understanding that web respondents include those approached by email and those approached by mail. In this comparison, results are controlled for demographics including age, gender, race/ethnicity, general health, education, and the type of form (long form vs. short form).

To compare differences in long form versus short form results, our evaluation matches results for both versions at the doctor level. The matched dataset only uses doctors who are in both. It is worth noting that doctor standard deviations in the full long form survey response set were similar to the results seen in the matched data above.

The convergence statistic seen in Table 3 is a measurement of how well two concepts being measured would correlate if there were no sampling variaton. In this table, the two concepts come from the two groups being compared, i.e., the mean values for mail responses and web responses. The correlation is across the results for doctors. Since conversion is a correlation coefficient, +1 represents perfect agreement, 0 is no agreement, and -1 is perfect inverse agreement.

As in reliability, convergence depends on the concept of a "true" doctor or practice mean, which provides the "context" of the study. The two global averages are shown as the "matched averages" and the convergence is the correlation. Results in Table 3 illustrate that the convergence statistics between the mail response to the survey and the web response to the survey are all high; some are as high as 100%; one measure (Information about care after hours) is 0.84, which is still high but has less favorable convergence. The sample N required to achieve 70% reliability for both web and mail shows that except for the item asking about the doctor's knowledge of a patient's complete history, the minimum sample sizes needed for 70% reliability decreased or stayed the same with web responses.

These findings suggest: 1) to a large extent web and mail responses provide similar information about providers; 2) the means of the variables (both for the statewide long form and the matched short form survey) are also fairly comparable; 3) the web is somewhat more efficient because it better distinguishes performance of individual providers in the matched short form sample. That is, the cross-doctor standard deviation for web responders is larger than the cross-doctor standard deviation for mail responders; and 4) the N sizes required to make 70% reliability are therefore smaller, but not small enough to reduce sample sizes and achieve cost savings.

The following table looks at responses based on having received a paper survey vs. an email survey. As previously noted, a long form email survey was not fielded in MA.

Table 4: Doctor-Level Psychometric Properties of Receiving Paper vs. Email (short form only)

	Full Datasets		Matched Datasets				
Item	Paper Mean	Email Mean	Paper Mean	Email Mean	Required N Paper	Required N Email	Convergence
Getting care right away	89.3	87.3	88.0	87.1	33	36	1.000
Information about care after hours	78.9	74.7	76.1	73.5	52	78	0.666
Doctor gives clear explanations	94.6	95.4	94.0	95.3	28	86	1.000
Doctor listens carefully	94.4	95.1	94.6	95.0	55	63	1.000
Doctor knows medical history	92.4	93.3	92.5	92.8	61	60	1.000
Doctor spends enough time	92.8	93.3	92.5	92.6	48	39	1.000
Doctor is informed about specialty care	85.0	86.2	85.6	84.7	56	51	1.000
Doctor follows up after tests	89.2	87.7	88.6	86.7	42	36	1.000
Doctor Talks About Goals	67.7	68.4	65.7	67.4	63	224	1.000
Doctor talks about things that make care hard	41.6	43.0	40.7	42.9	102	144	1.000
Doctor talks about stress	57.8	59.8	56.9	57.2	75	59	1.000
Doctor rating	91.9	92.6	91.7	92.1	29	31	1.000
Willingness to recommend	91.6	92.7	91.7	92.4	33	46	1.000

Table 4 suggests similar findings to the web versus mail results depicted in Table 3. To a large extent, responses for those who were sent a paper survey compared to those who were sent an email survey provide similar information about doctors as the analysis of mail versus web responses. The means of the variables (for short form only) for paper vs. email and the matched paper vs. email are also fairly comparable; with the emailed survey being somewhat more efficient than the paper survey because it better distinguishes performance of individual providers in the matched short form sample. Finally the N sizes required to make 70% reliability are more variable than in the web versus mail comparison, given the smaller sample size for this analysis. In some cases the N needed would be smaller, but not in all.

SURVEY LENGTH

After confirming that results were similar whether responses were received through mail or through the web, and whether sent by email or paper, analysis proceeded to test short vs. long form. In this test, the controls include the same set of demographics as used in the previous analyses above and we have added a control for web response vs. mail response (as a proxy for email approach vs. mail).

Table 5: MA Reliability Details at the Doctor Level for Long Form vs Short Form

	Within Doctor Standard Deviation		Doctor Level Standard Deviation	
Item	Long Form	Short Form	Long Form	Short Form
Getting care right away	21.9	20.9	5.8	6.0
Information about care after hours	41.5	41.4	6.8	7.1
Doctor gives clear explanations	14.5	14.1	2.5	2.8
Doctor listens carefully	15.7	14.9	3.0	3.2
Doctor knows medical history	18.6	16.8	3.4	3.6
Doctor spends enough time	17.1	16.6	3.2	3.4
Doctor is informed about specialty care	25.5	25.3	4.8	5.9
Doctor follows up after tests	26.1	24.3	6.3	6.9
Doctor talks about goals	46.1	46.4	7.3	7.2
Doctor talks about things that make care hard	49.0	49.2	7.8	7.4
Doctor talks about stress	48.0	49.1	8.8	8.5
Doctor rating	13.7	12.8	3.6	3.5
Willingness to recommend	19.6	17.8	4.3	4.3

Again, matched datasets using only results for doctors in both surveys are required because the full statewide survey is a different context than the short form. As seen in Table 5 above, doctor standard deviations in the long form were similar to the matched short form data.

The within-doctor SD stayed the same or decreased (which raises reliability), and the cross doctor SD increased significantly, raising reliability for most measures but not the measures related to soliciting information about the patient's goals, abilities and stress levels.

Table 6: MA Doctor-Level Psychometric Properties of Long Form vs. Short Form

	Full Datasets		Matched Datasets				
Item	Long Form Mean	Short Form Mean	Long Form Mean	Short Form Mean	Required N Long Form	Required N Short Form	Convergence
Getting care right away	87.7	88.8	87.3	87.6	33	29	1.000
Information about care after hours	77.3	78.1	76.6	77.4	88	79	0.855
Doctor gives clear explanations	94.8	94.7	95.2	94.4	82	61	1.000
Doctor listens carefully	94.3	94.5	94.5	94.1	65	50	1.000
Doctor knows medical history	91.4	92.4	91.8	92.0	73	51	1.000
Doctor spends enough time	93.0	92.8	92.9	92.4	66	55	1.000
Doctor is informed about specialty care	83.7	85.1	83.9	84.4	67	43	1.000
Doctor follows up after tests	87.3	88.7	87.2	87.6	41	29	1.000
Doctor talks about goals	67.4	67.3	66.5	67.0	93	98	1.000
Doctor talks about things that make care hard	42.1	41.2	40.2	41.1	94	105	1.000
Doctor talks about stress	63.2	58.1	60.7	57.8	70	78	1.000
Doctor rating	91.4	92.0	91.6	91.4	34	31	1.000
Willingness to recommend	91.0	91.8	91.5	91.2	48	40	1.000

Results in Table 6 indicate that relative to the statewide long form survey, short forms are slightly more efficient as a lower number of responses are needed to get reliable information with the exception of survey questions related to providers engaging patients to talk about their care where sample sizes need to be larger. It is not clear why this difference occurs for these questions but these findings suggest an unknown interaction that should be explored further.

Convergence is also very high for almost all measures except for the question about patients having information about care on weekends. Differences in results for this question may be due to the fact that the long form survey includes more questions about access and that may change the context of the question for respondents.

An earlier and similar CA study in 2014 took a more radical approach to sampling, using outreach to Consumer Reports subscribers to drive response. In that study, convergence statistics were significantly lower than the lowest finding for this pilot. This difference in study outcomes suggests that using a similar method of sampling for both the short-form experiment and the long form was a major factor in achieving convergence in the short form pilot study.

MHQP's statewide survey is designed to be fielded and reported at the practice level. The doctor-level results are more precise than the practice results, but due to the importance of practice measurement in MA, we included an analysis of results at the practice level.

These convergences are very high, again except for the question regarding evening and weekend information. Larger practice standard deviations in Table 7 result in lower required sample sizes as seen in Table 8.

It is important to note that this analysis does not control for practice size. The number of completed surveys required is in terms of the actual number of responses to a survey question, so actual sample sizes required for 70% reliability will be higher.

Table 7: MA Reliability Details at the Practice Level

	Within Site Standard Deviation		Site Level Standard Deviation	
Item	Long Form	Short Form	Long Form	Short Form
Getting care right away	22.0	21.1	6.2	6.3
Information about care after hours	41.7	41.6	6.0	6.7
Doctor gives clear explanations	14.5	14.2	2.1	2.2
Doctor listens carefully	15.7	15.1	2.4	2.7
Doctor knows medical history	18.8	16.9	2.8	2.9
Doctor spends enough time	17.2	16.8	2.7	2.8
Doctor is informed about specialty care	25.8	25.5	4.1	5.0
Doctor follows up after tests	26.3	24.6	6.4	6.8
Doctor talks about goals	46.5	46.5	6.2	7.0
Doctor talks about things that make care hard	49.0	49.2	5.6	6.3
Doctor talks about stress	48.0	49.1	6.4	7.2
Doctor rating	13.9	13.0	3.4	3.1
Willingness to recommend	19.7	17.9	3.9	3.7

Table 8: MA Practice-Level Psychometric Properties of Long Form vs. Short Form

	Full Datasets				Matched Datasets				
Item	Long Form Mean	Short Form Mean	Long Form % Missing	Short Form % Missing	Long Form Mean	Short Form Mean	Required N Long Form	Required N Short Form	Convergence
Getting care right away	87.7	88.6	45.6	48.2	87.8	88.2	30	27	1.000
Information about care after hours	77.3	78.0	2.1	3.1	76.8	78.0	115	90	0.861
Doctor gives clear explanations	94.8	94.7	1.0	1.1	95.0	94.4	116	101	1.000
Doctor listens carefully	94.3	94.4	1.0	1.1	94.2	93.9	100	74	0.962
Doctor knows medical history	91.4	92.2	1.3	1.3	91.4	91.8	108	78	1.000
Doctor spends enough time	92.9	92.7	1.2	1.3	92.8	92.4	96	83	1.000
Doctor is informed about specialty care	83.7	84.9	38.7	39.4	83.3	84.1	93	60	1.000
Doctor follows up after tests	87.2	88.6	15.6	9.4	86.7	87.7	40	31	0.981
Doctor talks about goals	67.4	67.3	1.9	1.6	66.3	67.1	133	103	0.934
Doctor talks about things that make care hard	42.0	41.3	2.8	2.4	40.2	41.4	178	143	0.910
Doctor talks about stress	63.1	58.2	2.1	2.6	60.6	57.9	130	110	1.000
Doctor rating	91.4	91.9	1.4	2.5	91.3	91.3	40	42	0.997
Willingness to recommend	91.0	91.7	1.4	2.1	91.1	91.0	61	56	1.000

COMPOSITES

As noted in the California analysis section, there are some potential composites in the short form. The following MAP results derive five composites that were initially hypothesized for Massachusetts.

