ELSEVIER

Patient Education and Counseling 67 (2007) 333–342

Patient Education and Counseling

www.elsevier.com/locate/pateducou

Measuring patient views of physician communication skills: Development and testing of the Communication Assessment Tool

Gregory Makoul [a,*], Edward Krupat [b], Chih-Hung Chang [a]

[a] *Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[b] *Harvard Medical School, Boston, MA, USA*

Received 14 March 2007; received in revised form 1 May 2007; accepted 3 May 2007

Abstract

Objective: Interpersonal and communication skills have been identified as a core competency that must be demonstrated by physicians. We developed and tested a tool that can be used by patients to assess the interpersonal and communication skills of physicians-in-training and physicians-in-practice.

Methods: We began by engaging in a systematic scale development process to obtain a psychometrically sound Communication Assessment Tool (CAT). This process yielded a 15-item instrument that is written at the fourth grade reading level and employs a five-point response scale, with 5 = excellent. Fourteen items focus on the physician and one targets the staff. Pilot testing established that the CAT differentiates between physicians who rated high or low on a separate satisfaction scale. We conducted a field test with physicians and patients from a variety of specialties and regions within the US to assess the feasibility of using the CAT in everyday practice.

Results: Thirty-eight physicians and 950 patients (25 patients per physician) participated in the field test. The average patient-reported *mean* score per physician was 4.68 across all CAT items (S.D. = 0.54, range 3.97–4.95). The average proportion of *excellent* scores was 76.3% (S.D. = 11.1, range 45.7–95.1%). Overall scale reliability was high (Cronbach's alpha = 0.96); alpha coefficients were uniformly high when reliability was examined per doctor.

Conclusion: The CAT is a reliable and valid instrument for measuring patient perceptions of physician performance in the area of interpersonal and communication skills. The field test demonstrated that the CAT can be successfully completed by both physicians and patients across clinical specialties. Reporting the proportion of ''excellent'' ratings given by patients is more useful than summarizing scores via means, which are highly skewed.

Practice implications: Specialty boards, residency programs, medical schools, and practice plans may find the CAT valuable for both collecting information and providing feedback about interpersonal and communication skills.

Keywords: Physician–patient communication; Communication skills; Quality; Assessment

1. Introduction

Over the past decade, accreditation, certification, and quality-improvement initiatives in many countries have highlighted the importance of examining the communication skills of physicians-in-training and physicians-in-practice. In the United States, the Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties (ABMS) jointly identified *interpersonal and communication skills* as one of the six general competencies for physicians [1,2]. As noted by Duffy and colleagues, ''while communication skills are the performance of specific tasks and behaviors by an individual, interpersonal skills are inherently relational and process oriented [3].'' Many assessment options are available, but there is no clear guidance on how interpersonal and communication skills should be measured [3].

Patient surveys are a viable option with a long history. However, the scale development process for these surveys is often unclear and items initially developed decades ago may not seem relevant to contemporary patients. Moreover, extant instruments sometimes mix communication items with satisfaction items, bundle multiple communication elements

* Corresponding author at: Center for Communication and Medicine, Northwestern University Feinberg School of Medicine, Division of General Internal Medicine, 676 North Street Clair, Suite 200, Chicago, IL 60611, USA.
E-mail address: makoul@northwestern.edu (G. Makoul).

0738-3991/$ – see front matter
doi:10.1016/j.pec.2007.05.005

into single items, ask patients to consider communication over a relatively long time-period (e.g., the past 12 months), or have modest internal reliability. With reliability, validity, and feasibility as central goals, we developed and tested a tool that can be used by patients to assess interpersonal and communication skills. Our objective was to create an instrument that captures patient views soon after inpatient or outpatient medical encounters, rather than impressions over a period of time. We reasoned that aggregating these ''snapshots'' of patient perceptions could provide specialty boards, residency programs, medical schools, and practice plans with the basis for providing focused feedback to physicians-in-practice and physicians-in-training.

2. Methods

We engaged in a systematic scale development process to obtain a psychometrically sound Communication Assessment Tool (CAT). We then conducted a field test to assess the feasibility of having physicians and patients use the CAT. Data were not connected with physician or patient names, and all aspects of the project were approved by the Northwestern University Institutional Review Board.

