cahps® transcript

# Public Reporting of Patients' Comments with Quality Measures: How Can We Make It Work?

June 2014 • Webcast


**Speakers**

Steven Martino, PhD, Behavioral Scientist, RAND, Pittsburgh, PA

Rachel Grob, PhD, Senior Scientist, Center for Patient Partnerships and Department of Family Medicine, University of Wisconsin, WI

Mark Schlesinger, PhD, Professor of Health Policy, Yale School of Public Health, New Haven, CT

**Moderator**

Dale Shaller, Managing Director, CAHPS Database; Shaller Consulting Group, Stillwater, MN


## Dale Shaller

*Shaller (opening), Slide 1*

Good afternoon, or good morning, and welcome to our Webcast on Public Reporting of Patients' Comments with Quality Measures: How Can We Make It Work? My name is Dale Shaller, and I'll be the moderator for today's Webcast.

*Shaller (opening), Slide 2*

Our Webcast today is one in a series on CAHPS, which stands for Consumer Assessment of Healthcare Providers and Systems, produced by the CAHPS User Network. We know that many of you are familiar with the CAHPS program, but for those of you who aren't, just a few words of background. Funded primarily by the Agency for Healthcare Research and Quality, or AHRQ, the CAHPS program develops standardized surveys to assess patients' experiences with their health care in multiple settings, including ambulatory practices, and facilities such as hospitals and nursing homes. The CAHPS team also conducts research and develops guidance on how the results of CAHPS surveys can be used for public reporting, to support consumer choice of health plans and providers, and for guiding quality improvement efforts by health care organizations.

*Shaller (opening), Slide 3*

The focus of today's Webcast is on the reporting research conducted by the CAHPS team. We'll be addressing specifically the research that we've done on the use of patient experience narratives, or comments, in Web-based public reports on physician quality, new techniques that we've been developing and testing for collecting or eliciting patient comments, and possible approaches that we'll soon be testing for integrating patient comments along with other measures of quality in public reporting Web sites, and that can also be used in private feedback reporting by health care provider organizations.

Members of the CAHPS Reports Team include researchers from RAND, Yale University, and the University of Wisconsin at Madison, supported by both CAHPS funding and an R21 grant to Yale from AHRQ.

*Shaller (opening), Slide 4*

So why is this kind of reporting research important? Well, for starters we've seen a tremendous growth in recent years in the number of physician reporting Web sites that invite and post patient comments about their experiences with their doctors. Just a few examples are shown here and include Vitals, and RateMDs, Angie's List, Yelp. There are many others, and a recent estimate suggests that there are some 40 to 50 online sites in the U.S. alone that include doctor reviews or ratings.

*Shaller (opening), Slide 5*

At the same time, consumer interest in going to these kinds of Web sites with patient narratives is also taking off. For example, this chart shows a huge spike in U.S. traffic to the HealthTalkOnline Web site based in the UK, which features patient stories about their experiences with illness and getting care for treatment. According to a recent study by researchers at the University of Michigan and published in the Journal of the American Medical Association, about 25% of U.S. adults consulted online physician rating sites in the previous year. And more than a third of them went to a physician, or avoided one, based on the ratings.

*Shaller (opening), Slide 6*

For many of us that are concerned about the slow uptake of interest in public reporting sites based on standardized performance measures, such as CAHPS, or HEDIS, or Patient Safety, this surge in consumer interest insights with patient comments is actually quite heartening, but also poses some serious concerns. Since these comments are usually very small in number for any given doctor, and they're virtually never drawn from a representative sample of patients, so they can easily be skewed in either a positive or negative direction. Although most studies indicate posted comments are overwhelmingly positive about their doctors. And these comments while powerful and increasingly attracting a large audience, represent only a partial picture of overall physician quality, and should therefore be balanced with other measures.

So these concerns, along with the emerging demand for narrative comments, led the CAHPS Reports Team to want to look further into ways that both the collection and the reporting of comments can be improved. And we know that a large number of today's participants on the Webcast are from provider organizations. So I want to just point out that our research findings on the collection and reporting of comments can be applied as well to internal feedback reporting for improvement. And this is an area that we'll be looking at explicitly in the future.

*Shaller (opening), Slide 7*

To help explore our research and findings today we're pleased to feature three members of the CAHPS Reports Team today: Steven Martino, a behavioral scientist at the RAND office in Pittsburgh; Rachel Grob, Senior Scientist at the Center for Patient Partnerships and the Department of Family Medicine at the University of Wisconsin in Madison; and Mark Schlesinger, Professor of Health Policy at the Yale School of Public Health. And, again, I'm Dale Shaller. I'm a member of the Yale Reports Team, and Managing Director of the National CAHPS Database, serving as your moderator.

*Shaller (opening), Slide 8*

So before we begin, just a few of the standard housekeeping details. If you need help at any time during the Webcast, you can use the "Q&A" icon the lower right hand part of your screen. And you can join us by phone at any time using the number shown here, and entering the conference ID also shown here. Some of you may have trouble with your computer freezing during the presentations, and if that happens, you can hit your F5 button,

and your screen will refresh. And you may just be experiencing a lag in the advancement of the slides due to your internet connection speed. So you can try logging in and out, and that may help.

