What Patients Tell Us about Primary Healthcare Evaluation Instruments: Response Formats, Bad Questions and Missing Pieces

Detailed Report of published article

Corresponding Author:
Jeannie L. Haggerty
Associate Professor
Department of Family Medicine
McGill University

Postal Address:
Centre de recherche de St. Mary
Pavillon Hayes – Bureau 3734
3830, av. Lacombe
Montréal (Québec) H3T 1M5
Canada

Contact:
Tel : (514) 345-3511 ext 6332
Fax : (514) 734-2652
jeannie.haggerty@mcgill.ca
What Patients Tell Us about Primary Healthcare Evaluation Instruments: Response Formats, Bad Questions and Missing Pieces

Abstract

Instruments have been developed that measure consumer evaluations of primary healthcare using different approaches, formats and questions to measure similar attributes. In 2004 we concurrently administered six validated instruments to adults and conducted discussion groups with some of them to explore how well the instruments allowed them to express their healthcare experience and to get their feedback about questions and formats.

Method: We held 13 discussion groups (110 participants): nine in metropolitan, rural and remote areas of Quebec; four in metropolitan and rural Nova Scotia. Participants made notes about their healthcare experience over the previous year, then responded to all six instruments under direct observation, then participated in guided discussions for 30 to 40 minutes. The instruments were: the Primary Care Assessment Survey; the Primary Care Assessment Tool; the Components of Primary Care Index; the EUROPEP; the Interpersonal Processes of Care survey; and part of the Veterans’ Administration National Outpatient Community Satisfaction Survey. Discussion transcripts were analyzed for content by two team members. Item response theory analysis was conducted to verify quantitative behaviour of qualitative insights.

Results: While respondents appreciated consistency in response options, they preferred options that vary to fit the question, as in the Primary Care Assessment Survey. Likert response scales functioned best; agreement scales were least appreciated. Questions that average experience over various providers or over many events diluted the capacity to detect critical negative or positive incidents. Respondents tried to answer all questions but stressed that they were not able to report accurately on elements outside their direct experience or in the provider’s world. They liked short questions and instruments, except where these compromise clarity or result in crowded formatting. All the instruments were limited in their capacity to report on the interface with other levels of care.

Conclusion: Each instrument has strengths and weaknesses and could be marginally improved, but respondents accurately detected their intent and use. Their feedback offers insight for instrument development.
Health service planners and providers administer evaluation instruments to patients to capture their experience systematically and reliably in order to guide healthcare improvements and monitor innovations over time. Selecting the most appropriate instrument involves choices about measurement approaches and response formats. The measurement approach is categorized generally as rating or reporting. Rating questions ask the respondent to make a judgment about the acceptability or quality of an attribute, judging whether standards or expectations – usually not made explicit – are met. The reporting approach elicits the occurrence of indicator events during care and then evaluators judge whether services conform to an explicit standard or expectations of service acceptability or quality.

Instruments often have different response scale formats. Likert scales provide response options arranged ordinally with a label (e.g. very good, often) attached to each. Semantic differential scales attach labels only to opposite poles of a continuum (disagree strongly, agree strongly) separated by discrete but unlabelled response options. In self-administered questionnaires, formatting decisions affect question comprehension and influence instrument performance.

In 2004 we concurrently validated six instruments that assess primary healthcare from the patient perspective. We report here on qualitative insights gained from debriefing a subgroup of respondents who responded to the questionnaires in a group setting then shared their reactions and discussed how well the instruments allowed them to express their primary healthcare experience. We describe each instrument briefly, summarize the discussion groups’ reactions and link these to recurrent findings from an analysis of the performance of responses for individual items.

**Method**

The method, described in detail elsewhere (Haggerty et al., 2009), consisted in administering validated and widely-used evaluation instruments to 645 respondents in Quebec and Nova Scotia, in French and English respectively. Study subjects were healthcare users with a regular provider, recruited from previous healthcare surveys, newspaper advertisements and community posters.

The six study instruments compared were (in order of presentation in the questionnaire): 1) the Primary Care Assessment Survey (PCAS); 2) the Primary Care Assessment Tool (PCAT, short, adult version); 3) the Components of Primary Care Instrument (CPCI); 4) the EUROPEP; 5) the Interpersonal Processes of Care (IPC) survey; and 6) the Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCCS). We retained only subscales of attributes addressed in more than one instrument. We depicted each instrument as closely as possible to the original version in format, font, and instructions to participants.