2.1. Initial item generation

The process of developing the CAT began with a review of models and instruments that became prominent in the context of teaching and assessing communication skills through sustained use at multiple institutions or adoption by major specialty boards. These include the SEGUE Framework [4] and the Four Habits Model [5,6], which the authors were involved in creating. Other tools reviewed at this stage were the American Board of Internal Medicine (ABIM) Patient Satisfaction Questionnaire [7], Calgary-Cambridge Guides [8,9], Consumer Assessment of Healthcare Providers and Systems (CAHPS) [10], Essential Elements of Communication in Medical Encounters [11], Patient-Centered Clinical Method [12], and the Royal College of General Practitioners Consulting Skills Module [13,14]. Although some of these are operationalized in the form of performance criteria or checklists rather than patient surveys, our review focused on identifying key communication tasks. The value of the task approach is derived from acknowledging the individuality of providers and recognizing that they may have different ways of accomplishing the same communication task [15]. Optimally, providers will tailor their communication skills and strategies to meet the needs of both the patient and the situation at hand, so the key is capture how well – rather than to specify how – a task was accomplished. The review yielded a list of 30 communication tasks, some with alternate wordings (e.g., ''Did not interrupt or cut me off when I was talking'', ''Let me talk without interruptions'').

2.2. Lay-person focus groups

The initial set of tasks was refined through a series of four focus groups with an average of eight patients each. These focus groups were conducted in the Chicago area, and included a relatively equal proportion of males and females, a diversity of ages and racial/ethnic backgrounds, and a broad range of literacy levels (i.e., one group of people with low literacy was recruited through an adult-learning program). The focus groups were videotaped to facilitate analysis of participants' response to existing items and ideas for new items. This process eliminated 15 items that were perceived as redundant or too narrow to be relevant across different specialties and visit types. For many of the 15 items that remained, wording was refined based on individual cognitive interviews and group discussion (e.g., ''Greeted me appropriately'' was changed to ''Greeted me in a way that made me feel comfortable''). Ideas for new items either modified the wording of existing items or focused on communication with staff.

Focus group participants were also asked to review several potential response scales: agreement scales with four, five, or six points; rating scales with five points (5 = excellent), six points (5 = excellent and 6 = could not be better), 10-points, and 100-points; and a dichotomous scale (i.e., no–yes). Focus group participants expressed a preference for the five-point agreement scale (strongly disagree to strongly agree), five-point rating scale (poor to excellent), and the dichotomous scale. These three options were pilot tested as described in Section 2.6.

2.3. National survey to determine item importance

To determine the importance Americans attached to these communication tasks, the 15 items were included on a national survey conducted by the University of Wisconsin Survey Center. This survey used a list-assisted frame and random digit dialing with a two-stage Mitofsky–Waksberg design, which gives all households a known chance of inclusion whether or not their phone number is listed [16,17]. After confirming that they had reached a household, telephone interviewers determined how many residents were at least 18 years old, and randomly selected a target respondent from all adult residents. Only the target respondent could be interviewed; no substitutions were allowed. The survey achieved a 41% response rate for a total of 1011 completed interviews. Interviewers were trained to ask about the importance of each communication task in a clear and consistent manner. Survey respondents gauged the importance of each task on a four-point scale ranging from ''not at all important'' to ''very important''. While all items were highly valued, we retained only the 12 items deemed ''very important'' by at least 70% of respondents.

2.4. Addition of items to ensure a comprehensive instrument

Three items were added because careful review of the list by the study team revealed significant gaps. The first item focused on giving information. During the focus groups, we had discussed three very specific tasks related to giving information: one on information about tests or procedures, another on information about diagnosis, and still another on information about treatments. None of these were included in

the national survey because focus group participants found them too narrow. Accordingly, we combined them into one item that is linked to individual patient expectations and, thus, ''very important'' by definition (gave me as much information *as I wanted*). The second item focused on perceptions regarding time, and was also worded to be ''very important'' by definition (Spent the *right amount* of time with me). We reasoned that getting feedback on this item would be useful for physicians. The third additional item targeted whether the *staff* treated the patient with respect. This was a consideration voiced during the focus groups as well as in discussions with experts at national and international meetings. While not a direct measure of a physician's interpersonal and communication skills and, thus, not included in the survey, we expected that the opportunity to collect feedback about staff would be valuable for physicians. In sum, the CAT that emerged for further testing was comprised of 15 tasks, 14 of which focus specifically on physician–patient communication. The 15-item version is designed for use with practicing physicians, it would be appropriate to drop the staff item if the CAT is used to gauge the interpersonal and communication skills of medical students or residents.