*Shaller (opening), Slide 9*

Given the large number of participants on today's Webcast, we had over 800 registrations, and that's terrific, we're going to be taking questions submitted online only. And to do that you can click that "Q&A" icon, again, at the lower right hand console on your screen. And the box will appear, and all you have to do is type up your question in the text box and select "Submit." So please do feel free to send your questions in at any time during the presentations, and we will address them during the Q&A session.

*Shaller (opening), Slide 10*

Today's slides are also available for downloading by clicking on the icon that says "Download Slides," again, on your console, and this will give you a PDF version of the presentation that you can download and refer to at any time.

*Shaller (opening), Slide 11*

We've also some additional resources posted that are available to you for your review under the "Resources" icon. And you'll find several things here, a couple of articles that have been published by the CAHPS Reports Team describing our research on patient narratives, a review article on the experience with patient comments in the UK, and a few other additional resources on public reporting.

*Martino, Slide 12*

So we've organized the Webcast in four parts today, and we're going to start with part one with an overview of what we've learned about the effects of including patient comments in public reports. And for this first segment I'm pleased to turn things over now to Steven Martino.

**Steven Martino**

*Martino, Slide 13*

Hello everyone. This is Steven Martino from RAND. I'm going to spend some time talking about an experiment that the CAHPS grantees recently completed in which we examine the impact of including patient comments in a Web-based public report of physician performance. Our focus was on physician performance, but the lessons learned are likely to apply equally to hospitals and other aspects of the health care system.

First, I will give you some additional background on this experiment. We know that systematically gathered and standardized data on patients' encounters with the health care system are valuable to consumers. But we also know that consumers tend not to seek out these data or rely heavily on them for making health care decisions despite the substantial investment that has been made in developing and fielding these standardized surveys. The same is true of standardized clinical performance measures.

We assume that patient comments also have some value to health care consumers. We make that assumption because of a great deal of research in the areas of decision making and consumer behavior. And this is research that is not necessarily focused on health care decisions. That research shows that narratives tend to evoke emotion in ways that numerical data do not, and that narratives may therefore engage people in ways that numerical data cannot. In addition, the detail that is characteristic of patient comments may promote increased understanding of the ways in which physician performance varies.

However, adding patient comments to a report that displays more objective numerical data on physician performance might also have certain drawbacks. It may, for example, increase the complexity of reports to the

extent that consumers become overwhelmed and withdraw. It may also distract consumers from paying attention to and using more objective and numerical data.

*Martino, Slide 14*

To address these questions we constructed a fictitious public reporting Web site to display comparative information on the performance of fictitious doctors. This Web site represents the context of our experiment. This Web site, which we called SelectMD, was designed to be consistent with real world public reporting sites in terms of its content, format, and functionality. Participants in this experiment were recruited from a Web based panel that is known to be representative of US households with internet access, which is to say almost all households.

There were about 850 participants in all, and these participants were randomly assigned to one of six experimental arms or conditions that involved different combinations of performance measures. So on SelectMD, some participants saw only CAHPS data on patient experience. Some saw CAHPS data along with clinical performance measures of the kind measured by HEDIS. And some saw CAHPS data, HEDIS data, and patient comments. We also varied the number of physicians for whom performance data were displayed, which explains why we have six conditions rather than three. But that's a nuance that I'm not going to focus on today.

Participants engaged with the SelectMD Web site in their own homes. They were told to look through the information that was provided, and to make a hypothetical choice of a doctor. There was a hidden tracking system that monitored what they did on the site and how long they spent in each area of the site. Participants also completed a survey before accessing the SelectMD site, which measured their prior experience with comparative performance data, and a survey after accessing the SelectMD site, which measured their satisfaction with the information provided on the site, their understanding of the information that was there, and their overall evaluation of the site.

*Martino, Slide 15*

Here is a picture of the site. The page that you're looking at is the "Performance Overview" page, which shows a summary of ratings for each doctor in the areas of service quality, or CAHPS, and treatment quality, which are clinical performance or HEDIS like measures. Quality was shown as ranging from one to five stars. As you can see there's a legend at the top of the page that explains what each number of stars means.

You can see that this "Performance Overview" page is layered over the top of three other pages of information. One of those pages presents detailed data on CAHPS. Another presents detailed data on clinical performance. And the last presents the patient comments, which we call patient reviews. Remember that not all participants were presented with all of these pages. You can also see on the upper right of the page that we provided the ability to filter doctors by gender and amount of experience, and to sort doctors by the doctor's last name or by either of the standardized performance measures.

We designed the site so that there was a small positive correlation between the CAHPS and clinical performance measures, and a stronger positive correlation between CAHPS and patient comments. We also designed the sites so that there were clear, better, and worse performers based on the standardized data.

*Martino, Slide 16*

This is what participants saw if they clicked on the "Patient Reviews" tabs. For this study we created about 150 comments. We modeled these fictitious comments on actual comments that we collected from RateMDs, and we tested these comments extensively to make sure that they were perceived as realistic, and to make sure that

they conveyed the level of positivity or negativity that we intended. So here on this page you can see a few examples of comments on Doctor Orson Alban.