Thirteen discussion groups were held: three each in metropolitan, rural and remote areas in Quebec and two each in metropolitan and rural areas of Nova Scotia. Participants first had a few minutes to jot down notes about their experience of healthcare over the previous year, and then they responded to the questionnaire, while their reactions and time were directly observed. They then participated in a guided group discussion lasting 30 to 40 minutes, which was recorded. The discussion guide, the same for all groups, asked which questionnaires and questions were most and least liked, were some questionnaires or questions confusing or difficult and how well did the questions allow them to express the essence of any critical incidents. The group facilitators made briefing notes, and recordings were transcribed and
analyzed independently for content by three members of the research team who did not participate in the discussions (CB, MF, JH). Issues raised in most group discussions were retained as important.

We also conducted non-parametric item response theory (IRT) analysis of the items based on item response theory (IRT) for a fine-grained examination of how participants responded to individual items and response options, including “don’t know” responses (Ramsay, 2000). IRT graphs model the probability of endorsing the response options of a single item (e.g., “how would you rate the wait time for appointments?”) as a function of the total score on the latent variable or construct being assessed (e.g., accessibility). Ideally, the probability of endorsing each response option is highest in a distinct zone of the construct (1 = poor is distinctly concentrated in low accessibility; 6 = excellent, in high). We adapted this method to examine the response patterns of both missing values and “don’t know” responses to see if their probability is higher in one region of the latent variable.

Results
In total, 110 subjects participated in 13 focus groups (average nine per group); 64.6% were female, average age was 49.7 (SD: 12.8) years and 67% had a post-secondary education. Overall experience of care was excellent for 49%, average for 38% and poor for 13%, as assessed by the question: “In general, would you say that the care you receive at your regular clinic is: excellent, average or poor?”

In the results presented below, citations are identified as being from Nova Scotia [NS] or Quebec [QC], and all citations from Quebec have been translated from French.

Formats and response scales by instrument
1. Primary Care Assessment Survey (PCAS)
The PCAS (Safran et al., 1998) (Figure 1) was developed by Dana Safran to measure the achievement of the Institute of Medicine’s definition of primary care: “the provision of integrated, accessible healthcare services by clinicians who are accountable for addressing a large majority of personal healthcare needs, developing a sustained partnership with patients, and practicing in the context of family and community.” The instrument applies only to respondents with a “regular personal doctor” and uses a predominantly rating approach. It has a large, easy-to-read font; readability is at grade-4 level. The Likert scale labels change by question context, as does item presentation.
**Primary Care Assessment Survey:** Sample items from different subscales demonstrating the structure of rating-type Likert responses on a 6-point scale.

1a. How would you rate the usual **wait** for an appointment when you are sick and call the doctor's office asking **to be seen**?

- Very poor
- Poor
- Fair
- Good
- Very good
- Excellent

<table>
<thead>
<tr>
<th>Very poor</th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Very good</th>
<th>Excellent</th>
</tr>
</thead>
</table>

Thinking about **talking** with your regular doctor, how would you rate the following?

<table>
<thead>
<tr>
<th>Very poor</th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Very good</th>
<th>Excellent</th>
</tr>
</thead>
</table>

1b. Doctor's **explanations** of your health problems or treatments that you need

<table>
<thead>
<tr>
<th>Always</th>
<th>Almost always</th>
<th>A lot of the time</th>
<th>Some of the time</th>
<th>Almost never</th>
<th>Never</th>
</tr>
</thead>
</table>

1c. How often do you leave your doctor's office with **unanswered questions**?

<table>
<thead>
<tr>
<th>Always</th>
<th>Almost always</th>
<th>A lot of the time</th>
<th>Some of the time</th>
<th>Almost never</th>
<th>Never</th>
</tr>
</thead>
</table>

Thinking about how much you **TRUST** your doctor, how strongly do you **agree** or **disagree** with the following statements:

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>Not sure</th>
<th>Disagree</th>
<th>Strongly disagree</th>
</tr>
</thead>
</table>

1d. My doctor sometimes **pretends** to know things when he/she is really not sure