2.5. Lexile analysis for readability

We sought to keep the reading skills needed to comprehend the CAT items at or below an eighth grade level, a decision consistent with the Institute of Medicine's observation that individuals who have difficulty reading above this level may face problems understanding and acting upon healthcare information [18]. Accordingly, each of the 15 items was subjected to a Lexile analysis for readability [19,20]. Lexiles are based on sentence length and word frequency in popular literature; a Lexile value of 1000 is a level at which people can read eighth grade texts with more than 80% comprehension. The individual CAT items have Lexile values ranging from 260 to 760. Taken together, the 15 CAT items have a Lexile value of 510. This corresponds to a fourth grade reading level, which increases the likelihood that the scale can be appropriately understood and used whether self-administered or interviewer-administered.

2.6. Selection of response scale

We conducted a pilot test to determine which of three response scales proved the most psychometrically sound: the five-point Likert scale (strongly disagree–slightly disagree–neither agree nor disagree–slightly agree–strongly agree), five-point rating scale (poor–fair–good–very good–excellent), or the dichotomous scale (no–yes). Data were collected at clinical practices affiliated with Northwestern University Feinberg School of Medicine in Chicago, where paper versions of the CAT were completed by 30 patients for each of 17 physicians (9 general internists, 4 pediatricians, and 4 orthopaedic surgeons) immediately after their visits. Within each group of 30 patients, 10 used the *disagree–agree* scale, 10 used the *poor–excellent* scale, and 10 used the dichotomous *no–yes* scale.

We applied Andrich's rating scale model (RSM) [21,22] to psychometrically analyze the structures of response categories using the WINSTEPS software program [23]. RSM is measurement model based on Item-Response Theory. It specifies that all items in a test or scale measure the same underlying trait (i.e., the test or scale is unidimensional). This model was selected because it also allows examination of the category structure of the rating scales. The RSM specifies two facets (person latent trait, B_n; item location, D_i), and the step threshold (F_i). In this study, B_n refers to the latent trait measure (i.e., communication perception) of person n. The item location (D_i) measures the degree to which item i is likely to be endorsed in a manner reflecting a high score, with higher values indicating that an item is harder to endorse. The step threshold (F_i) is the point on the latent trait scale at which two consecutive category response curves intersect. For instance, for the *poor–excellent* scale, F_1 is the transition from intensity category 1 (poor) to category 2 (fair), F_2 is the transition from category 2 (fair) to category 3 (good), F_3 is the transition from category 3 (good) to category 4 (very good), and F_4 is the transition from category 4 (very good) to category 5 (excellent). We expected values of F_1, F_2, F_3, and F_4 to be distinct and in ascending order.

The RSM analyses indicated that the five-point *poor–excellent* scale was optimal for collecting data. In contrast to the five-point *disagree–agree* scale, the intersection between adjacent response categories in the *poor–excellent* scale were ordered from less to more, and equally spaced in terms of distance between the two threshold parameters, indicating equal intervals between any two adjacent response categories (see Fig. 1). Moreover, in contrast to the dichotomous *no–yes*