*Martino, Slide 17*

This table shows the effect of including patients' comment on the SelectMD site along with CAHPS and clinical performance measures. As you can see at the bottom of the table, when comments appeared on the site participants spent a third more time on the site, and performed more than twice as many actions. There was also a non-significant but consistent tendency for participants to evaluate the site more positively when comments were presented versus when they were not.

*Martino, Slide 18*

However, when comments were present, participants spent less time probing for detail on the CAHPS and clinical performance measures. Thus, the increased time spent on the site by participants who saw comments was not spent drilling down to the components of the standardized performance measures. It was spent exploring the content of the comments.

*Martino, Slide 19*

Perhaps as a result, participants who viewed the site with comments versus without comments consistently made worse choices. For example, in conditions in which there were no clinical performance measures present on the site 61% of participants selected the doctor with the best CAHPS scores when comments were excluded, and 49% selected the doctor with the best CAHPS scores when comments were included. As you can see, these are fairly substantial differences in decision quality.

*Martino, Slide 20*

The results of this first experiment illustrate the potential promise and the potential danger of incorporating patient comments into a public reporting Web site. This led us to want to explore these two questions that you see here in depth. How can we obtain comments that are representative of patients' experiences, balanced in how they reflect that experience, and aligned with valued aspects of patient experience? And how can we report patient comments in a way that promotes integration with standardized measures and minimizes report complexity? Our hope is that by answering these questions we will discover ways to maximize the value enhancing aspects of patient comments and minimize their drawbacks.

So now, Rachel is going to say more about the first of these two questions.

**Dale Shaller**

*Grob, Slide 21*

Thank you, Steven. And just a reminder that you may submit questions at any time. We've received several already using the "Q&A" icon on the console at the bottom of your screen. And as Steven just mentioned, we now turn to part two of our story, which focuses on our research to understand how we can improve the collection or the elicitation of patient comments. And for this segment I'm delighted to turn things over now to Rachel Grob.

**Rachel Grob**

*Grob, Slide 22*

Good afternoon, everybody. This is Rachel Grob at the University of Wisconsin, Madison. And I'm going to focus with you on moving from anecdote to science with our elicitation study. Our goal for this research is to collect patients' reports of their health care that are representative, that are balanced, that are fulsome, by

which we mean really full and well developed, and that are understandable. And I'm going to walk you through each of these aspects.

I also want to note from the outset, that while this is our goal for the research, it's also an aspiration for public reporting. And I'm guessing that many of you on the call may share that aspiration, both for the promise of this work, and also for overcoming some of the difficulties that Steven just described for you.

*Grob, Slide 23*
OK, so we've got our goals defined here. And before I walk through the specifics of how we're addressing each of those, I just want to give you a sense of our experimental design.

*Grob, Slide 24*
Our elicitation study is focused on creating five to seven open ended questions. We want to limit response burden. So we were shooting for a completion time of less than 10 minutes. In the first round of our research it ran about seven minutes. We're now in the field with a second round of testing this design.

We have experimented with placing these qualitative questions that are eliciting the narrative data, the comments, both at the beginning of the CAHPS survey, and also at the end to see what effect that would have both on the quality of elicitation data and on the CAHPS survey itself. We have experimented with doing the elicitation on the telephone and on the Web, and we are going through multiple rounds of-- I alluded to, we have finished one complete round and analyzed this data, which Mark will be telling you about in terms of the results in a few minutes. And we are currently live with round two.

*Grob, Slide 25*
We are comparing the short elicitation, the seven question sequence, against hour-long, intensive interviews conducted by trained and experienced interviewers to create what we refer to as the "gold standard." So we're comparing each participant's response on the short elicitation to their own response to a long interview. In other words, how much data can we get in a very short amount of time relative to what we can get if we talk to folks for an hour and really find out a lot of detail about their experience.

We are assessing the quality of the elicitation according, both to its fidelity, how much, again, are we capturing of the scope, of the breadth, of the balance of positive and negative, and how useful is the elicitation data for someone else, a third party, who is reading this account that has been generated through the elicitation. And to create this comparison we're both doing traditional, textual, coding, which most qualitative studies employ as a methodology, and we're also doing a more narrative, or synthetic form, of analysis where we're looking at the whole elicitation, the whole story that was told in this seven question sequence of answers in terms of emotional expressivity, completeness, concreteness, chronology, consistence, and coherence, a whole series of narrative codes.

*Grob, Slide 26*
So I'm not going to spend time here walking you in detail through the round one design of the elicitation questions and where we got in round two, although we'll welcome your questions on that in the Q&A if you're curious, because instead I really want to give you a sense of how this elicitation is reading.

*Grob, Slide 27*
So if I were conducting this short elicitation with you on the telephone, I would start out by asking you, "What are the most important things that you look for in a health care provider and his or her staff?" So we begin there with the expectations.

Second, I would ask, "When you think about the things that are most important to you, how do your provider and his or her staff measure up?" So we want to know in the patient's own words what the match is between what they were expecting and what they were experiencing.