<table>
<thead>
<tr>
<th>Agree</th>
</tr>
</thead>
</table>

1e. I completely trust my doctor’s **judgments** about my medical care

<table>
<thead>
<tr>
<th>Agree</th>
</tr>
</thead>
</table>
This instrument was most consistently preferred by respondents. It is easy to read and the adaptation of response scale options to the nature of the questions adds variety to the questionnaire—“made me more alert to answering” [NS]—without feeling “tricky.” The item frame (e.g. “Thinking about talking with your regular doctor, how would you rate ...”) is in close proximity to the relevant question so that respondents are not required to keep in mind a frame presented only at the beginning of the questionnaire. However, participants did find the Trust subscale questions irrelevant or requiring too much guessing (Figure 1d). They thought it unlikely that patients who did not trust their doctor would stay long enough to evaluate him/her: “if you thought that you wouldn’t be with that doctor, and if the doctor hides it from you, you don’t know. So you couldn’t answer “yes.” [NS]

The IRT analyses showed that, with few exceptions, the response scales work well. Figure 7a illustrates the response behavior for an item from PCAS Organizational Accessibility (corresponding to Figure 1a), which is typical many PCAS items. Response options are most likely to be endorsed in specific regions of the latent variable, such that the response options are clearly ordinal, mostly with equivalent intervals between them.

2. Primary Care Assessment Tool (PCAT)

The PCAT (Figure 2) was developed by Barbara Starfield and colleagues to assess the extent to which care is consistent with four unique and three essential attributes of primary care. Initially a pediatric tool (Starfield, Cassady, Nanda, Forrest, & Berk, 1998; Cassady CE, 2000), it was adapted and validated for adult care (Shi, Starfield, & Xu, 2001). The items refer to the “Primary Care Provider,” but respondents (and evaluators) are guided to identify this provider who is the regular source of care and/or knows the patient best and/or takes responsibility for most care (Strength of Affiliation scale). It can be used by those without a regular provider.
Figure 2
Primary Care Assessment Tool: Sample items from two subscales showing a 4-point Likert response scale with the same option labels for all items. Italics denote emphasis and contrast between similarly worded items in a subscale.

**FIRST CONTACT – ACCESS**

<table>
<thead>
<tr>
<th>Please check the one best answer.</th>
<th>Definitely</th>
<th>Probably</th>
<th>Probably not</th>
<th>Definitely not</th>
<th>Not sure/don’t remember</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a. When your Primary Care Provider is <em>open</em> and you get sick, would someone from there see you the same day?</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2b. When your Primary Care Provider is <em>closed</em>, is there a phone number you can call when you get sick?</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

**ONGOING CARE**

<table>
<thead>
<tr>
<th>Please check the one best answer.</th>
<th>Definitely</th>
<th>Probably</th>
<th>Probably not</th>
<th>Definitely not</th>
<th>Not sure/don’t remember</th>
</tr>
</thead>
<tbody>
<tr>
<td>2c. When you go to your Primary Care Provider, are you taken care of by the <em>same</em> doctor or nurse each time?</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2d. If you have a question, can you call and talk to the doctor or nurse who knows you best?</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2e. Does your Primary Care Provider know you very well as a <em>person</em>, rather than as someone with a medical problem?</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>
Our respondents did not easily understand “primary care provider,” which is used to refer to both a person and a site. This usage in French of a single term proved too confusing in our cognitive tests, so the French version differentiated “usual source of care” (referring to a healthcare professional) and “clinic” (for site) depending on the context of the question to achieve appropriate subject-verb concordance. However, this subject-verb dissonance persisted in English, with some respondents wondering “how [their] doctor could be open or closed.” (e.g. Figure 2a and 2b) Respondents in various focus groups felt this instrument covered more ground or was more in-depth than others, but many found the font too small and the matrix formatting difficult to follow.

The measurement scale, eliciting probability of occurrence, is neither purely reporting nor purely rating. Some questions are posed hypothetically (e.g. Figure 2a), whereas others relate frequency of occurrence (e.g. Figure 2c). The probability response options for questions eliciting frequency were particularly problematic in French, where poor syntax concordance between the declarative tense in the question and the conditional response options created confusion (e.g. Figure 2c). The two extremes were easy to understand – definitely not means no or never, definitely means yes or always – but the intermediate options were neither clear nor equivalent between respondents.