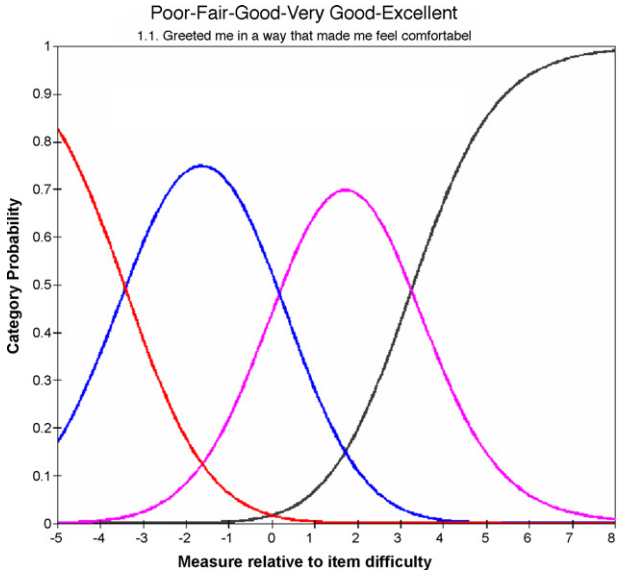

Fig. 1. Response category characteristic curve. Red represents F_1, the transition from intensity category 1 (poor) to category 2 (fair); blue represents F_2, the transition from intensity category 2 (fair) to category 3 (good); pink represents F_3, the transition from category 3 (good) to category 4 (very good); black represents F_4, the transition from category 4 (very good) to category 5 (excellent). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

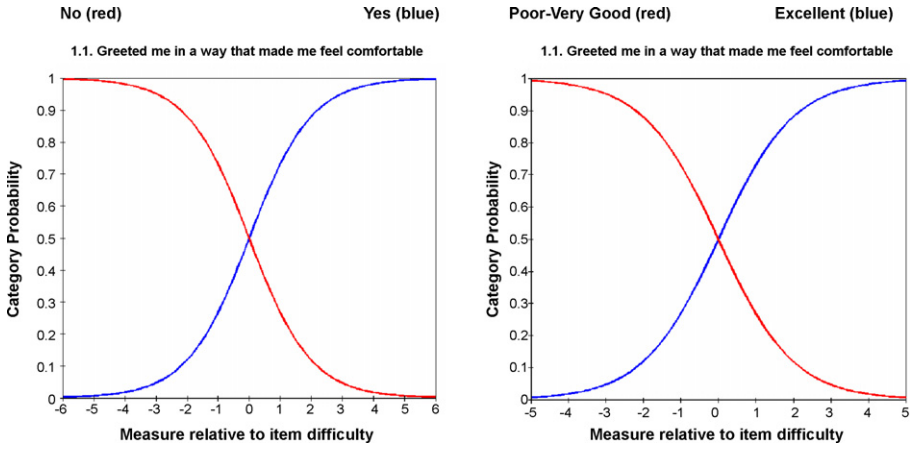

Fig. 2. Response category curves: ''excellent'' maps onto ''yes''. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

scale, the rating scale provided more information. Fig. 2 illustrates an important finding: when the five-point rating scale was dichotomized by comparing the four lowest categories (poor–fair–good–very good) to the highest category (excellent), the scale performed almost exactly like the dichotomous no–yes scale. This was not the case when the five-point rating scale was divided in other ways, suggesting that a rating of ''excellent'' is akin to ''yes'', while even ''very good'' is closer to ''no'' than ''yes''.

2.7. Scale reliability and validity testing

After establishing that the *poor–excellent* response scale should accompany the CAT items, we conducted a second pilot test to examine the psychometric characteristics of the items. Data were collected using paper versions of the CAT at practices within the Colorado Permanente Medical Group (CPMG) in Denver. The CAT was completed immediately by a total of 600 patients immediately after their visits: 30 for each of 20 physicians in a variety of specialties (Dermatology, General Surgery, Head and Neck Surgery, Obstetrics and Gynecology, Ophthalmology, Orthopaedic Surgery).

Exploratory factor analysis with principal components extraction and Varimax rotation reveals one factor that accounts for 78.8% of the variance, although the staff-oriented item has the lowest communality. Results of the pilot test indicate that the 15-item CAT is internally consistent, with a high scale reliability (Cronbach's coefficient alpha = 0.98). We also conducted differential item functioning (DIF) analyses to determine if the CAT yields unbiased data for participants with different sociodemographic and clinical characteristics [24]. Results of DIF analyses clearly showed that these items perform similarly across physician specialty as well as across patient sex, race/ethnicity, education level, self-reported health status, and previous visits to the physician ($r \geq 0.95$, $p < 0.001$ for each DIF analysis).