Table 9: MA Potential Composite Measures

Composite	Content	Alpha	% Scaling Success
Communication	• Doctor gives clear explanations • Doctor listens carefully • Doctor knows medical history • Doctor spends enough time	0.88	100
Access	• Getting care right away • Information about care after hours	0.26	88
Talk	• Doctor talks about stress • Doctor talks about goals • Doctor talks about things that make care hard	0.65	100
Coordination	• Doctor is informed about specialty care • Doctor follows up after tests	0.57	75
Summary	• Doctor rating • Willingness to recommend	0.84	100

In this table, Alpha refers to Cronbach's alpha which is a measure of the internal consistency of the items. Scaling success is the percentage of possible item reassignments where the hypothesized (listed) composite assignment is better than an alternative assignment. The composite analysis considers how well aligned each item is to its own composite and all of the other composites. With respect to a particular alternative composite, if the item is closer to its own composite, that is considered a scaling success; if closer to the alternative composite, it is a scaling failure. The percent scaling success is the percentage of successes across all of the items and all of the other possible composites these items might have been assigned to.

Three of the five composites hypothesized worked extremely well. The items in the communication composite are highly correlated in the short form similarly to what we find in the in the long form. Creating a new "engagement"/"talk" composite based on measures that ask about whether a doctor (or doctor's office) talked with a patient about care is a logical first step toward measuring whether doctors and patients interact. All three items involve a two-way communication between the provider and the patient and elicit information from the patient. Further work needs to be done to better understand the quality of the interaction.

The access composite for the short form fails however, and there are no other possible items in the survey to group with "getting care right away". "Information about getting care after hours" is an item that predicts the summary scores, but has no appropriate companion in this structure. An access composite that has a higher percentage scaling success could be created, using additional different questions about access. More analysis is needed to determine what items would need to be included in the survey to form a strong access composite. Similarly, the items in the care coordination composition did not align. Consideration should be given to reporting these two items separately and new items that might better align to assess care coordination should be considered.

While we have seen that scores on individual items are comparable in the short form and the long form and between email and mailed surveys, the composite scores from the long form and the short form will not in most cases rank providers or groups similarly if the items in the composites have changed. For example, the access composite is missing items from the access composite on the long form and would not rank providers or groups similarly. When any composite is revised whether in the current long form or in a new, short form, this potential impact on ranking must be acknowledged; where possible, a pilot study should be conducted to determine how to adjust for the change in composite composition.

COMPARISON OF RESPONSE

When a survey is changed, it can affect both the response rates (whether sample members choose to respond at all) and the responses (that is, what they say if they do respond). This section is about what the respondents say. In particular, it is about their relationship to external factors; the relationship they have to each other and their statistical behaviors as described under psychometrics. The most important question in this section is whether or not the short form made the responses more or less favorable.

RELATIONSHIP OF RESPONSES TO SURVEY METHODS

The following is a table of item means (scored on a scale of 0-100) broken down by the combination of possible response protocols, including type of survey, approach method, and response mode. The results are adjusted for age, gender, race/ethnicity, general health, and education.

Table 10: MA Sample Means of Items by Response Group

Item	PES Paper Mail	PES Paper Web	SF Paper Mail	SF Paper Web	SF Email Web
N	35426	4807	5963	668	2521
Getting care right away	87.7	89.3	87.9	89.4	87.2
Information about care after hours	77.5	77.4	78.1	80.4	75.4
Doctor gives clear explanations	94.9	95.6	93.9	93.7	95.0
Doctor listens carefully	94.4	94.6	93.9	94.0	94.6
Doctor knows medical history	91.5	92.1	91.5	92.9	92.4
Doctor spends enough time	93.0	93.5	92.4	92.8	92.9
Doctor is informed about specialty care	83.7	84.7	84.3	84.9	84.6
Doctor follows up after tests	87.4	87.8	87.9	88.8	87.6
Doctor talks about goals	67.4	66.3	67.6	72.1	68.5
Doctor talks about things that make care hard	41.3	43.3	42.5	50.3	44.6
Doctor talks about stress	62.7	63.1	58.5	70.1	61.7
Doctor rating	91.6	91.1	91.1	91.1	91.8
Willingness to recommend	91.2	91.5	90.5	91.4	91.6

Although differences with the long form mail group are sometimes significant due to the large sample sizes, the only large differences in response means are associated with short form web responses to questions related to engaging patients in talking about their care (Doctor Talks About Goals, Doctor Talks About Things That Make Care Hard, and Doctor Talks About Stress). These differences are only partially confirmed by California data, but they suggest that some respondents might have had difficulty with the web short form for these questions. More research is needed to understand this difference.

Table 11 shows the coefficients of a regression equation for scored items (scale is 0-100) where the independent variables describe the method of approach, the response mode, and the type of survey instrument in a combined sample that includes short form and long form and the control variables available from the short form: age, gender, race/ethnicity, general health, and education. The results are also controlled for provider ID, meaning that the comparisons are based on within-provider data only. For example, the comparison of email vs. paper mail outreach is based on comparing respondents who had email and paper mail within each provider, and then summing up those comparisons across all providers.

Table 11: MA Coefficients of Key Variables Predicting Item Mean

Item	Email vs mail	Group vs Plan	Web vs Mail	Short Form vs Long form
Getting care right away	-1.4	1.0	0.8	-0.1
Information about care after hours	**-2.7**	-2.4	-1.2	1.3
Doctor gives clear explanations	0.9	0.3	0.3	**-1.1**
Doctor listens carefully	**1.0**	0.5	-0.4	**-0.6**
Doctor knows medical history	0.8	0.1	-0.0	-0.1
Doctor spends enough time	0.5	0.3	-0.1	**-0.7**
Doctor is informed about specialty care	-0.0	-0.4	-0.4	0.7
Doctor follows up after tests	-0.2	0.6	-0.4	0.4
Doctor talks about goals	2.6	-0.6	-2.7	0.6
Doctor talks about things that make care hard	-0.9	-0.6	2.2	1.8
Doctor talks about stress	-0.1	-2.2	**2.0**	**-2.7**
Doctor rating	**1.6**	**0.9**	**-1.1**	**-0.6**
Willingness to recommend	**1.2**	0.7	-0.2	-0.7

Note: bolded coefficients are significant at the .05 level.

We established in previous tables that doctors and practices would be ranked in similar ways by type of response (web vs. mail), approach method (email or paper survey) and type of instrument (long form and short form). Here, we are examining whether results would go up or down in unison depending on these same factors. If the rankings are the same, but the results go up or down in unison, then comparative analyses can be made. However, trending of results will need to include adjustments so that we do not, for example, suggest that providers' performance has gotten worse when in fact their performance has improved at the time we changed survey methods.

Table 11 uses multiple regression analysis to examine all types of differences simultaneously. The table shows the regression coefficients, which are the differences in mean score attributable to the methods differences described in the header. For example, the difference in the mean score for how well doctors listen increased by 1 point simply due to having received an email survey rather than a mail one. There are some significant differences in the table, more than one would expect at random, but in general they are small. In contrast to previous studies of web vs. mail which have found no differences, there was a slight shift in the relative response by web compared to mail for a number of variables, but without a consistent sign. There were significant differences for the email vs. paper survey sample frame, mostly indicating higher mean response by email. That is, if one were to use email alone, one would expect to see an upward trend in all results due to the method shift. For example, a 1.2 point improvement on the 0-100 provider rating question could be attributed to methods effect. There were fewer differences for short form vs. long. The short form surveys seem to have a different result for the question that asks about the doctor talking about stress that was traced to an anomaly in the short form web administration.

Since the differences are small and of mixed sign, we can feel confident in making these "consistent" (no change in item wording or response choices) changes to the survey length. For practices or plans that might be very sensitive to small shifts up or down, these pilot results could be used to adjust trend scores if this particular short form is adopted. If a similar survey, such as CG-CAHPS 3.0, were adopted, one would not expect to see many changes, but the case mix might differ, and this could affect the item trends.

COMBINED ENTITY ANALYSIS

A primary aim of the pilot study was to test a shorter survey to determine if it gives comparable answers and ranks providers in a comparable way to existing statewide surveys. The demographic characteristics of the two regions differed significantly and substantially. California's sample was more racially and ethnically diverse, less educated and had more females. Differences in demographics were controlled for within each state's analysis so that comparisons could be made.

SHORT FORM ANALYSES

Our hypotheses were: 1) the short form will have better response rates than long forms within each state; 2) the short form and long form will produce comparable distribution of responses in both California and Massachusetts at the provider level; and 3) the short form and long form will rank doctors (in both states), practices (in MA) and medical groups (in CA) in a similar way.

Table 1: Massachusetts and California Short Form Results

Short Forms	Massachusetts	California
Response rate	Short form higher than long form (same time in field) (26.9% v.24.5%)	Short form same response as long form (responses arrived quicker but short form was fielded a shorter time than long form) (25.4% v. 26.3%)
Results	Overall findings same for short and long forms with small exceptions at the item level*	Overall findings same for short and long forms with small exceptions at the item level*
Ranking of providers	Same for short and long form at the item level*	Same for short and long form at the item level*

*** This study did not test comparability of results or ranking at the composite level.**

In addition to the high level results in the table above, the data provides several interesting findings:

- In Massachusetts, where the short and long forms were kept in the field for a similar amount of time, mail response rates were substantially higher for the short form. In California, short form responses were returned more quickly than the long form, however, the total response rates were the same. It should be noted that the short form fielding time was little more than half that of the long form, and therefore, it is possible that had the short form been in the field longer, California's response rates would have been higher for the short form.

- In general, both states' responses were comparable at the provider level regardless of survey length although there were some small differences according to the mode of response, particularly with regard to talking about stress in MA where mean scores were higher with the short form web responses.

- Doctors, practices (MA), and medical groups (CA) were almost universally ranked similarly in both states. All of the convergences (agreement of "true" answers) in both states were high, and in fact higher than in other pilots of this type.

EMAIL ANALYSES

Secondly, we wanted to test email approaches to surveying. Our hypotheses for this arm of the pilot were: 1) short form email response rates will be better than long form email response rates; 2) emailed short forms will give comparable responses to emailed long forms and mailed short forms; 3) emailed forms will rank providers in a comparable way to mailed forms that have similar content; and 4) patients in the sample frame who have emails will have different demographics than patients who do not have emails.