Looking at my answers again, I realize that I answered those in a speculative way, which is not an accurate experience. So I think the word “probably” is more of a speculative nature, and perhaps that should be changed. [NS, rural]

For me, “probably”, “probably not” and “not sure” all mean the same thing. [QC, remote]

The IRT analysis (Figure 7b) indicates that although the “probably = 3” option has the highest probability of being endorsed in the mid range of accessibility, which corresponds well to the numeric value assigned to this option; it is still endorsed over the entire range of the latent variable. The insight of participants comments suggests “probably” may reflect expectations or hope as much or more than lived experience. For some items eliciting uncommon experiences (e.g. for Figure 2b), only “definitely” and “definitely not” are endorsed, indicating that a binary response scale (yes/no) would be most appropriate.

Most problematic for evaluators is the “not sure / don’t remember” response option. Although respondents appreciated this option rather than having to guess, it counts as a missing value in analysis. For unusual scenarios such as having a phone number to call when the office is closed (Figure 2b), 16% of respondents endorsed this response. The coding manual advises interpreting this negatively as “probably not,” based on the rationale that essential attributes of effective primary healthcare should be known to patients. However, the probability curve for this option, illustrated in the IRT graph in Figure 7b as NA, shows it is endorsed more in the positive range of accessibility, contrary to what is suggested by the developer. We found this pattern for other items in the PCAT as well.
3. Components of Primary Care Instrument (CPCI)

The 20-item CPCI (Flocke, 1997) (Figure 3) was developed by Susan Flocke and colleagues based on direct observations of care processes in 138 primary healthcare clinics. Although items are grouped approximately within constructs, no grouping is identifiable in the formatting. The same disagree/agree semantic difference response scale is used for all item statements.

**Figure 3**
Components of Primary Care Instrument: Sample items demonstrating a 6-point semantic differential response scale of the reporting type, with occasional reverse wording.

Mark the response that best describes your **regular doctor**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3a. I go to this doctor for almost all of my medical care.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3b. This doctor handles emergencies.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3c. This doctor can take care of almost any medical problem I might have.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3d. I could go to this doctor for help with a personal or emotional problem.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3e. This doctor does not know my medical history very well.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3f. This doctor knows a lot about my family medical history.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
<tr>
<td>3g. This doctor and I have been through a lot together.</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
</tr>
</tbody>
</table>
Respondents appreciated this instrument’s shortness and simple language, but overall, it was the least liked. They found the format cramped and had difficulty linking statements to the correct response line. Many were observed physically tracking the statements to the response scale. Reverse-worded items were seen as being “tricky” or as tests that made responding more difficult.

_I think these questions are put there to make sure that the person takes the time to read the questionnaire, not just marking anything._ [QC, rural]

Respondents did not like the semantic difference response scale as much as the Likert scales of other instruments and they did not like responding to the statement with an agree-disagree response. They tended to endorse extreme responses rather than use the full response scale. Combined with the crowded formatting, this produced significant halo effects. For instance, many respondents strongly agreed with both positively and negatively worded items in the same construct (e.g. items Figure 3e and 3f). IRT analysis, illustrated in Figure 7c for an item from the Accumulated Knowledge subscale (Figure 3f), show that respondents overwhelmingly endorse the “strongly disagree (1)” or “strongly agree (6) options, meaning that items function as a dichotomous response scale (yes/no), providing good discrimination between negative and positive experiences but with a loss of information. In addition, most respondents selected the most positive option, providing little capacity to discriminate between above-average assessments. Figure 7c shows that all response options other than 6 correspond to a below-average experience.

4. EUROPEP
The EUROPEP (Figure 4) was developed by Richard Grol, Michel Wensing and colleagues (Grol R, 1999) to compare the performance of general practice in different European countries and to incorporate patient perspectives in care improvement initiatives. The 23-item instrument uses a rating approach to assess Clinical Behaviour, which includes interpersonal communication and technical aspects of care, and Organization of Care, which principally addresses issues around accessing care. The instrument has been translated into 15 languages.
**Figure 4**
EUROPEP: Sample items demonstrating 5-point semantic differential response scale of the rating type that is the same for all items, with a “not applicable” option.

<table>
<thead>
<tr>
<th>How would you rate the following care provided by your general practitioner in the last 12 months?</th>
<th>1=Poor</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th></th>
<th>NA</th>
<th>(does not apply)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a. Making you feel you had time during consultations</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4b. Keeping your records and data confidential</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4c. Quick relief of your symptoms</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4d. Thoroughness</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4e. Help in dealing with emotional problems related to your symptoms</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4f. The helpfulness of staff (other than the doctor)</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4g. Getting an appointment to suit you</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4h. Getting through to the practice on the phone</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4i. Providing quick services for urgent problems</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td></td>
</tr>
</tbody>
</table>
Like the CPCI, the EUROPEP uses a semantic differential response scale, but its presentation is clearer. Respondents found it easy to read and answer but not as specific as the other questionnaires. Some item statements are very short (e.g. Figure 4d “Thoroughness”), with different interpretations producing lack of specificity.