We used an existing patient satisfaction data routinely collected by the Colorado Permanente Medical Group to test the validity of the CAT. The validity test compared CAT ratings for three physicians with the lowest patient satisfaction scores on CPMG's own measure (78%, 78%, 79%) to ratings for three physicians with the highest scores (98%, 99%, 99%). The CAT ratings were markedly different between these two groups of physicians, with an average CAT rating of 4.28 (S.D. = 0.67) for the low patient-satisfaction physicians and an average of 4.92 (S.D. = 0.23) for the high patient-satisfaction physicians (df = 173, $p < 0.001$). This analysis reinforces the validity of the CAT, as well as the point made above that even a mean rating of 4 (i.e., ''very good'') indicates considerable room for improvement.

2.8. Field test to assess feasibility

The systematic scale development process and pilot tests generated a streamlined and psychometrically sound version of the CAT, with 15 items (14 physician-oriented, 1 staff-oriented) coupled with a five-point response scale ranging from ''poor'' to ''excellent'' (see Appendix A). We field tested this scale to determine the extent to which problems arise when delivering the CAT in a less controlled environment. We evaluated key logistical aspects of implementation in the field: (1) whether physicians who volunteered to participate in the field-test would use the CAT for self-assessment; (2) whether office staff would ask patients to complete the CAT in everyday clinical practice; (3) whether patients and/or caregivers would use the CAT survey; (4) whether respondents encountered problems when completing the CAT, either through an automated telephone response system or via the Internet. We also analyzed how scores differed by item and by doctor; and examined different ways of analyzing and presenting scores.

The American Board of Medical Specialties (ABMS) asked member boards to volunteer for field-testing the CAT, resulting in a convenience sample of 40 physicians representing six boards: Dermatology, Family Medicine, Neurosurgery, Ophthalmology, Orthopaedic Surgery, and Physical Medicine & Rehabilitation. As one goal of the field test was to implement the CAT in venues without research assistants, the ABMS made systems available that allowed patients to choose between

completing the CAT via either the Internet or interactive voice response (i.e., telephone).

3. Results

A total of 38 of the 40 physicians (95%) were successful in completing the CAT once as a self-assessment as well as having their office staff recruit 25 patients to complete the CAT within 1 day of their visit. Accordingly, a total of 950 patients (i.e., 38×25) were involved in the field test (see Table 1). The patient sample included a broad range of ages, from children through patients age 75 and over. Patients sometimes had help with the CAT: 9.8% were completed by family caregivers. As one of the participating specialties was family practice, 1.2% of the patients were age 14 and under; the CAT was completed by a parent in these cases. The modal age category was 45–54, accounting for nearly a quarter (23.2%) of the sample. More than half (59.4%) of the patients were female. While all racial/ethnic groups were represented, most patients (85.7%) were Caucasian. The majority of patients (69.7%) had seen their physician more than once before the study visit, and most reported their health status as ''good'' (37.1%) or ''very good'' (31.2%). Other than specialty, we do not have descriptive information about the physicians.

3.1. Telephone versus Internet administration of the CAT

More than half (55.8%) of patients or caregivers used the telephone version, while 44.2% went online to complete the CAT. In a multivariate logistic regression, age, sex, and education level predicted use of the telephone system, controlling for race/ethnicity, previous experience with the study physician, and self-reported health. Specifically, women were more likely than men (OR = 1.50, $p < 0.005$) and older participants were more likely than younger (OR 1.32, $p < 0.001$) to use the telephone version. In contrast, people with higher levels of education were less likely to use the telephone system (OR = 0.86, $p < 0.001$). In other words, people with more education were more likely to use the Internet version.

Overall, both the telephone and Internet versions appear to have been easy to use and navigate, with at least 95% of participants reporting that instructions were clear and that they had no problems entering their user ID and password, responding to survey items, changing answers, and completing the survey. One method-related difference did arise: 5.5% of participants who responded via telephone found it difficult to enter their ID and password versus 1.7% of participants who used the web version ($\chi^2 = 7.70$, df = 1, $p < 0.01$). In addition, a trend emerged when data were analyzed by survey method (i.e., phone versus web), although differences were not statistically significant: Whether reported as means or percent-excellent, scores collected through the telephone survey mechanism were slightly, but consistently, higher. In other words, more patients used the ''5'' (i.e., excellent) option when responding to items via the phone, a pattern that held across differences in patient sex, race/ethnicity, education, and health status.