Table 2: Massachusetts and Californian Email Results

Email results	Massachusetts	California
Response rates (RR) emailed vs. mailed short form only	Short form email RR was not as high as short form mail (21.5% vs. 29.7%)	Short form email with mixed mode follow-up RR was higher than short form mail (33.6% vs. 24.1%
Response rates emailed short vs. emailed long form	Not tested in MA	No significant difference
Comparability of results	Comparable to both long and short form mail survey.	Comparable to both long and short form mail survey.
Email population demographics vs. mailed population	Slightly older and female	Slightly older and female

Email results in the table above were interesting to us for the following reasons:

- In MA, email yield rates (usable response/surveys attempted) were higher than expected (21.5%) but not as high as mail yield rates (29.7%). They were also not as high as the total mixed mode yield in CA which was 33.6%. This is comprised of a yield of 9.8% for email responses, 21.2% for follow up with mailed paper surveys, and 2.4% for web responses using the link from the paper survey.

- There was not an appreciable increase in email yield rates associated with the short form in CA where a comparison to long form email could be made.

- Email short forms did provide comparable results to both long and short form mail surveys in both states.

- In both states, the demographics of those with email have shifted from what they were in previous studies. Patients with email now tend to be slightly older and more female, whereas in earlier email studies, they were younger and overwhelmingly male.

In addition to testing the hypotheses related to our key questions, we used the data collected through this pilot to do additional analyses and these results are included in our detailed results. Based on adjusted results, we looked at the relationship of responses by survey approach (long form vs. short form and email vs. mail) and response mode (mail vs. online). Differences found at this level were small. However, for organizations in MA and CA that are sensitive to small changes that could affect financial incentives, these pilot results could be used to adjust for trending.

We also considered different ways to summarize results of the short form survey through question groups or composites. Grouped communication questions and grouped questions about providers engaging with patients in talking about care showed high internal consistency among the items and fit together well. Two questions related to care coordination and two questions related to access did not perform as well when grouped together as composite measures. Additional questions could improve the access composite but the care coordination questions are best kept separate.

Lastly, in response to the growing interest in patient narratives about services, we included open-ended questions that were fielded in both states' electronic short form pilots to test the feasibility of eliciting narrative responses from patients. We chose to test two different sets of open-ended questions with respondents randomly assigned to receive either a three-item elicitation or a five-item elicitation. Preliminary findings indicate that response rates were similar for both sets of questions. However those receiving the five-question protocol gave longer responses than did those receiving the three-question protocol. There were differences in response rates by patient characteristics in both MA and CA samples, but none of these varied significantly between the two sets of questions. In general, non-respondents were younger, less educated, and more often identified as Asian. There were no differences in length of narrative (i.e., word count) across patient characteristics within the CHPI data, but within the MHQP data, younger adults and women gave longer responses. Responses were largely positive and correlated with responses to the close-ended items.

CONCLUSIONS

The results of this pilot indicate that new and more innovative approaches to surveying are evolving and promising for efforts to reduce survey costs and burden. A move in this direction will help make high value performance information available to the public. Along with the promise of new approaches, we encountered the reality of the current state of patient information systems and the limited experience many physician organizations have in engaging their patients in new ways. We found that full implementation of the email modes of survey is most hindered by the lack of systematic collection, verification, and maintainance of patient emails. With this current state, email surveying is not a viable survey option by itself. However, the results in both states, including a response rate of almost 40% for phone responses to the PAS long form survey in California, suggest that multi-mode surveys that reach patients in numerous ways are the best option. Indeed, multi-mode surveys must be fielded to achieve results that are reliable enough for high stakes use. We do believe that if systematic collection of patient contact information is widely adopted, and the concerns about privacy protection are addressed for patients and providers, the ability to survey through electronic modes can improve substantially.

With the constraints noted, the generally positive results we achieved in testing our hypotheses through this pilot are strengthened by the fact that overall, similar results were found in two regions with significantly different health systems and patient characteristics. Contrary to prevailing wisdom, the Massachusetts response rates were better for a short form survey than they were for the longer 61-question survey, and California might have had higher response rates had the fielding period been longer. In addition, the short form survey instrument appears to be a viable alternative to longer form surveys with the finding that that relative scores for doctors, practices, and groups to the same question are generally comparable for the short form based on convergence statistics with case mix control. Further, response to the email short form survey produced results that are comparable to long form surveys and mail short form surveys.

It will be important for organizations that are considering whether to use a short form to understand that the composite scores from the current long form and the new short form may or may not rank providers, practices or groups similarly depending on the items included in the respective composites. In this study, we were not able to test our composites to determine if they score and rank providers the same. However, as the items that comprise the composites rank providers the same, the new short form composite rankings will be valid. If results for some composites are used for high stakes purposes, it would be prudent to do one of the following:

- Keep the composites construction stable until other changes are made to the instrument;
- Remove items that are not in the short form composites from the previous year's long form survey composites and run an analysis to see if the composite results, now having identical items, are comparable (note that these should be comparable since the items from both years are convergent);
- Or complete a small pilot project using the old and new composites so that adjustments can be made if needed.

Our detailed analysis of adjusted results found small differences in responses by survey approach (long form vs. short form and email vs. mail) and response mode (mail vs. online) that can be adjusted for in CA and MA for provider organizations that might be sensitive to small changes as they relate to financial incentives. Other markets can use the same adjustments noted in this report as they are likely to follow the same pattern as in CA and MA. However, if stakes are sufficiently high in a given market, a small test with 2500 respondents using the new survey prospectively (to compare with a large fielding of the old survey) or the old survey restrospectively (to compare with a large fielding of the new shorter survey) would be suggested so that local adjustment factors can be determined.

With our short form pilot and the release of CG-CAHPS 3.0, we are headed toward the broader use of shorter surveys that will be less expensive to mail and less burdensome to patients. If these shorter surveys are fielded with a scientific sampling frame and a full multi-modal survey approach including email, they will yield comparable information to current surveys with better response rates. Altogether, these steps should increase confidence in the ability to achieve reliable results for a lower price.

Finally, given the numerous avenues consumers have available to them to comment on a variety of products and services they receive, including health care services, it is important that a systematic way to elicit and analyze such comments be designed. The development of nationally recognized protocols for collecting and reporting narratives is necessary to increase provider acceptance and the likelihood of effective quality improvement efforts. The public reporting of such comments on health care quality websites could help consumers make more informed choices and potentially increase their interest in engaging with their providers in conversations about how they would like to receive care. The work we have done in this pilot is a first step in the direction of systematic collection and reporting. We hope to continue this work and advance the field further in the coming months and years.

NEXT STEPS

The Short Form Patient Experience pilot has provided CHPI and MHQP with a timely opportunity to evaluate our current survey practices and consider more innovative alternatives. This is especially important as we seek to reach more of our patient population and ease the response burden for those patients. If we can increase the response rates for individual physicians, we will have more meaningful information for participating medical groups and physicians who use these results to monitor performance, identify improvement opportunities, and learn from best practices. We will also have the potential to publicly report patient experience results at the physician level to provide more information to patients who are choosing a doctor.

Moreover, modernizing the current standard survey is important because patient experience survey results are used for pay for performance programs in both states. In California patient experience results comprise 20% of the pay for performance (P4P) formula administered by the Integrated Healthcare Association (IHA), the largest non-governmental physician incentive program in the US. Similarly, in Massachusetts, Blue Cross Blue Shield of Massachusetts includes patient experience measures in the incentive formula for its Alternative Quality Contract and other health plans are beginning similar programs. These programs are evidence of the system-wide move toward value based reimbursement and highlight the importance of including the patient perspective when assessing value.

Our findings strongly indicate that short form electronic surveys are comparable to long form surveys fielded through traditional survey administration protocols and point the way toward future research activities. We have identified these key areas needing further testing to both validate our findings and further improve measurement:

- **Implementation and testing of short form electronic surveys on a larger scale** - The results of this study strongly support the value of a shorter electronic survey through lowered costs for both paper and electronic survey versions and improved response rates. Because results are used for high stakes P4P, further evaluation of a short form and/or electronic survey is recommended to assess the impact of changes on results used for performance incentives. Our existing statewide measurement programs offer the opportunity to expand upon what was done in the pilot study to evaluate whether adjustments are needed to ensure that performance levels and changes in performance are accurately measured as we move to the shorter electronic version. Such a study would further validate findings and accelerate adoption of advances in this area.

- **New composite measures.** Summary composite measures make public reporting easier to understand and use. Through the pilot we were able to create new composites and found there are several areas needing further development:

    - **Access** – Our analysis of results indicated that the access related questions in the pilot study survey could not be grouped together to create a summary composite measure. Testing the performance of other access related survey questions and question groupings will help us identify those questions that work best together to create a summary composite measure for this important aspect of care.

    - **Care Coordination** – As we have seen with previous efforts, creating a summary measure for care coordination is challenging because individual questions are designed to address different points of care and the character of the experience measured can vary significantly (e.g., receiving test results vs. a PCP knowing about a patient's visit with a specialist). Coordination of care is an area of high interest for patients, providers and policy makers. We want to identify the best measures of this concept by working with survey research experts and then, through focus groups with providers and patients, determine the best way of presenting these results so that they can be used by providers to improve coordination and by patients to understand the quality of coordination they should expect to receive.

    - **Patient Engagement** – By grouping questions that asked patients about talking with providers into a composite measure, the pilot study was a first step in identifying a group of questions that work together to measure engagement. Further work should also evaluate the best way to ask patients about the quality of their interactions with providers.

- **More frequent surveying for accountability and other high stakes use** - Providers tell us that survey results are more meaningful and actionable for quality improvement when they are timely. Patients also find questions about recent experiences easier to answer. It has become common practice to survey patients for both quality improvement and accountability on separate tracks which is burdensome and duplicative for patients and providers alike. To reduce redundancy and improve efficiency, we want to explore ways to leverage the results derived from more frequent surveys to meet accountability requirements. Testing would evaluate the comparability of more frequent, quarterly surveys with the annual CAHPs survey in terms of usefulness and statistical validity.

• **Testing and evaluation of a pediatric version of the short form survey** – Our short form pilot was tested on the adult population only and families caring for children also need information about care quality. Therefore we need to confirm that questions perform the same way for the pediatric population and also need to test inclusion of survey questions in areas specific to pediatrics, such as Growth and Development.

• **Establishing best practices for multi-mode surveys** - Finally, this study indicated that the best way to employ new modes of surveying is to integrate them into a multi-modal survey protocol. Our findings support the need to design approaches that strategically consider respondent preferences to achieve higher response rates and more efficiently reach patients. Further research into the communication preferences of different populations and the effect of different survey administration and response modes on different populations will help programs implement protocols that more effectively and efficiently capture results for a wide diversity of patients.