*It was very quick and easy to fill out. But I’m not sure that it really describes my experience with my doctor.* [NS, rural]

Like the PCAT, the EUROPEP offered a “not applicable” response option. This option was endorsed by more than 10% of respondents for only three items: help with emotional problems (Figure 4e); preparation to see specialists; and quick services for urgent problems (Figure 4i). In the first two scenarios, the probability of endorsing “not applicable” is higher among those with a positive score in the rest of the construct; for quick service its response profile is highest in the negative experience zone. IRT analysis showed that response options were not always clearly differentiated, but they were ordinal and functioned as intended (Figure 7a shows item 4e). Like the CPCI, any responses other than the most positive correspond to below-average experience.

5. **Interpersonal Processes of Care (IPC)**

The IPC (Figure 5) was developed by Anita Stewart and colleagues to measure interpersonal aspects of quality of care as a counterpoint to technical components. The 45-item instrument was designed to be particularly sensitive to issues of equity and discrimination against patients with limited language proficiency or capacity to advocate strongly in their own care. It has been mostly administered by telephone, so typeface and formatting were discretionary. The questions, which use a frequency of response scale, refer to care received from all doctors, not only those in primary care. The items are regularly separated by phrases that frame the next set of questions. (We slightly changed all questions with the developers’ agreement in order to be compare items in this questionnaire with items from all five questionnaires used in the study and therefore be able to address care given by the regular doctor or by all doctors seen at the regular clinic). Usually, items are asked in the following form item 5a, figure 5: “How often did doctors speak too fast?”
Figure 5
Interpersonal Processes of Care: Sample items demonstrating the 5-point Likert response scale of the reporting type using a frequency scale is the same for all options. (Format used was discretionary)

The following questions ask about your experience with your regular doctor or doctors at your regular clinic over the past 12 months.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5a. How often did the doctor(s) speak too fast?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5b. How often did the doctor(s) let you say what you thought was important?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5c. How often did the doctor(s) seem bothered if you asked several questions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The next questions ask about how you and your regular doctor or doctors at your regular clinic decide about your health care. Over the past 12 months …

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5d. How often did you and your doctor(s) work out a treatment plan together?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The next four questions ask about the doctor’s front office staff, meaning the receptionist or the person you talk to on the phone to make appointment. In the past 12 months…

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5e. How often was office staff rude to you?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5f. How often did office staff have a negative attitude toward you?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Respondents found the content of the questions highly relevant to their healthcare experience. This instrument focuses on respectfulness, and one item in particular evoked comment in several groups: “How often did doctors treat you as an equal?”

I want [a professional doctor-patient relationship], I don’t want to be treated as an equal, I don’t want him to talk to me like I’m an equal because I don’t understand – I want him to talk to me in plain language but to talk to me with respect and caring and understanding. [NS, rural]

One strength of this instrument is that it includes a separate subscale for assessing office staff behaviour (e.g. Figure 5e and 5f), whereas others such as the EUROPEP (Figure 4f) embed the question among items about the doctor or clinic. Patients often have clearly differentiated experiences between staff and doctors.

The staff has a great impact on your entire experience. If the receptionist is not someone you feel you can approach, you feel blocked from your Doctor. [NS, rural]

The major difficulty with this instrument was the frequency response scale applied to all providers and all visits. The vast majority of respondents selected the two most positive categories of experience. IRT analyses show that anything less than the most positive category indicates a negative experience, as shown in Figure 7e (representing Figure 5c). The response scale is not fully exploited. It is especially poorly adapted to elements of care that are not expected to occur frequently or with every provider, as shown in Figure 5d. This suggests that a benchmark approach to scoring should be used, whereby any category other than “always” for positive events or “never” for negative events would be interpreted as a problem.

**Figure 7**

Item response graphs of typically performing items from each instrument showing option characteristic curves (solid Lines) and expected total scores (broken line) modelled as a function of total subscale scores (top axis) and standard normal quantiles (bottom axis).