Then the third question probes on positive experiences, and it reads like this: "Now, we'd 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." And that probe is really designed to get a lot of that detail and quality.

Similar for question four, only we're looking for some of the things that don't go so well. So it reads, "Next 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."

And finally, our fifth question probes around the relationship between the patient and the provider, because in round one we know that our data in that respect were not as robust as they need to be. So we ask, "Please describe how you and your provider relate to and interact with each other." So you got a sense there of how this is being designed, and what we're taking into the field right now.

*Grob, Slide 28*

So going back to our goals and our aspirations for public reporting, what do we mean by representative? Well, in the study as we tried to create these elicitations that are maximizing benefit and minimizing danger we are collecting data from a nationally representative internet panel. We work with a company called GFK to do that. And we are stratifying sampling in the first round. We had an under representation of people with chronic or serious illnesses. So we oversampled for that in round two. We know that those folks have a lot of contact with the health care system, and a lot to say. We want to make sure it's working well.

And I know because I'm doing some of the interviewing we really are getting a nationally representative sample. We're talking to people with mental illness, with all levels of literacy. It's a good and robust panel. And we're also aspiring to induce higher future participation rates by making this engaging for the participants themselves.

*Grob, Slide 29*

What do we mean by balance? Well, we're really talking here about balancing the positive and negative experiences. Dale referred to this a little bit in his introduction in terms of how the field is looking. So clinicians fear the disgruntled patient. We know that actually comments are largely positive. We're trying to get a really good balance in that regard.

*Grob, Slide 30*

So what do we mean by fulsome accounts? We're talking here about a relative rather than an absolute standard, each patient really describing what matters to them, and comparing that elicitation against the interview. And we are, again, developing a coding process that can really capture that.

*Grob, Slide 31*

And here's just a little screen shot of one portion of our very complex and comprehensive coding scheme, the content of experiences. You can see the categories there.

*Grob, Slide 32*

And then as we zoom in, for example on emotional rapport, you can see some of the detail here.

*Grob, Slide 33*

And you may be thinking, as you look at this, what's the difference between warm and caring, and friendly and nice, and respectful and professional? But, in fact, there really are a lot of nuances in the data, and it matters quite a bit. We've worked hard to parse things carefully so that we're really capturing the fulsomeness of the accounts we're hearing.

*Grob, Slide 34*

Finally, we want these accounts to be understandable for the reader, for that third party, so they really can perceive what the patient experienced, and a little bit about who that patient is, and why what matters to them matters. So I'm going to give you two brief illustrations here.

The first is an illustration of the distance traveled from the quantitative CAHPS survey alone to the survey with the qualitative elicitation attached. And our participant, who we'll call Jane, responded to our question about the survey in general as follows. She said, "Based on your questions, it appears I am getting what I need, yet I don't feel satisfied." That was for the CAHPS survey.

Then when she proceeded to respond to the elicitation here is what she actually said about her experience with her provider and his or her staff. "I saw my provider for my yearly physical. I find her aide to be personable and friendly, easy to talk to. My doctor is somewhat remote. She is not terribly sympathetic, and has little time. I believe she is competent, which is why I still go to her. I have no major health problems, but if I needed to see a doctor often I'm not sure I would be satisfied. I do always get a response when I call with a problem. Medicine today is so messed up in this country it is distressing."

Okay, so you can see the difference between her saying she didn't feel the survey captured her own feeling, and then what happened when she got a chance to speak for herself. That's kind of the distance we feel we've traveled. Where we're still going is illustrated by my final example here, which is a comparison between what we got on the written version of the Web elicitation, again in round one, and what we heard in the interview.

So Timmy, as we'll call him, wrote very briefly in his elicitation, his response to our seven questions, "I had my physical and it went well. Staff was professional. Everything was fine." And he gave a high rating to his provider. However, when we interviewed Timmy we found about 40 minutes in that he actually had more than a physical. He had had an elective minor surgical procedure. He has had a vasectomy, which he didn't want to tell us about until far into the interview. And he had a lot more to say about his doctor. Here's what he said: "I appreciated it, because we don't have children. And we made a decision not to have children. And we were both concerned going in to talk to a doctor about doing this that they might consider not doing it since we didn't have kids. And he didn't. He asked us our reasons, and we talked through it. And that was about it. There wasn't any kind of judgment, or anything like that. So I was very, very pleased with that kind of a consultation. I also liked the fact that he brought my wife in, and it wasn't just a conversation with me and him. He wanted her there at the consultation." So you can see how understandable that would be, why this gentleman had given his provider a high rating, which wasn't really captured by him saying everything was fine, but really was described in this kind of story.

We know we can't get all the way to what we can get in an hour-long interview with the short elicitation, but we're going to push it as far as we can. And I know Mark is going to tell you a lot more about our results for round one. Back to you, Dale.

**Dale Shaller**

*Schlesinger, Slide 35*

Great. Thanks, Rachel, so much for the overview of the elicitation protocol. We now follow up to part three of our Webcast. And Mark Schlesinger will lead us through a look at how some of our preliminary findings are showing how well elicitations compare to an in depth set of "gold standard" interviews. Mark.