STUDY LIMITATIONS

While the findings provided insights into the strengths of the short form survey, there were also limitations to the study that should be noted here.

- This study only included primary care providers and the application of its findings to surveys that include specialists should be made with caution.

- In California, the field time for the mailed short form surveys (8 weeks) was not as long as for the regular PAS surveys (13 weeks). In MA, both were 11 weeks. Based on observed survey tail-off we estimate that the short form yield rates could have been as much as 3% higher. This result might not be this high, however, due to short form responses being slightly accelerated relative to long forms (respondents were more likely to respond quickly).

- A third limitation was that surveys were only administered in English. By not offering alternate language versions we are missing the unique opinions and perspectives of individuals from other cultures that can make up significant portions of a physician's patient population. The long forms included Spanish and (in California) Chinese and Vietnamese. Had this been done for short forms, their response rates would have been higher, but controls for race and ethnicity limit differences in the response values due to those factors.

- A fourth limitation was that phone surveys were not done as a follow up means to reaching patients who had not filled out a paper or email-based survey. Because phone surveys are done in the standard long form PAS fielding, it would have been helpful to compare this aspect of the survey. For purposes of analysis, phone responses were treated as non-responses. The phone might have increased or decreased the number of regular mail responses.

- Results strongly suggest that the shorter CG-CAHPS 3.0 instrument would also have better response rates. However, that might require shortening the physical document as well as reducing the number of questions. On the basis of these results we would hypothesize high convergence statistics for CG-CAHPS 3.0.

PHASE ONE ANALYSIS OF OPEN-ENDED NARRATIVE RESPONSES

An important aspect of the MHQP/CHPI pilot was that it included open-ended questions that would allow patients to express in their own words their perspective on the care they had received. In this part of the pilot project two different sets of open-ended questions were tested. Respondents were randomly assigned to receive either a three-item elicitation or a five-item elicitation.

The RAND Corporation, which is engaged in an overall analysis of the CAHPS Open-Ended Questions, partnered with MHQP and CHPI to complete a sub-analysis of the de-identified, open-ended responses received for both sets of questions used in this pilot project. The 3-question protocol used in this study was developed by researchers UCLA and is based on the Kano model of customer satisfaction, developed by the Japanese professor Noriaki Kano. The wording of the 3-question protocol follows this model to the extent that it asks patients to report on the two extremes of things about the provider or office staff that either "delighted" or "disappointed" them. The 5-question protocol was developed by the CAHPS Narrative Elicitation team as part of work supported by the Agency for Healthcare Research and Quality.

The questions in the three-item elicitation were:

- Please tell us how this doctor's office could have improved the care and services you received in the last 12 months.
- Please describe something about this doctor or health care provider that delighted or disappointed you.
- Please describe something about the staff at this office– the receptionist or nurses – that delighted or disappointed you.

The questions in the five-item elicitation were:

- What are the most important things you look for in a healthcare provider and his or her staff?
- When you think about the things that are most important to you, how do your provider and his or her staff measure up?
- Now we would like to focus on anything that has gone well in your experiences with your provider and his or her staff over the past 12 months. Please explain what happened, how it happened and how it felt to you.
- Now we'd like to focus on any experiences with your provider and his or her staff that you wish had gone differently over the past 12 months. Please explain what happened, how it happened and how it felt to you.
- Please explain how you and your provider relate to and interact with each other.

These questions were provided to all respondents in the MHQP and CHPI studies that either received an email invitation to complete the survey or who received the mail survey but chose to answer it via the web (MHQP N=4807, CHPI N = 1523) Respondents were not told how many questions they would receive prior to making this choice, so those agreeing were randomly assigned to receive either the three-item elicitation or the five-item elicitation.

In this first phase of the analysis four basic questions were addressed. Each is noted below with RAND's overall findings.

KEY FINDINGS FROM PHASE ONE

• Overall, 24.5% of experiment-eligible patients given the CHPI survey and 17.8% of experiment-eligible patients given the MHQP survey answered at least one open-ended question. Response rates were similar for the two versions of the protocol within each sample.

• Response rates did not drop across questions, nor were they lower at the end of the 5-question protocol than at the end of the 3-item protocol, as might be expected from respondent fatigue.

• Those receiving the 5-question protocol gave longer (and in the case of MHQP, much longer) responses than did those receiving the 3-question protocol.

• There were differences in response rates by patient characteristics in both samples, but none of these varied substantially between the two protocols. Specifically, in both samples, lower educational attainment, younger age, and being of Asian background were associated with a lower likelihood of providing a narrative comment. In addition, within the MHQP sample, those missing race/ethnicity information were less likely than Whites to provide a response.

• There were no differences in length of narrative (i.e., word count) across patient characteristics within the CHPI data, but within the MHQP data younger adults and women gave longer responses. The gender effect was strongest within the 5-question protocol.

DETAILED CHPI RESULTS

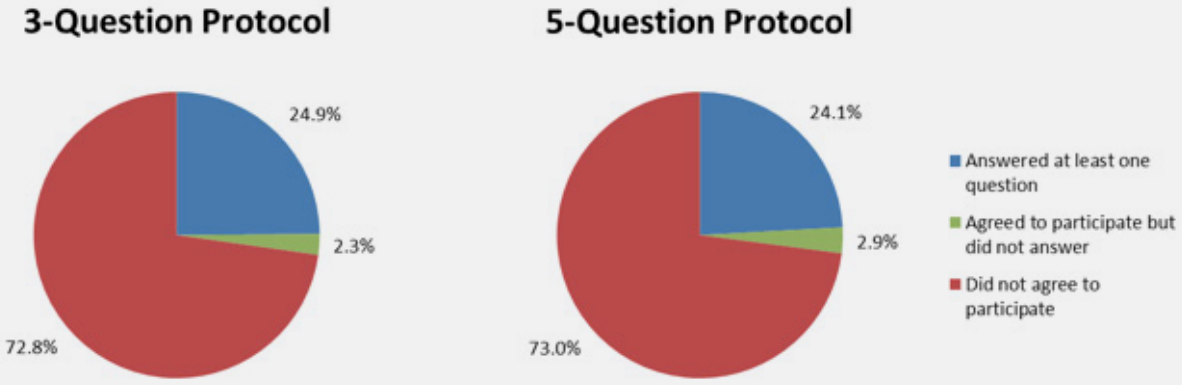
I. How frequently will patients provide narrative commentary?

Overall, 1543 patients participated in the web version of the CHPI survey. Of these, 1489 stated that they received care from the named provider and had one or more visits to that provider in the past year. Overall, 365 patients (24.5%) answered at least one of the open-ended questions, 39 (2.6%) agreed to leave a comment but then answered no open-ended questions, and 1085 (72.9%) did not choose to leave a comment.

II. Is the extent of commentary affected by whether the elicitation questions are in a shorter (3-question) or longer (5-question) protocol?

Of the 1489 participating patients, 768 received the 3-question open-ended protocol and 721 received the 5-question protocol. The rates of response, displayed in Exhibit 1, did not differ either substantively or statistically between the two protocols.

Exhibit 1: Open-ended CHPI Response Rates by Version of the Protocol



III. Do response rates decline after the initial open-ended questions, perhaps indicating greater fatigue?

Exhibit 2 demonstrates that the percentage of patients that gave a response did not generally drop over the course of the questions, for either protocol. Furthermore, the response rate for the final question was quite similar across the two protocols, and in fact was a bit higher in the longer protocol. This suggests that fatigue across questions was not substantial, and that it was no greater in the longer elicitation.

Exhibit 2: Percentage Responding per Question, CHPI

		Question				
Protocol	Sample size	1	2	3	4	5
3-Question	768	18.0	22.7	19.3	n/a	n/a
5-Question	721	21.5	20.4	17.8	15.4	20.0

In contrast, Exhibit 3 shows that those receiving the 5-question protocol provided more information, in terms of larger word counts, than did those receiving the 3-question protocol. The difference in total word count was statistically significant ($p = 0.01$).

Exhibit 3: Mean Word Count per Question, among Those Agreeing to Leave a Comment, CHPI

	Question					
Protocol	1	2	3	4	5	TOTAL
3-question	24.8	28.1	16.5	n/a	n/a	69.3
5-question	18.7	16.1	16.8	15.0	20.3	86.8

IV. Does participation vary across different subgroups of patients?

The specific concern here are groups that historically have been less inclined to voice their health care experience, such as minorities or those with less education. Exhibit 4 presents the percentage of respondents (out of 1489) who answered at least one open-ended question, broken down by which protocol they received. The final two columns indicate whether there was a statistically significant difference ($p < .05$) between the demographic categories and whether or not this difference varied by which protocol they received (using logistic regression with patient characteristic, protocol, and their interaction as predictors). For simplicity sake, some characteristics with multiple levels (educational attainment, self-rated health, age) are collapsed into two categories in Exhibit 5, but were treated as continuous predictors in the statistical models (see technical appendix for details).[3]

Higher education and older age were associated with a greater likelihood of providing a response to the open-ended questions. As seen elsewhere, Asians were less likely to provide a response than were Whites.

Exhibit 4: Percentage Responding by Patient Characteristics and Version of the Protocol, CHPI

		Protocol			
Patient characteristic	N[a]	3-Question	5-Question	Difference by patient characteristic?[b]	Did difference vary by protocol?
Gender				No	No
Male	601	23.4	22.1		
Female	888	25.7	25.4		
Educational attainment				Yes	No
High school degree, GED, or less	172	18.9	15.6		
More than high school degree or GED	1273	26.4	25.7		
Self-rated health				No	No
Excellent or very good	758	26.8	22.2		
Good, fair, or poor	691	23.9	27.8		
Age				Yes	No
19 to 53	715	20.8	25.3		
54 to 80	774	28.2	22.9		
Race/ethnicity[c]				Yes	No
White	808	26.4	25.8		
Hispanic	252	27.1	29.3		
Asian	197	16.3	14.0		
Other race/ethnicity	83	29.5	25.6		
Missing race/ethnicity	76	15.4	10.8		

[a] Not all respondents provided all patient characteristic data, so sample sizes do not always sum to 1489; [b] Yes indicates a statistically significant difference, $p < .05$; [c] Three race categories (Black; Native Hawaiian or Other Pacific Islander; American Indian or Alaskan Native) were excluded due to low frequency.

[3]These analyses were conducted independently of each other, such that other patient characteristics were not controlled for in each regression model. Future analyses should consider the effect of adjusting for other patient characteristics, in order to account for relationships among characteristics.

Exhibit 5 shows that, across the board in the CHPI data, there were no statistically significant differences in word count across any of the patient characteristics (among the 404 who agreed to leave a comment). Furthermore, there were no significant interactions with which version of the protocol the patient received, suggesting that the differences in word count between protocols (described above) are fairly consistent across patient groups.