Figure 7a: PCAS: How would you rate the usual wait for an appointment when you are sick and call the doctor’s office asking to be seen? (Item Figure 1a) Modelled on Organizational Accessibility. Probability of responding to each option (1 to 6) is clearly differentiated, ordinal, with approximately equivalent intervals.
Figure 7b PCAT: you’re your Primary Care Provider is open can you get advice quickly over the phone if you need it? (Figure 2a modelled on First-contact Access)
The option characteristic curve are differentiated and ordinal; 3=probably is endorsed most in the mid-range but also over the entire range of Access. The NA=don’t know option tends to be endorsed by respondents with above average access

Figure 7c: CPCI “This doctor knows a lot about my family medical history.” (Figure 3f, modelled on Accumulated Knowledge subscale)
The extreme options are predominantly endorsed - 1=strongly disagree; 6=strongly agree – functioning almost as a dichotomous response scale.

Figure 7d: EUROPEP How would you rate “Help in dealing with emotional problems related to your symptoms” (Figure 4e modelled on Explaining in Clinical Behaviour subscale)
The option characteristic curves are not all clearly differentiated, and ordinality is not always clear in the lower end of the scale.
DKN=not-applicable selected across all levels of Clinical Behaviour, and especially in the positive zone.
Missing pieces across instruments

6. Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCCSS)
The VANOCCSS (Borowsky et al., 2002) (Figure 6) is based on the Picker-Commonwealth approach to performance assessment (Gerteis, Edgman-Levitan, Daley, & Delbanco, 1993). Although it was designed as a postal questionnaire, we did not receive an original copy, so format and typeface were at our discretion. Most items pertain to the last visit and were therefore excluded from our study; we retained only the usual-care subscales of Overall Coordination and Specialist Access. The instrument elicits the frequency of critical incidents but applies a benchmark-type binary scoring to each item to detect problems. The occurrence of at least one problem is considered problematic, so it is very sensitive to negative experiences.

Figure 6

Veteran’s Affairs Outpatient Customer Satisfaction Survey (VANOCCS). Sample of items demonstrating a response scales that varies across items. * indicates response options that are considered problematic for dichotomous scoring (not indicated on the respondent version)

6.a. Were the providers who cared for you always familiar with your most recent medical history?
[ ] [1] No *
[ ] [2] Yes, sometimes *
[ ] [3] Yes, always

6.b. Were there times when one of your providers did not know about changes in your treatment that another provider recommended?
[ ] [1] No
[ ] [2] Yes *
[ ] [3] No changes in the past 12 months

6.c. Were there times when you were confused because different providers told you different things?
[ ] [1] No
[ ] [2] Yes *
These next questions are about visits in the past 12 months to specialists that your regular doctor knows about or has recommended you see.

IF YOU HAVE NOT SEEN A SPECIALIST IN THE PAST 12 MONTHS, PLEASE GO TO…

6.d. How often did you get to see specialists when you thought you needed to?

- [ ] [1] Never *
- [ ] [2] Sometimes *
- [ ] [3] Usually
- [ ] [4] Often
- [ ] [5] Didn’t need any specialists

6.e. How often did your specialists have the information they needed from your medical records?

- [ ] [1] Never *
- [ ] [2] Sometimes *
- [ ] [3] Usually
- [ ] [4] Often
- [ ] [5] Didn’t need any specialists
Because the instrument applies only to those who had seen providers other than the regular doctor, we placed it last in the questionnaire. Not all respondents answered it, and we obtained little qualitative information. However, it addresses whole-system issues that respondents identified as important.

**Response formats**
Among the various instruments and response scale formats, respondents consistently preferred Likert response scales, sans serif fonts in large sizes, squares as opposed to circles for responses, grouping questions under headings and varying response scales and formats throughout the questionnaire. They found matrix formatting (e.g. PCAT and some PCAS) difficult. Respondents appreciated being able to indicate that an item was not applicable or that they did not know, as in the PCAT and EUROPEP. Several expressed confusion around scoring information being indicated in response options (e.g. “Always 5”). For some, the clear link to scoring induced a more positive response.

**Bad questions**
When asked about questions they did not like, respondents consistently stated they did not like assessing what was in the provider’s mind or evaluating processes they did not directly observe or experience. Almost all questionnaires contained such questions. The items in the PCAT Trust subscale about the doctor pretending to know something (Figure 1d) and in the EUROPEP about the confidential treatment of the medical record (Figure 4b) are examples. Respondents attributed the benefit of the doubt but cautioned that their best guess did not necessarily translate to an accurate assessment.