**Mark Schlesinger**

*Schlesinger, Slide 36*

Thanks, Dale. Good afternoon, everyone. Now we get to the exciting bottom line. How well can we actually do with this six-question, seven-minute little sequence of open-ended questions compared to the hour-long interview done by very sophisticated interviewers like Rachel? And so for much of the criteria that we're looking at, our measures of fulsomeness, of balance, of understandability, what we're going to do, as Rachel suggested, is to compare what we can get from the elicitations, either over the Web or the phone, against the content we can get from these detailed, hour-long interviews.

But when we focus on the question of representativeness we're going to look somewhat more broadly than that, because we had limited resources. We conduct only about 54 intensive interviews. So that's going to be the basis for the first of these tests. But in looking at questions of representativeness we were able to do an additional set of Web elicitations that gave us a large enough sample size we could compare across sociodemographic categories, across racial groups, across age groups, across gender, and get a better sense of who was being more or less responsive to these elicitations.

*Schlesinger, Slide 37*

So let's start with the bottom line. All told we did pretty well, remembering that this is a multi-stage process, as Rachel described, and the results we're reporting to you this afternoon are from the first round of the elicitation, from which we learned and revised our elicitation protocol, and are back in the field as we speak right now with the second round. But these results are all from the first round. What we found was the following, that in terms of fulsomeness, in representativeness, we did, I would say, not badly. And I'll give you a little more sense of that in just a moment.

But in terms of balance, that is of getting the right positive and negative mix of comments, that is a mix on the elicitations that match those from the interviews, we did remarkably well I think, as well as doing really, pretty, quite well on the kind of understandability or narrative coherence of these comments. And, again, we'll go into details on all four of these in just a second.

The one consistent pattern that came through all four of these different criteria are that there was a striking difference between the telephone elicitation and the Web elicitation, even though they both used the exact same questions, the exact same wording, and the exact same sequence. As you will see, phone elicitation, that is talking to someone over the phone, vastly outperformed, at least in round one, the Web elicitation with a little bit of variation that we'll see in just a moment.

So we'll start with fulsomeness.

*Schlesinger, Slide 38*

We had a number of measures, but our basic measure of fulsomeness was to look across those 10 different categories of experience that Rachel had listed for you earlier and just ask the simple question. If someone talked about having had an experience with the doctor that touched on one of those 10 categories when they

were describing things in their hour-long interview, did we also pick that up in the elicitation? And overall, as you can see from the dark purple bar, just over 40% of the time we did.

So even though we had these seven-minute elicitations compared to an hour-long interview, we were getting about 40% of the episodes being reported, with, again, somewhat better performance, the dark blue bar from the phone elicitations. About 45% of the categories are being reported, captured in the elicitation that were reported in the interview, compared to about 35% for the Web elicitation. But what was really striking was the variation you saw when you looked across the categories of patient experience.

*Schlesinger, Slide 39*

What you find are some categories where we did quite well across the board. So if you look, for example, at the categories of caring, at competence, at interactions with staff, what you'll see is we were picking up with the elicitations about 50% to 60% of the events that were reported in the interviews, and we did so very consistently between phone and Web. Those bars are all about at the same height. So for some domains of experience we were doing really well on the first run.

Then we had another set of domains of experience, captured here by access, by communication, by how much time the patient reported having in communicating with their doctor, where in some ways we did even better on the phone elicitations, those bright blue bars. We were getting 55%, 60%, 70% of the events reported in the interview were being captured by the elicitation. But in these domains you could see there's a big drop off for the Web elicitations, that white purple bar, which suggested, even though the questions were exactly the same, we weren't getting the same response with that mode.

And then we had a third set of domains of experience captured by orientation. That's basically physician practice style, thoroughness, shared decision making, and coordination of care, where we weren't doing all that great with either of the elicitation modes. Now, you might look at thoroughness and shared decision making and say, "Oh my gosh, they're doing terribly." Turns out that people didn't talk much about either of those two domains. So even though we asked about it, even though we probed for it in the interviews, they just didn't have that much to say. So those two domains not quite so salient for the public. And we'll come back to that as we move forward.

But clearly what we have that is a protocol that's performing pretty well, but in at least some selected areas not quite as well on the Web side. And so when we moved to the round two elicitation we redesigned the elicitation, partially in the ways Rachel described to address that.

*Schlesinger, Slide 40*

Second domain, or second measure of performance, balance. Here, as I suggested earlier, we really did quite well. But here we have a somewhat different metric for comparison. So you see across this particular graph there's a bright green line right at the 1.00 level. That's because what we're doing is comparing the mix of positive and negative responses that you find in the elicitation to that same mix that you find in the interview. So by the way this measure is constructed, if the bar goes above the green line that means the elicitation is biased positively. Get more positive relative to negative in the elicitation than you got in the interview. If it's below the green line, then you're getting a negative bias, or more negative comments relative to the positive in the elicitation than you got in the interview.