Exhibit 5: Mean Word Count by Patient Characteristics and Version of the Protocol, CHPI

Patient characteristic	N[a]	Protocol		Difference by patient characteristic?[b]	Did difference vary by protocol?
		3-Question	5-Question		
Gender				No	No
Male	146	71.3	79.3		
Female	258	68.1	90.5		
Educational attainment				No	No
High school degree, GED, or less	36	73.0	82.3		
More than high school degree or GED	366	69.0	87.7		
Self-rated health				No	No
Excellent or very good	214	69.3	83.9		
Good, fair, or poor	190	69.4	89.8		
Age				No	No
19 to 53	189	66.5	92.7		
54 to 80	215	71.4	80.3		
Race/ethnicity[c]				No	No
White	239	70.2	88.0		
Hispanic	75	70.6	79.4		
Asian	35	55.2	69.5		
Other race/ethnicity	23	86.2	95.0		
Missing race/ethnicity	11	45.0	58.0		

[a] Not all respondents provided all patient characteristic data, so sample sizes do not always sum to 404; [b] Yes indicates a statistically significant difference, $p < .05$; [c] Three race categories (Black; Native Hawaiian or Other Pacific Islander; American Indian or Alaskan Native) were excluded due to low frequency.

DETAILED MHQP RESULTS

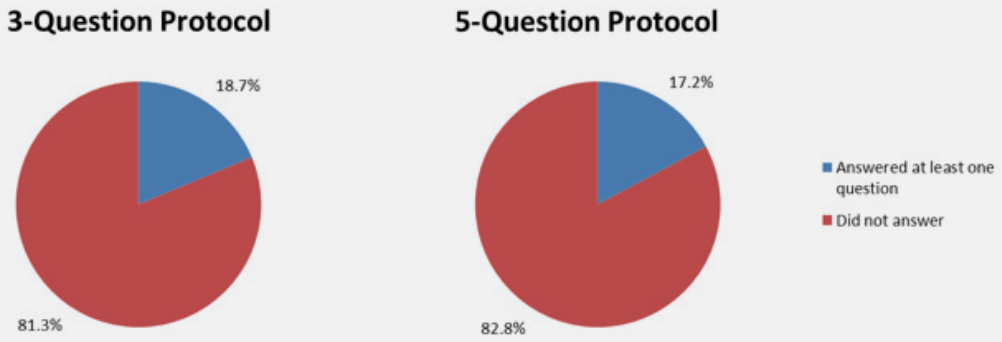
V. How frequently will patients provide narrative commentary?

Overall, 4807 patients participated in the MHQP survey. Of these, 4793 stated that they received care from the named provider and had one or more visits to that provider in the past year. Overall, 857 patients (17.8%) answered at least one of the open-ended questions and 4078 (82.2%) did not choose to leave a comment. Like the CHPI survey, the MHQP survey also first asked patients if they would like to leave a comment. However, there was no indicator for this in the MHQP dataset, so those results are not presented here.

VI. Is the extent of commentary affected by whether the elicitation questions are in a shorter (3-question) or longer (5-question) protocol?

Of the 4793 patients, 2334 received the 3-question open-ended protocol and 2459 received the 5-question protocol. The rates of response, displayed in Exhibit 6, did not differ either substantively or statistically between the two protocols.

Exhibit 6: Open-ended MHQP Response Rates by Version of the Protocol



VII. Do response rates decline after the initial open-ended questions, perhaps indicating greater fatigue?

Exhibit 7 demonstrates that the percentage of patients that gave a response did not generally drop over the course of the questions, for either protocol. Furthermore, the response rate for the final question was quite similar across the two protocols, and in fact was a bit higher in the longer protocol. This suggests that fatigue across questions was not substantial, and that it was no greater in the longer elicitation.

Exhibit 7: Percentage Responding per Question, MHQP

		Question				
Protocol	Sample size	1	2	3	4	5
3-Question	2334	11.3	16.8	12.1	n/a	n/a
5-Question	2459	14.8	14.4	12.6	11.7	14.3

In contrast, Exhibit 8 shows that those receiving the 5-question protocol provided a good deal more information, in terms of larger word counts, than did those receiving the 3-question protocol. The difference in total word count was highly statistically significant ($p < .0001$).

Exhibit 8: Mean Word Count per Question, among Those Leaving a Comment, MHQP

	Question					
Protocol	1	2	3	4	5	TOTAL
3-Question	21.0	38.9	15.0	n/a	n/a	74.8
5-Question	18.7	20.7	28.3	21.9	25.9	115.5

VIII. Does participation vary across different subgroups of patients?

We were interested in understanding if groups that historically have been less inclined to voice their health care experience, such as minorities or those with less education, responded as often as others when given the opportunity to provide comments in their own words or if the pattern seen in close-ended survey questions continued. Exhibit 9 presents the percentage of respondents (out of 4793) who answered at least one open-ended question, broken down by which protocol they received. The final two columns indicate whether there was a statistically significant difference ($p < .05$) between the demographic categories and whether or not this difference varied by which protocol they received (using logistic regression with patient characteristic, protocol, and their interaction as predictors). For simplicity sake, some characteristics with multiple levels (educational attainment, self-rated health, age) are collapsed into two categories in Exhibit 9, but were treated as continuous predictors in the statistical models (see RAND technical appendix for details).[4]

As shown in Exhibit 9, those with higher education, and older individuals were more likely to provide a narrative response. In addition, Asians and those missing ethnicity were less likely than others to provide a response to the open-ended questions.

Exhibit 9: Percentage Responding by Patient Characteristics and Version of the Protocol, MHQP

		Protocol			
Patient characteristic	N[a]	3-Question	5-Question	Difference by patient characteristic?[b]	Did difference vary by protocol?
Gender				No	No
Male	2590	17.4	15.6		
Female	2203	20.2	18.9		
Educational attainment				Yes	No
High school degree, GED, or less	464	13.5	14.1		
More than high school degree or GED	4232	19.6	17.8		
Self-rated health				No	No
Excellent or very good	3100	19.7	17.1		
Good, fair, or poor	1604	17.8	18.1		
Age				Yes	No
19 to 53	2720	15.3	15.4		
54 to 80	2070	22.9	19.5		
Race/ethnicity[c]				Yes	No
White	4040	19.4	17.9		
Hispanic	148	12.2	12.2		
Asian	256	8.4	10.2		
Other race/ethnicity	120	26.7	21.7		
Missing race/ethnicity	114	5.3	3.5		

[a] Not all respondents provided all patient characteristic data, so sample sizes do not always sum to 4793; [b] Yes indicates a statistically significant difference, $p < .05$; [c] Three race categories (Black; Native Hawaiian or Other Pacific Islander; American Indian or Alaskan Native) were excluded due to low frequency.

[4]These analyses were conducted independently of each other, such that other patient characteristics were not controlled for in each regression model. Future analyses should consider the effect of adjusting for other patient characteristics, in order to account for relationships among characteristics.

Exhibit 10 shows the relationships between patient characteristics and word count (among the 857 who left a comment). Among MHPQ respondents, women provided significantly longer narrative responses than did men, and this difference was significantly larger among those receiving the 5-question protocol. Younger patients gave longer responses, on average, regardless of protocol.

Exhibit 10: Mean Word Count by Patient Characteristics and Version of the Protocol, MHQP

Patient characteristic	N[a]	Protocol		Difference by patient characteristic?[b]	Did difference vary by protocol?
		3-Question	5-Question		
Gender				Yes	Yes
Male	426	65.2	91.4		
Female	431	84.4	139.1		
Educational attainment				No	No
High school degree, GED, or less	64	72.4	91.5		
More than high school degree or GED	789	74.4	117.9		
Self-rated health				No	No
Excellent or very good	569	71.2	111.3		
Good, fair, or poor	288	82.3	123.4		
Age				Yes	No
19 to 53	418	84.4	123.6		
54 to 80	438	67.0	106.5		
Race/ethnicity[c]				No	No
White	753	72.7	116.0		
Hispanic	18	76.9	105.2		
Asian	24	69.2	95.6		
Other race/ethnicity	29	87.4	110.6		
Missing race/ethnicity	5	153.3	59.5		5

[a] Not all respondents provided all patient characteristic data, so sample sizes do not always sum to 857; [b] Yes indicates a statistically significant difference, $p < .05$; [c] Three race categories (Black; Native Hawaiian or Other Pacific Islander; American Indian or Alaskan Native) were excluded due to low frequency.

PHASE TWO ANALYSIS OF OPEN-ENDED NARRATIVE RESPONSES

This phase of the RAND analysis focuses on the content of the patient narratives, with special emphasis given to comparing patients who were assigned to complete the 3-question narrative elicitation protocol versus patients who were assigned to complete the 5-question protocol.

ANALYSIS OF NARRATIVE CONTENT

Each patient's responses to the open-ended questions were aggregated to create a single narrative, and this narrative was then coded on fifteen dimensions. Coders rated the overall valence of the narrative according to the following scheme: 1=mostly negative, 2=more negative than positive, 3=equal mix of negative and positive, 4=more positive than negative, 5 = mostly positive. Coders then identified the **number of positive statements** and the **number of negative statements** in the narrative that pertained to the following seven aspects of care: provider communication, office staff, access to care, technical competence of the provider, time spent with the patient, caring on the part of the provider, and provider thoroughness. From this information, RAND created the following additional variables: any mention of each of the seven aspects of care, and breadth of the narrative as a whole (which is equal to the sum of the number of aspects of care mentioned in the narrative).

Coding of the narrative information was performed by two coders. To ensure inter-rater reliability in use of the coding scheme, two initial subsets of 47 CHPI patient narratives were sequentially coded by both coders. After each subset, inter-rater reliability was checked using intra-class correlation (ICC) coefficients, the two coders conferred to achieve consensus on any discrepancies, and as necessary the coding scheme was clarified. At the end of the second subset, the coders had achieved substantial to outstanding inter-rater reliability (ICC > .6) on twelve of the fifteen codes, with two being very close to this benchmark (negative statements made about ample time was .57 and positive statements made about thoroughness was .59), and a third (negative statements made about thoroughness) being rarely used by one coder and never by the other (ICC = 0). All subsequent patient narratives were coded by a single coder. An additional subset of 47 patient narratives was double-coded at the transition from CHPI to MHQP data, to check for any drift in application of the coding scheme. Results were similar to whose with the CHPI data, and overall inter-rater reliability (across all three subsets) was substantial or better (ICC > .6) for all but negative statements made about time spent with the patient (ICC = .57) and negative statements made about thoroughness (ICC = 0).