*One respondent: There aren’t many people who know what their doctor [knows]…do you all know what your doctor thinks? Another respondent: Probably not. [QC, urban]*

Items about the depth of the physician–patient relationship provoked divergent and strong reactions. Some participants had few expectations that the physician would know anything beyond their medical history; others attested to the importance of whole-person knowledge in making care appropriate:

*They ask questions that don’t make any sense: “this doctor and I have been through a lot together,” “I could consult this doctor for a personal or emotional problem.” Come on. You’re not close to your doctor, to consult for a personal problem. You go… you go to see the doctor. [QC, rural]*

*I find the more your doctor knows you and your particulars, the better he’s going to be able to adjust things. [NS, urban]*

Most respondents said the instruments elicited general aspects of their healthcare experience that they considered important. Several groups felt 12 months was too short a reference because it only encompassed one or two visits, whereas their experience was informed by a longer period. Some missing pieces related to reporting on the system as a whole and others to the regular physician.

Respondents consistently expressed frustration at being limited to their regular doctor, as many of the most important critical incidents occur in specialty or hospital care or outside the clinical encounter. They wanted more questions on delays for access to tests and specialists and elicitation of technical mistakes. Many wanted to express their frustration or worry over finding a family physician.

Regarding reporting on the regular physician, respondents clearly welcomed an opportunity to give anonymous feedback and expected this information to go directly to their physician. They would have
liked to comment on how stressed or tired their physicians appear during clinical encounters. They also wanted to communicate to their physicians their desire for more input into decision-making around choices of specialist or treatment options. They also expressed frustration at not being able to report on technical aspects, especially mistakes.

They felt there were not enough questions addressing respectfulness: respecting the appointment time, responding to voice mail messages, recognizing personal worth and office staff respecting their privacy and confidentiality. They spoke of feelings of tension or confusion in the waiting room. Respondents wanted more questions about the physical set-up of the waiting room (privacy, toys for children), cleanliness of the bathroom or the clinic in general and wheelchair accessibility.

Discussion
Every instrument had features respondents appreciated: the readability of the PCAS; the PCAT’s “don’t know” response option (the most problematic feature from an evaluator’s perspective); the CPCI’s clear language; the EUROPEP’s conciseness; the IPC’s focus on respectfulness and office staff; and the VANOCCS’ whole-system perspective. The PCAS was considered the most readable and applicable overall, though its placement as first in the questionnaire may have affected this perception.

For self-administered instruments, formatting is as important to item performance as is wording. Our results reinforce suggested good practice such as large typeface, instructions placed close to responses, presenting questions together by thematic grouping and minimizing response format changes between sections (Dillman, 2007). Additionally, we recommend removing numeric scoring information from response options, using matrix formatting sparingly and using sans serif fonts.

All instruments included questions patients find difficult to evaluate, about things not directly observed or occurring in the provider’s mind. Patients can accurately report critical incidents like breaches in confidentiality or trust, but if they do not, it does not mean these have not happened. Offering a “don’t know” option gives patients more options but results in missing values for analysis. Like Schuman and colleagues (Schuman H. & Presser S., 1996), we found no consistent patterns to indicate how, “don’t know” options, should be interpreted. Though, they tended to be endorsed more in the positive zone of constructs of interest.

Respondents not only tolerate variation, but appreciate variety and response scales that correspond well with the line of questioning. The PCAS is exemplary in this regard; the response scale changes across dimensions while remaining consistent within a block of questions, facilitating the response task. Likewise, validity may be compromised by using a graded response scale for occurrences that are rare or binary experiences. The VANOCCS offers an interesting combination of graded reporting of events with binary scoring; an approach that may be relevant for instruments with highly skewed responses.

Finally, respondents affirmed the importance they accord to evaluating healthcare services and their expectation that this information be used to communicate suggestions for improvement anonymously and clearly to their providers. They desire to be “good respondents” and will respond to questions even when they are not sure of the answer or of what the question means. The information in this study can be used to ensure that evaluators and clinicians select instruments that not only demonstrate acceptable psychometric properties, but are also well accepted by patients.
What Patients Tell Us about Primary Healthcare Evaluation Instruments: Detailed Report

Reference List