Now, we measured this particular dimension of performance in a couple different ways. We could use textual codes that Rachel responded to. We just count up the number of mentions there were. Or we count up the number of lines in the transcript that were devoted to positive versus negative characterizations. And so those

are the two clusters on the left. And what you see is that we did pretty well overall, with the phone elicitations being slightly negatively biased, the Web elicitations slightly positively biased, and on average they turned out to be pretty consistent.

But we can also measure them in a different way, using the narrative codes that Rachel described. Those are having our coders go back, and just kind of assess everything that got said. How did that sound overall when you were talking about the doctor, or talking about the doctor's office. Those are the two clusters of results in the middle of the graph. And there you can see we did quite well indeed. Very strong fidelity between the elicitations and the interviews with still a little bit of positive bias on the Web elicitations.

*Schlesinger, Slide 41*

On to our third measure, understandability. This is, in a nutshell basically, how well can you understand the story of what happens to people when they go to the doctor's office. And we measured this in a variety of different ways, as Rachel mentioned, which are summarized for you on the left in this aggregate category of coherence. Coherence capturing a combination of a half dozen different aspects of thinking about how well the story is told.

And so, here again, we did pretty darn well. Indeed, in the telephone elicitations we got about 70% of the kind of story content using the elicitations that we did even with the hour-long interviews. But here you see this quite large drop off once again with the Web elicitations. You see that equally much if we look at components of coherence, texture, or how much detail is conveyed, or the completeness of the story, how much do you kind of know what happened to the patient kind of from soup to nuts when they first get to the doctor, versus the outcome of their doctor's visit. So overall pretty good, but, again, a little bit of drop off with the Web elicitation mode.

*Schlesinger, Slide 42*

Finally, for our fourth and final measure of performance, representativeness. How well can we get people to actually play the game, to participate? Here the answers look pretty good with one or two weak points. In terms of pure participation, just getting people to respond to elicitation we got good across the board responses from people in various educational and sociodemographic groups. So people participated well.

But if you look at the depth of their participation, the extent to which they were talkative or responsive, people like me, man, the more taciturn gender, were a little less responsive than women were. People with lesser education, a little less responsive to people who had more education. And those were concerns. And we identified what we believed to be the reasons for some of those biases. Both those groups needed a little more scaffolding, a little more superstructure to the questions. And so in the round two elicitations we built those two things in.

*Schlesinger, Slide 43*

So recall that all these results are coming from the first round of elicitations, which is essentially our test bed where we were first trying out the elicitation protocols. We took a variety of lessons from those, from the comparative performance, from where things worked well, and where things worked less well, and we incorporated all of those into the second round elicitation. Now, because we're just in the field with that now we don't have much to report on this, but the preliminary results we do have suggests that the second round is looking pretty good, that we're getting about 40% more content with the Web elicitation. Remember, that was kind of the weak point of the first round. And that content involves substantially more about expectations, emotions, and the relational aspects of people interactions with their doctors, all of which were somewhat weak points in the first round.

All right, that takes us through the results. And now, we turn things back to Steven.

### Dale Shaller

*Martino, Slide 44*

All right. And let me just interject that our plans are to move forward now to look at ways we can improve the public reporting of comments. Assuming we can get more better improved comments collected or listed through the techniques that Mark and Rachel have described. So Steven, we're going to ask you to round out the story with part four. And we have a lot of questions that we'd like to get to. So we'll try to get to those as soon as possible. Steven.

### Steven Martino

*Martino, Slide 44*

Okay, thanks, Dale. I'm going to end by giving a very brief introduction to the next version of the fictitious SelectMD Web site, which we refer to as SelectMD 2.0.

*Martino, Slide 45*

This site mimics its predecessor in many ways, although we've updated the design and functionality of the site to be consistent with how Web sites have evolved in the past few years. We still use the site to present data on CAHPS, clinical performance, or HEDIS, and patient comments. As in our prior experiment, participants will be randomly assigned to one of several experimental arms or conditions. The arms that I will focus on here are ones that involve various ways of incorporating patient comments.

As Dale said, a main aim of this experiment is to find a way or ways to report comments in such a way that maximizes their value enhancing characteristics, and minimizes their drawbacks, as identified in our first experiment. So there are three ways that we are incorporating patient comments in the SelectMD 2.0 Web site. One way is just as we did in SelectMD 1.0. A second way is in a style that's reminiscent of how Amazon provides comments on its consumer goods. And a third way is with tags, content tags, that allow users of the site to choose to see topics of particular interests. And I'll show you examples of each of these in a moment.

*Martino, Slide 46*

First, here is the landing or introductory page of the new SelectMD Web site. We present this just to give you a sense of the ways in which the site has been updated stylistically.

*Martino, Slide 47*

Here is a shot of the SelectMD 2.0 Performance Overview page. You can see that we are now making use of icons. We've renamed the measures. So, for example, what was "Service Quality" is now "Use of Effective Treatments." And we've added the dimension of "Patient Safety," or as you can see here, "Methods to Reduce Medical Errors."