CHPI RESULTS

Overall Valence of Narratives

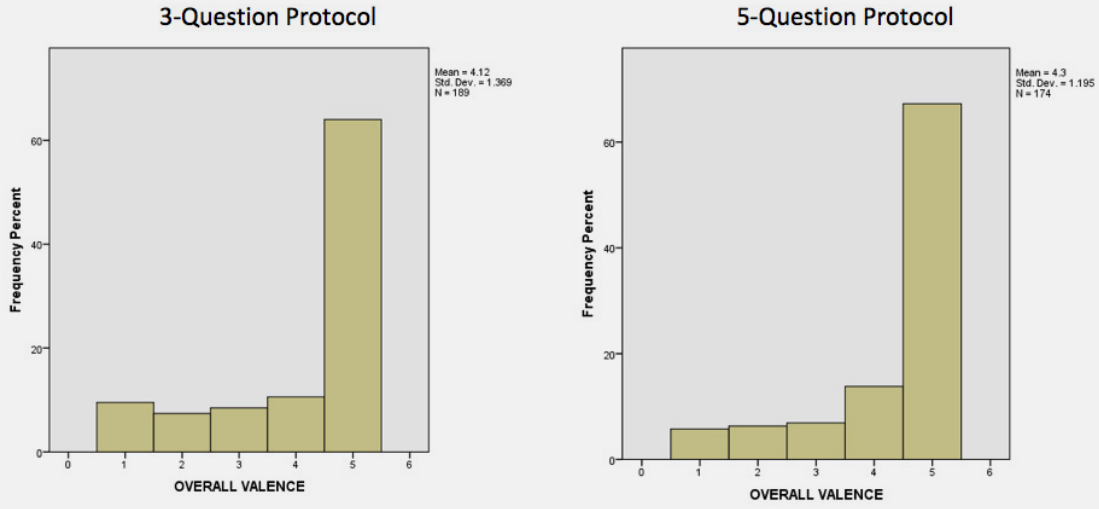
Across elicitation protocols, the mean overall valence of comments was 4.21 (SD = 1.29), indicating a high degree of positivity in patients' comments. The overall valence of narratives did not differ (p = .18) between the 3-question and 5-question protocols, as can be seen in Exhibit 1.

Exhibit 1: Overall Valence of Narratives by Condition (Type of Elicitation Protocol)

Narrative Elicitation Protocol	*N*	Overall valence, *M* (*SD*)
3-question protocol	189	4.12 (1.37)
5-question protocol	174	4.30 (1.20)
Total	363	4.21 (1.29)

The distribution of overall valence was also roughly similar across the two conditions (see Exhibit 2).

Exhibit 2: Distribution of Overall Valence



Was narrative valence associated with scores on the CAHPS measures and did associations differ by which protocol patients received?

There were strong positive associations between the overall valence of narratives and all CAHPS measures. That is, more positive commentary was associated with higher scores on the CAHPS doctor communication composite (p < .0001), higher scores on the access composite (p < .0001), higher scores on the care coordination composite (p < .0001), higher global ratings of one's provider (p < .0001), and a greater willingness to recommend one's provider (p < .0001).

The association between overall valence of the narratives and scores on the CAHPS doctor communication composite was stronger with the 5-question protocol than with the 3-question protocol (p = .02), as was the association between the overall valence of the narratives and global provider ratings (p < .0001) and between the overall valence of the narratives and patients' willingness to recommend their providers (p = .002). The association between the overall valence of the narratives and scores on the CAHPS care coordination composite was marginally stronger with the 5-question protocol than with the 3-question protocol (p = .07). The association between overall valence of the narratives and scores on the CAHPS access composite did not differ by protocol (p = .81).

Were patient characteristics associated with narrative valence?

Race/ethnicity, gender, and educational background were not associated with the overall valence of patient narratives (p's = .12 or greater). Older patients provided more positive narratives than younger patients (p = .03). Patients with better self-rated health provided more positive narratives than patients with worse self-rated health (p < .0001), but this association was only evident with the 3-question protocol.

Any Mention of Specific Aspects of Care

Did the likelihood of a patient mentioning a specific aspect of care differ by protocol?

Patients were more likely to comment on provider communication, access to care, technical competence of the provider, the amount of time spent with the patient, and caring on the part of the provider when given the 5-question protocol than when given the 3-question protocol (see Exhibit 3). Office staff was, however, more often mentioned by patients given the 3-question protocol than when given the 5-question protocol. It is worth mentioning, however, that the 3-question protocol explicitly calls out office staff, which may account for this result.

Exhibit 3: Percent of Narratives Containing Any Mention of Specific Aspects of Care by Protocol

Aspect of care	3-question protocol (N = 189)	5-question protocol (N = 174)	*p*
Provider communication	49%	68%	<.0001
Office staff	76%	52%	<.0001
Access to care	34%	43%	.08
Technical competence	28%	50%	<.0001
Time spent with patient	19%	28%	.03
Caring	59%	76%	.001
Thoroughness	9%	9%	.99

Positive and Negative Statements about Specific Aspects of Care

Did the number of positive and negative statements made about specific aspects of care differ by protocol?

The number of positive statements made about each aspect of care was greater with the 5-question protocol than with the 3-question protocol for all but two aspects of care: thoroughness and office staff. For thoroughness, the number of positive statements made was similar across protocols. For office staff, the number of positive statements made was greater with the 3-question protocol than with the 5-question protocol (see top half of Exhibit 4). In contrast, the number of negative statements made about each aspect of care was equivalent across protocols (see bottom half of Exhibit 4).

Exhibit 4: Number of Positive and Negative Statements Made about Specific Aspects of Care, By Condition

Number of positive statements (*M, SD*)			
Aspect of care	**3-question protocol (N = 189)**	**5-question protocol (N = 174)**	***p***
Provider communication	0.50 (0.70)	1.13 (1.15)	<.0001
Office staff	1.00 (0.89)	0.59 (0.73)	<.0001
Access to care	0.17 (0.43)	0.44 (0.78)	<.0001
Technical competence	0.26 (0.51)	0.66 (0.83)	<.0001
Time spent with patient	0.14 (0.36)	0.33 (0.65)	.001
Caring	0.69 (0.77)	1.33 (1.20)	<.0001
Thoroughness	0.10 (0.36)	0.11 (0.40)	.73
Number of negative statements (*M, SD*)			
Aspect of care	**3-question protocol (N = 189)**	**5-question protocol (N = 174)**	***p***
Provider communication	0.12 (0.38)	0.11 (0.42)	.87
Office staff	0.20 (0.60)	0.21 (0.58)	.93
Access to care	0.31 (0.66)	0.25 (0.56)	.36
Technical competence	0.07 (0.33)	0.03 (0.17)	.11
Time spent with patient	0.08 (0.33)	0.06 (0.29)	.51
Caring	0.10 (0.37)	0.13 (0.44)	.47
Thoroughness	0.02 (0.13)	0.01 (0.08)	.36

Breadth of Narrative

Did the breadth of patient narratives differ by protocol?

As can be seen in Exhibit 5, the breadth of patient narratives (i.e., the number of aspects of care mentioned by the average patient) was greater with the 5-question protocol than with the 3-question protocol ($p < .0001$).

Exhibit 5: Number of Aspects of Care Mentioned by the Average Patient (Breadth of Narrative) By Condition

Narrative Elicitation Protocol	*N*	Number of aspects of care mentioned, *M* (*SD*)
3-question protocol	189	2.73 (1.24)
5-question protocol	174	3.26 (1.38)
Total	363	2.98 (1.33)

MHQP RESULTS

Overall Valence of Narratives

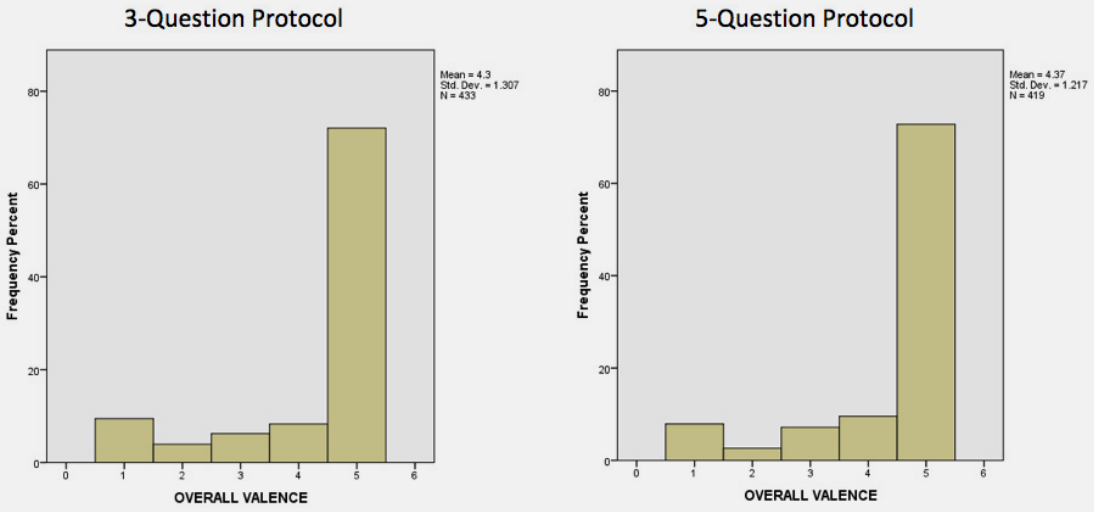
Across elicitation protocols, the mean overall valence of comments was 4.33 (SD = 1.26), indicating a high degree of positivity in patients' comments. The overall valence of narratives did not differ (p = .41) between the 3-question and 5-question protocols, as can be seen in Exhibit 1.

Exhibit 1: Overall Valence of Narratives by Condition (Type of Elicitation Protocol)

Narrative Elicitation Protocol	*N*	Overall valence, *M* (*SD*)
3-question protocol	433	4.30 (1.31)
5-question protocol	419	4.37 (1.21)
Total	852	4.33 (1.26)

The distribution of overall valence was also roughly similar across the two conditions (see Exhibit 2).

Exhibit 2: Distribution of Overall Valence



Was narrative valence associated with scores on the CAHPS measures and did associations differ by which protocol patients received?

There were strong positive associations between the overall valence of narratives and all CAHPS measures. That is, more positive commentary was associated with higher scores on the CAHPS doctor communication composite (p < .0001), higher scores on the access composite (p < .0001), higher scores on the care coordination composite (p < .0001), higher global ratings of one's provider (p < .0001), and a greater willingness to recommend one's provider (p < .0001).

The association between overall valence of the narratives and scores on the CAHPS doctor communication composite was stronger with the 5-question protocol than with the 3-question protocol (p = .02), as was the association between the overall valence of the narratives and the CAHPS care coordination composite (p = .04). The association between overall valence of the narratives and scores on the CAHPS access composite did not differ by protocol (p = .15), nor did the association between the overall valence of narratives and patient's global ratings of their providers (p = .32) or their willingness to recommend their providers (p = .33).

Were patient characteristics associated with narrative valence?

Race/ethnicity and educational background were not associated with the overall valence of patient narratives (p's = .17 or greater). Male patients provided more positive patient narratives than female patients (p = .04), older patients provided more positive narratives than younger patients (p = .001), and patients with better self-rated health provided more positive narratives than patients with worse self-rated health (p < .003). These associations did not differ by protocol.

Any Mention of Specific Aspects of Care

Did the likelihood of a patient mentioning a specific aspect of care differ by protocol?

Patients were more likely to comment on provider communication, access to care, technical competence of the provider, the amount of time spent with the patient, caring on the part of the provider and thoroughness of the provider when given the 5-question protocol than when given the 3-question protocol (see Exhibit 3). Office staff was, however, more often mentioned by patients given the 3-question protocol than when given the 5-question protocol. It is worth mentioning, however, that the 3-question protocol explicitly calls out office staff, which may account for this result.

Exhibit 3: Percent of Narratives Containing Any Mention of Specific Aspects of Care by Protocol

Aspect of care	3-question protocol (N = 433)	5-question protocol (N = 419)	*p*
Provider communication	47%	72%	<.0001
Office staff	65%	46%	<.0001
Access to care	38%	47%	.006
Technical competence	30%	59%	<.0001
Time spent with patient	21%	31%	.001
Caring	51%	76%	<.0001
Thoroughness	5%	12%	<.0001

Positive and Negative Statements about Specific Aspects of Care

Did the number of positive and negative statements made about specific aspects of care differ by protocol?

The number of positive statements made about each aspect of care was greater with the 5-question protocol than with the 3-question protocol for all but one aspect of care: office staff. For office staff, the number of positive statements made was greater with the 3-question protocol than with the 5-question protocol (see top half of Exhibit 4). In contrast, the number of negative statements made about each aspect of care was equivalent across protocols for all but one aspect of care: caring on the part of the provider. For that one aspect, the number of negative statements made was greater with the 5-question protocol than with the 3-question protocol (see bottom half of Exhibit 4).

Exhibit 4: Number of Positive and Negative Statements Made about Specific Aspects of Care, By Condition

Number of positive statements (*M*, *SD*)			
Aspect of care	**3-question protocol (N = 433)**	**5-question protocol (N = 419)**	***p***
Provider communication	0.47 (0.62)	1.20 (1.13)	<.0001
Office staff	0.85 (0.87)	0.57 (0.82)	<.0001
Access to care	0.24 (0.51)	0.50 (0.77)	<.0001
Technical competence	0.25 (0.45)	0.88 (0.98)	<.0001
Time spent with patient	0.19 (0.40)	0.32 (0.55)	<.0001
Caring	0.52 (0.59)	1.42 (1.23)	<.0001
Thoroughness	0.05 (0.23)	0.14 (0.40)	<.0001
Number of negative statements (*M*, *SD*)			
Aspect of care	**3-question protocol (N = 433)**	**5-question protocol (N = 419)**	***p***
Provider communication	0.09 (0.35)	0.11 (0.45)	.53
Office staff	0.17 (0.49)	0.18 (0.58)	.78
Access to care	0.27 (0.59)	0.24 (0.55)	.35
Technical competence	0.07 (0.30)	0.05 (0.26)	.27
Time spent with patient	0.04 (0.21)	0.05 (0.26)	.42
Caring	0.05 (0.21)	0.09 (0.36)	.03
Thoroughness	0.00 (0.00)	0.01 (0.08)	.08

Breadth of Narrative

Did the breadth of patient narratives differ by protocol?

As can be seen in Exhibit 5, the breadth of patient narratives (i.e., the number of aspects of care mentioned by the average patient) was greater with the 5-question protocol than with the 3-question protocol ($p < .0001$).

Exhibit 5: Number of Aspects of Care Mentioned by the Average Patient (Breadth of Narrative) By Condition

Narrative Elicitation Protocol	*N*	Number of aspects of care mentioned, *M* (*SD*)
3-question protocol	433	2.57 (1.33)
5-question protocol	419	3.44 (1.46)
Total	852	3.00 (1.46)

TECHNICAL APPENDIX:

THIS APPENDIX PRESENTS ADDITIONAL METHODOLOGICAL DETAILS INVOLVED IN THE RAND ANALYSIS.

CONSTRUCTION OF WORD COUNT

The data received by RAND were de-identified, which means that in places part or all of a narrative response was redacted as involving identifying or personal health information. In many of these cases, it was clear that this information was simply the name of a patient, location, provider, or health condition. In these cases, RAND treated the redacted information as a single word when computing word count. In a few cases, it appeared as if substantial text or the full response was redacted. In these instances, when the word count of a response to a given question (within a given protocol) was less than the mean word count for that question and protocol, RAND replaced the word count with the mean for that question and protocol. If the word count of the redacted response was longer than the mean word count for that question and protocol, RAND treated the redacted information as a single word when computing word count.

CODING OF DEMOGRAPHIC DATA

Within the logistic regression models, gender was coded as 0 for male and 1 for female. Educational attainment coded as 1 for less than 8th grade through 6 for more than 4-year college degree. Self-rated health was coded as 1 for excellent through 5 for poor. Race/ethnicity was coded as a series of indicator variables comparing the following groups to Whites: Hispanic, Black, Asian, other race/ethnicity, missing race/ethnicity.

APPENDIX I: INFORMATION ABOUT MHQP AND CHPI

MASSACHUSETTS HEALTH QUALITY PARTNERS (MHQP)

Massachusetts Health Quality Partners (MHQP) is a non-profit organization established in 1995 that provides reliable information to help physicians improve the quality of care they provide their patients and help consumers take an active role in making informed decisions about their health care. MHQP's mission is to drive measureable improvements in health care quality, patients' experiences of care, and use of resources in Massachusetts through patient and public engagement and broad-based collaboration among health care stakeholders, including physicians, hospitals, health plans, purchasers, patient and public advocates, government agencies, and academics. MHQP governance includes a multi-stakeholder Board and a Physician Council, Health Plan Council, and Consumer Health Council. All Councils are represented on MHQP's Board.

A trusted leader in health care quality measurement and evaluation, MHQP is at the national forefront in the development and implementation of valid and reliable measures of the patient care experience. MHQP's early work in developing the Ambulatory Care Experiences Survey (ACES) instrument, with researchers at the Tufts Health Institute, tested core survey content and was instrumental in establishing the feasibility and value of measuring patients' experiences with clinicians and practices. The findings from this work informed the development of the CAHPS Clinician/Group Survey which has become the national standard for ambulatory care patient experience measurement.
Massachusetts was the first state in the nation to publicly report statewide Patient Experience Survey results for primary care. Since 2005, MHQP has overseen a statewide commercial Patient Experience Survey program, fielding a biennial survey at the primary care physician practice site level (for sites with three or more physicians) and publicly reports the results on http://www.healthcarecompassma.org. In July 2015, MHQP completed the field period for its seventh round of what has now become an annual statewide commercial survey using the PCMH CAHPS survey instrument with content added for regional pay for performance programs.

This well established program has been supported by the five largest commercial plans in Massachusetts, representing 86% of the commercially insured population in the state and has become integral to quality improvement programs while also providing high quality information to the public. In 2014, MHQP was successful in moving to a multi-stakeholder funding model for the annual Patient Experience Survey with funding from health plans, physician organizations, and the state, demonstrating the value this survey provides to multiple players.

Sponsoring health plans receive results to support their quality improvement and pay for performance initiatives (e.g., the Blue Cross Blue Shield of MA Alternative Quality Contract (AQC)). A number of physician organizations opt to increase sample sizes to obtain provider level results and include their smaller PCP practices in the survey. The state's Health Policy Commission and Center for Health Information and Analysis (CHIA) rely on MHQP's Patient Experience Survey to monitor health care quality trends in MA.

To expand the reach of its public reporting, in 2012 MHQP jointly published the results of its statewide survey with Consumer Reports. The MHQP/Consumer Reports partnership was a pilot project funded by the Robert Wood Johnson Foundation (RWJF) to provide consumers with valid and reliable health information to support informed decision-making. Consumer Reports published a special 32-page print insert of MHQP's 2011 Patient Experience Survey (PES) results along with editorial content for 120,000 Massachusetts subscribers. This report was the first of its kind in the nation and significantly broadened the reach of public reporting of patient experience results. MHQP has subsequently produced a "plain language" version of the report to help patients better communicate with their doctors and better coordinate their care.
In addition to the work that MHQP has done to establish standard public reporting about health care quality, we are pursuing new ways of implementing surveys to provide ongoing feedback about care to practices.

[5]Safran, DG, Karp, M,Coltin, K, Chang, H., Li, A, Ogren, Rogers, W. 2005. Measuring Patients' Experiences with Individual Primary Care Physicians. *Journal of General Internal Medicine.*

CALIFORNIA HEALTH PERFORMANCE INFORMATION SYSTEM (CHPI)

CHPI is the only Multi-Payer Claims Database (MPCD) currently in operation in California and consists of both insured and self-funded claims from the state's three largest health plans and the Medicare fee-for-service program. CHPI builds on a history of physician performance measurement programs since 2007, involving many of the same collaborators as the California Better Quality Information (BQI) Pilot and the California Physician Performance Initiative (CPPI). CHPI is a public benefit corporation (501(c)4) that is governed by an independent board of stakeholder representatives from health plans, providers, purchasers, and consumers (see Appendix B for CHPI's Board of Directors and Advisory Committees).
The Pacific Business Group on Health (PBGH) administers CHPI through a professional services contract. PBGH is a non-profit 501(c)3 coalition of public and private purchasers whose mission is to act as an influential change agent, demanding increased value in the health care system through collaborative purchaser action and support for systemic change initiatives to drive improvement in affordability, quality, and service.

In mid-2013, the Patient Assessment Survey (PAS) program was brought under the governance of CHPI. Incorporating PAS into CHPI added a third, distinct aspect of quality to its performance information work. The PAS contains a set of patient experience measures on access, patient-doctor interactions, office staff interactions, coordination of care, health promotion, and overall ratings of care. The survey uses the industry-standard CG-CAHPS© instrument with some customization for topics of local interest.