*Martino, Slide 48*

Here you can see what the Performance Overview page looks like for those participants who will be assigned to the Amazon comment presentation style. If you look here in the column with the "What Patients Say" header, you can see that for each doctor a user gets to see not only the total number of comments that are available for that doctor, but also how they are distributed across categories from very positive to very negative.

*Martino, Slide 49*

If a user clicks to see comments for Doctor Dorinda Bekki, this is the page that they will see. So on the left you can see the distribution that was also present on the Overview page. And on the right you can see each of the comments with a header that indicates the valence of the comment: Mixed, Negative, Very Negative, and so

forth. You can also see at the top of this page that we have provided a way to filter comments by a particular valence. So, for example, a user may choose to see only the most positive and the most negative comments about a doctor get a sense for the extremes, or they can choose to see all of the comments.

Our hope is that by providing this functionality we are giving users a way to explore the comments that seems less overwhelming than when the comments are presented in a random fashion, as they were in SelectMD 1.0, and as they are in one of the arms of this experiment. Also we hope that showing the distribution of comments helps to get people thinking about the range of comments that are available and how coherent or variable that set is.

*Martino, Slide 50*

Finally, here you can see in the tagged comment arms, in the comment that presents, in the condition or arm that presents comments with content tags the page that shows the number of comments available for each doctor.

*Martino, Slide 51*

And then, if you were to click on a particular doctor, you can see exactly what the comments are. And underneath each of the comments are the content tags. So for this particular doctor, Doctor Dorinda Bekki, you could see that the first comment is, "If you can get beyond the wait time and the irritable office staff, Doctor Bekki is a nice, competent doctor." Underneath of that the content tags are "Clinical Knowledge and Skill," "Doctor-Patient Communication", and "Office Staff."

And you may notice that those content tags are reminiscent of the names of CAHPS composite measures. And we do that on purpose as a way to perhaps draw people's attention to the fact that there is that correspondence, and perhaps get people interested in exploring the CAHPS measures in more depth. You can also see at the top of this page that participants have the opportunity to see only comments about particular aspects of patient experience, or they can choose to see them all. As I said, these are just a few of the conditions to be included in this experiment, which we hope to begin in the fall. And we look forward to sharing the results of this experiment with you, in one way or another before too long.

**Dale Shaller**

*Shaller (closing), Slide 52*

Okay, great. Steven, thanks so much, and Mark and Rachel. We've covered a lot of ground. We have a lot of questions. Again, this is how you ask a question if you haven't done so already. We'll do our best to get to everyone that's come through here.

So let me start with one question that's kind of overarching, which is the goal to add additional open ended questions to the current CAHPS family of surveys or something separate from the actual survey end. Our thoughts are to do the kind of testing that's been described here for improving elicitation of comments, and find ways to embed the collection within the context of the close-ended CAHPS survey. And we still need to work out the details of how that would be done logistically and operationally, whether to include that as a supplemental set at the end. And this is why the further research is underway, to figure out what is the best way to methodologically place these questions into the survey, and ask the questions that yields the most complete and representative set of responses.

So we have a number of questions. I'm going to start with a few to Rachel. A couple are similar in nature. Some surprising kind of reactions that why, Rachel, did many subjects have no or few negative comments? And a

related question is, "I would've thought that the phone could generate a greater percent of positive comments than the Web".

**Rachel Grob**
Yeah, very good questions, and things that our team has spent a lot of time thinking about. So we felt that we didn't get a representative number of negative comments actually in round one. And we think there are a few reasons for that. One is that the way we were asking people which provider among the providers that they see they wanted to choose we thought might allow them to pick the one that they were proud of going to, rather than the one that they'd be embarrassed to say they still go to.

And, of course, this isn't a problem with the research, and not a problem in the real CAHPS world, where we tell people which physician or clinician group they are being asked to discuss on the survey. So it's a problem with the research. And the sampling frame that we showed you, where we're trying to go after people with serious, or life threatening, or chronic illnesses, was partly designed to compensate for that, as well as some of the changes that we made in how people are picking their focal provider, and especially in the interviews, this was tougher in the elicitation, just being sensitive to the fact that it may be embarrassing for people to admit that they're still with a provider about whom they have negative comments.

Mark, do you want to add anything about the negative versus positive balance on phone versus Web from the round one data?

**Mark Schlesinger**
Going in we had very mixed expectations. We could tell stories that would lead us to predict either way that the Web would be easier to say negative things about, or the phone would be easier to say negative things about. And so the results, although counterintuitive in some ways, were completely intuitive in other ways. So I think I'll just say that it's complicated enough how you elicit these things, that you can easily imagine how it could go in either direction.

**Dale Shaller**
Mark, here's a question for you with respect to the analyses that you presented. Are the differences statistically different or just random variation? There are no confidence intervals presented. So it's hard to determine.

**Mark Schlesinger**
Right. And those are good questions. For all of the results that we were describing as significant differences between the telephone and the Web, which you remember were often gaps of 20 or 30 percentage points in terms of performance. Those were all statistically significant. In cases where you had combinations where things were a little closer together, like in a few of the domains where things were 5% or 10% different, that's borderline in terms of statistical significance, recognizing we have relatively small sample sizes here.