First fielded in 2001, the PAS is a yearly survey that measures patient experience with medical groups among adult HMO and POS enrollees in California. In 2014, 10 major California health plans and 112 unique physician organizations (reporting on 174 units) collaborated on the PAS project. The 2013 participating groups served 9.9 million commercially insured HMO and POS patients, or almost 95% of the total HMO/POS commercial population in California. The participating health plans in 2014 were: Aetna Healthcare of California, Anthem Blue Cross, Blue Shield of California, CIGNA Healthcare of California, Health Net, Kaiser Foundation Health Plan of Northern California, Kaiser Foundation Health Plan of Southern California, Sharp Health Plan, United Healthcare, and Western Health Advantage.

The PAS information supports patients in choosing and using health care providers and services. Survey results comprise 20% of the pay for performance formula administered by the Integrated Healthcare Association (IHA). The results also are used in medical group quality improvement work and previously have been published online by the California Office of the Patient Advocate. In 2014, the PAS program included two new publishers Consumer Reports (California-specific insert) and CalQualityCare.org (a California based health care ratings web site run by the California HealthCare Foundation). Previous years' PAS results can be found on the results page of the CHPI web site. A number of physician groups also administer the PAS Doctor Survey, whose results are available only to the group, to obtain patient experience results at the individual physician level for performance improvement and recognition activities.

APPENDIX II: COMPARISON OF CHPI/MHQP SHORT FORM SURVEY AND CG-CAHPS 3.0 SURVEY

BACKGROUND

To implement this project the team was charged with developing a short form survey instrument by December 1, 2014 in order to implement survey administration on schedule. Fortunately, there was an established body of work to help guide our work in this area. At the outset of our survey development process, the project team consulted with National Committee for Quality Assurance (NCQA) and the Consumer Assessment of Health Plans and Systems (CAHPS) team. The CAHPS team is under the aegis of the federal Agency for Healthcare Research and Quality (AHRQ) and has developed the core set of Clinician-Group (C-G) questions that are used in a number of federal programs. NCQA has implemented accreditation and recognition programs that are being widely adopted across the country. The recognition program that is particularly relevant to our work is the Patient Centered Medical Home (PCMH) CAHPS instrument. These two organizations are the most influential in the country with regard to patient experience measurement and have longstanding expertise in patient experience measurement and survey development.

Both NCQA and the CAHPS team had released revised and shortened versions of their surveys for public comment in fall 2014. NCQA released a shortened instrument in September 2014 and the CAHPS team released their revised instrument in November 2014. CHPI and MHQP staff met with both teams to discuss and understand the changes being proposed. At the time of our meetings, the proposed instruments were not in alignment. There was important agreement around the questions most important to PCMH measurement, but NCQA's survey version was considerably shorter than the standard PCMH CAHPS instrument, which included C-G CAHPS core content (13 items for NCQA vs. 31 for CAHPS).

Our project team also considered project goals as we defined the final instrument. A primary objective of the project was to improve response rates through shortening the survey. The project team agreed that a multi-page survey would not be perceived as shorter by patients and therefore we endeavored to develop a one-page survey.

Both CHPI and MHQP added limited content to NCQA's proposed instrument after consulting with our stakeholders. Some key issues raised by our stakeholders as we finalized content included:

- The impact of reducing the number of questions to be used for high stakes P4P and limiting the focus of measurement

- Some stakeholders also felt that the survey was still longer than it needed to be.

- Certain questions are focused more on process than outcome. For example, a question asking whether information about getting weekend care was given by a provider's office does not measure whether patients know how to get care.

The project team decided to include all of the recommended PCMH content that both NCQA and CAHPS agreed upon, aligning more closely with the shorter NCQA version for our proposed draft instrument, but adding CAHPS content that both regions agreed were important measures of care. Regional summary rating items were added so that comparisons with statewide long form versions could be made.

Regions considering implementing a short form survey should consider how their survey results are used, specifically whether there is a need to align with surveys that rely upon CAHPS core content. In July 2015 the CAHPS team released its 3.0 version of its C-G CAHPS survey. The team has recently announced plans to do further work to understand how this standard survey may be shortened.

The following table compares the final set of questions for both short form surveys and highlights where the CHPI/MHQP short form survey diverges from C-G CAHPS 3.0.

CHPI-MHQP Short Form Survey	CG-CAHPS 3.0 Survey
1. Our records show that you got care from the doctor named below in the last 12 months. (Name of doctor) Is that right?	1. Our records show that you got care from the provider named below in the last 6 months. (Name of provider) Is that right?
	The questions in this survey will refer to the provider named in Question 1 as "this provider." Please think of that person as you answer the survey.
2. Is this the doctor you usually see if you need a check-up, want advice about a health problem, or get sick or hurt?	2. Is this the provider you usually see if you need a check-up, want advice about a health problem, or get sick or hurt?
YOUR CARE FROM THIS DOCTOR IN THE LAST 12 MONTHS	
These questions ask about your own health care. Do not incude care you got when you stayed overnight in a hospital. Do not include the times you went for dental care visits.	
	3. How long have you been going to this provider?
3. In the last 12 months, how many times did you visit this doctor to get care for yourself?	4. In the last 6 months, how many times did you visit this provider to get care for yourself?
4. In the last 12 months, did you phone this doctor's office to get an appointment for an illness, injury, or condition that needed care right away?	5. In the last 6 months, did you contact this provider's office to get an appointment for an illness, injury, or condition that **needed care right away?**
5. In the last 12 months, when you phoned this doctor's office to get an appoint for care you needed right away, how often did you get a appointment as soon as you needed?	6. In the last 6 months, when you contacted this provider's office to get an appointment for **care you needed right away**, how often did you get an appointment as soon as you needed?
	7. In the last 6 months, did you make any appointments for a **check-up or routine care** with this provider?
	8. In the last 6 months, when you made an appointment for a **check-up or routine care** with this provider, how often did you get an appointment as soon as you needed?
	9. In the last 6 months, did you contact this provider's office with a medical question during regular office hours?
	10. In the last 6 months, when you contacted this provider's office during regular office hours, how often did you get an answer to your medical question that same day?
6. Did this doctor's office give you information about what to do if you needed care during evenings, weekends, or holidays?	
MANAGING YOUR CARE	
7. In the last 12 months, how often did this doctor explain things in a way that was easy to understand?	11. In the last 6 months, how often did this provider explain things in a way that was easy to understand?
8. In the last 12 months, how often did this doctor listen carefully to you?	12. In the last 6 months, how often did this provider listen carefully to you?
9. In the last 12 months, how often did this doctor seem to know the important information about your medical history?	13. In the last 6 months, how often did this provider seem to know the important information about your medical history?
	14. In the last 6 months, how often did this provider show respect for what you had to say?
10. In the last 12 months, how often did this doctor spend enough time with you?	15. In the last 6 months, how often did this provider spend enough time with you?
11. In the last 12 months, did anyone in this doctor's office talk with you about specific goals for your health?	
12. In the last 12 months, did anyone in this doctor's office ask you if there are things that make it hard for you to take care of your health?	

There is also a difference in the suggested layout. The CHPI-MHQP short form is a 2-page survey with 1 cover sheet, whereas the CG-CAHPS 3.0 survey is laid out over 5 pages with a cover sheet.

CHPI-MHQP Short Form Survey	CG-CAHPS 3.0 Survey
YOUR EMOTIONAL HEALTH	
13. In the last 12 months, did you or anyone in this doctor's office talk about things in your life that worry you or cause you stress?	
COORDINATING YOUR CARE	
14. In the last 12 months, did this doctor order a blood test, xray, or other test for you?	16. In the last 6 months, did this provider order a blood test, x-ray, or other test for you?
15. In the last 12 months, when this doctor ordered a blood test, x-ray, or other test for you, how often did someone from this doctor's office follow up to give you those results?	17. In the last 6 months, when this provider ordered a blood test, x-ray, or other test for you, how often did someone from this provider's office follow up to give you those results?
16. Specialists are doctors like surgeons, heart doctors, allergy doctors, skin doctors and other doctors who specialize in one area of health care. In the last 12 months, did you see a specialist for any particular health problem?	
17. In the last 12 months, how often did this doctor (named in Question 1) seem informed and up-to-date about the care you got from specialists?	
OVERALL RATING OF DOCTOR	
18. Using any number from 0 to 10, where 0 is the worst doctor possible and 10 is the best doctor possible, what number would you use to rate this doctor?	18. Using any number from 0 to 10, where 0 is the worst provider possible and 10 is the best provider possible, what number would you use to rate this provider?
19. Would you recommend this doctor to your family and friends?	
OVERALL RATING OF CARE	
20. Using any number from 0 to 10, where 0 is the worst care possible and 10 is the best care possible, what number would you use to rate all your health care from all doctors and other health providers that you have seen in the last 12 months?	
	19. In the last 6 months, did you take any prescription medicine?
	20. In the last 6 months, how often did you and someone from this provider's office talk about all the prescription medicines you were taking?
	21. In the last 6 months, how often were clerks and receptionists at this provider's office as helpful as you thought they should be?
	22. In the last 6 months, how often did clerks and receptionists at this provider's office treat you with courtesy and respect?
ABOUT YOU	
21. In general, how would you rate your overall health?	23. In general, how would you rate your overall health?
	24. In general, how would you rate your overall **mental or emotional** health?
	25. What is your age?
	26. Are you male or female?
22. What is the highest grade or level of school that you have completed?	27. What is the highest grade or level of school that you have completed?
23. Are you of Hispanic or Latino origin or descent?	28. Are you of Hispanic or Latino origin or descent?
24. What is your race? Mark one or more.	29. What is your race? Mark one or more.
	30. Did someone help you complete this survey?
	31. How did that person help you? Mark one or more.

APPENDIX III: SURVEY AND DATA ELEMENTS TABLE

Survey and Data Elements Table

	MA STATEWIDE	MHQP PILOT	CHPI PILOT	CA PAS
SURVEY MATERIALS				
Total number questions	61	23	24	55
Patient comments	Web option	Web option/email	Web option/email	N/A
Email invite	N/A		✓	Optional/TBD by groups
SURVEY FIELDING				
Wave 1	4/30/2015	Email survey 5/13/15 with standard mail one week later.	Email survey 2/25/15 with reminder one week later. Non-respondents to email receive standard mailing two weeks following first email.	Email survey 12/15/14 with reminder one week later. Non-respondents to email receive standard mailing one month following first email.
Wave 2	6/2/2015	Email survey 5/20/15 with standard mail three weeks later.	Standard Mail 4/8/2015	Standard Mail 2/13/15
Wave 3	N/A	Email survey 5/26/15	N/A	N/A
CATI	N/A	N/A	N/A	3/13/15
Data collection closed	7/2/15	7/10/15	4/29/15	4/12/15
SENDER				
Health Plan		Email & mail	N/A	N/A
Group or Practice Site	N/A	N/A	✓	Email & mail