**Dale Shaller**
I'm going to go back to Rachel with this question. Rachel, when probing positive or negative experiences have you considered the impact of the order in which the experiences are asked? Say, for example, positive impressions first versus negative impressions first and vice versa, which is I know something we have done.

**Rachel Grob**
Yeah, we definitely have considered that. And we sort of began with the assumption that it would be good to sort of get people rolling with the positive experiences instead of going in right away to things that might be more sensitive or difficult for them to talk about. But it's actually a really interesting consideration. We just had so many variables that we're balancing in our design that trying to do, you know, part of the sample with

negative first, and part with positive first seemed like it would be confounding and not give us the statistical power that we needed.

But I think it's a good point, and something we should think about in future. I will say that in the intensive interviews, which I've done a lot of them myself, people will bring up whatever is most salient to them first. And even in this short elicitation, as you saw in that second question when we ask, how does your provider measure up, they can offer us their negative experiences first. So we will note that in our analyses and take account of it.

**Dale Shaller**
Thank you. I'm going to ask Steven to respond to this question. Steven, the question is, if there's any research that looks at the overall ratings and the likelihood to provide comments, are those that, for example, are more or less satisfied with their overall experience more likely to provide comments? Do we know anything about that?

**Steven Martino**
Yes, in fact, there are several teams of researchers who have done research in this area, and who have found that comments on public Web sites are overwhelmingly positive. The large majority of them are positive. And so in our elicitation research one of the things that we've tried to do is to craft questions in a way that gets people thinking about negative experiences, and providing them with an opportunity to tell us about them.

So in the elicitation protocol we tell people that we'd like to focus on any experiences that you wish had gone differently over the past 12 months, and to explain to us what happened, how it happened, and how it felt to you. And so we found that providing questions like this is an effective way to get people talking about things that they might not otherwise talk about.

**Dale Shaller**
Thank you. I want to pose this question with respect to the SelectMD and the reporting of comments that we've studied. I believe this question relates to that aspect of the research. The question is, how many different panels were used to select the respondents, and what criteria were used to screen the respondents if any? How many respondents participated in the SelectMD 1.0 setting?

**Mark Schlesinger**
So in the SelectMD 1.0 research, it was, again, a randomized panel of, in that case, people below Medicare age, so adult population age 21-64. In that particular research we also limited it to people who had computer based access to the Web on the ground that those would be the types of people most likely to encounter the kinds of Web sites like SelectMD 1.0. And we had a total of 130 in each of six arms. So a total of about 780 or 800 people all told participated.

**Dale Shaller**
All right. There's another question related to the reporting and the assignment in the 2.0 version, which we'll be testing soon. Steven, how do you determine the difference between a negative and a very negative comment?

**Steven Martino**
We have a set of raters who read the comments and provide ratings on their valence. We write the comments in a way that they have a certain expected valence, but we have independent coders code the comments. And we make sure that we have good agreement among those coders about what constitutes a negative or very negative comment.

**Dale Shaller**
Thank you. Here's an interesting question regarding the improving elicitation aspect of our work. The question is, does it matter if a sample is random or representative if folks are mentioning an aspect of care that really needs improvement, and the clinic or the facility actually knows it needs it?

I guess I'm throwing that to anybody who wants to grab it.

**Mark Schlesinger**
OK. Well, you could be just as useful someone to grab that one as well as any of us. But let me take a first stab at it. I think it's a very good question. Part of the power of these more detailed qualitative elicitations is that it's not a matter of just counting up the number of aggravated people and saying, oh gosh, we have a lot of unhappy people here. You can actually identify detailed information about what went wrong in various ways, suggesting how, at least arguably, it might be more constructively addressed. And so that's very much the premise of focusing on this sort of data, that the detail matters. And so one very powerful case may be a completely important ground for action.

**Rachel Grob**
Yeah, and I just wanted to add to that that we think it's important to be eliciting the experiences of those who won't necessarily volunteer them, because they're the most irate or aggravated, because their experience also matters. So it's kind of overcoming the only listening to the squeaky wheel approach.

**Dale Shaller**
I'm really sorry. We have to wrap things up. We've had so many really good questions, and we've only scratched the surface. We'll try to get back to you.

*Shaller (closing), Slide 53*
If you would like to send your question to us again, we'll definitely want to respond to that. If you want to sign up for regular email updates about the CAHPS program you can go to the AHRQ Web site, as shown here, and you can sort of find an icon that allows you to sign up for regular updates about all the activities under the CAHPS program.

*Shaller (closing), Slide 54*
We do have an evaluation survey I mentioned. You hit the icon, which is on your console. It's not an exit survey. You need to hit it before you leave the Web site. We really do ask you to fill that out, and to click the "Submit Survey" when you're done with the survey.

You can connect with us at any time by email at the address shown here, or by our toll free number shown here as well, and the CAHPS Web site, which is always open for business.

*Shaller (closing), Slide 55*
I just want to thank everyone on behalf of AHRQ, the CAHPS Consortium, Steven, Rachel, Mark for your presentations, and to all of you for joining us today.

This concludes our Webcast. Have a wonderful rest of your day.