Table 1
Number of doctors and patients by specialty

Specialty	# doctors	# patients
Dermatology	8	200
Family Medicine	11	275
Neurosurgery	3	75
Ophthalmology	5	125
Orthopaedic Surgery	3	75
Physical Medicine & Rehabilitation	8	200
Total	38	950

3.2. Scale properties in the field

Confirmatory factor analysis indicated that the first 14 CAT items, all of which focus on communication with the doctor, are properly considered one factor. In other words, it is appropriate to create a mean score for the 14 doctor-oriented items and consider the staff-oriented item separately because its communality is conspicuously low (0.28). However, it is important to include this staff-oriented item because it provides relevant information for physicians. Moreover, patient reaction to the staff appears to be associated with patient-reports regarding physician behavior ($r = 0.49$, $n = 950$, $p < 0.001$). In terms of overall scale reliability, Cronbach's coefficient alpha was very high (0.96) for the 14 doctor-oriented items. When scale reliability was examined per doctor, alpha coefficients were also uniformly high, ranging from 0.80 to 0.99 (mean = 0.95, S.D. = 0.03).

Table 2 displays CAT scores from the field test. Means for physician self-assessments ranged from 3.76 (S.D. = 0.88) for ''spent the right amount of time with [the patient]'' to 4.61 (S.D. = 0.50) for ''treated [the patient] with respect''. Physician self-assessments were consistently lower than patient-reported means, which ranged from 4.44 (S.D. = 0.31) for ''encouraged me to ask questions'' to 4.81 (S.D. = 0.13) for ''treated me with respect'' across all physicians. This pattern of patient-reported scores was consistent with both the pilot test and previous research on actual communication in medical encounters [4].

It is clear that mean scores are clustered toward the upper end of the scale. Given the finding from rating scale analysis that a score of 5 (i.e., ''excellent'') maps onto ''yes'', we also calculated the percentage of ''excellent'' ratings per item. This approach revealed a broader spectrum of scores, which were normally distributed. More specifically, across all physicians, average percent-excellent scores ranged from 62.7% (S.D. = 16.19) for ''encouraged me to ask questions'' to 84.4% (S.D. = 9.10) for ''treated me with respect''. While these are the same items highlighted by examining minimum and maximum mean scores, focusing on percent-excellent obviates the ceiling effects associated with use of patient-reported means.

At the individual item and summary-score level, the CAT detects significant differences between physicians, in terms of both mean scores and percent-excellent scores reported by patients ($p < 0.001$ for all comparisons). Fig. 3 illustrates patient-reported means for the first 14 CAT items. Across all 38

Table 2
CAT Scores from Field Test

ITEM	Doctor self-assess (n = 38)	Patient mean (n = 950)	Patient % excellent (n = 950)
1. Greeted me in a way that made me feel comfortable	4.37	4.74	80.2%
2. Treated me with respect	4.61	4.81	84.4%
3. Showed interest in my ideas about my health	4.26	4.67	74.4%
4. Understood my main health concerns	4.42	4.73	79.7%
5. Paid attention to me (looked at me, listened)	4.37	4.75	81.3%
6. Let me talk without interruptions	3.84	4.71	78.2%
7. Gave me as much information as I wanted	4.32	4.63	72.8%
8. Talked in terms I could understand	4.29	4.75	80.4%
9. Checked to be sure I understood everything	4.11	4.58	70.2%
10. Encouraged me to ask questions	4.16	4.44	62.7%
11. Involved me in decisions as much as I wanted	4.26	4.59	70.7%
12. Discussed next steps	4.55	4.70	77.7%
13. Showed care and concern	4.47	4.74	80.1%
14. Spent the right amount of time with me	3.76	4.63	74.9%
15. Staff treated me with respect	4.37	4.76	80.8%

Scale: 1 = poor; 2 = fair; 3 = good; 4 = very good; 5 = excellent.

study physicians, the patient-reported mean score was 4.68 (S.D. = 0.54). The lowest mean score for an individual physician was 3.97 (S.D. = 1.04) and the highest was 4.95 (S.D. = 0.16). The physician with the lowest score (3.97) was more than one standard deviation below the mean. As shown in Fig. 4, the mean percent-excellent score for the first 14 items was 76.3% (S.D. = 11.1) across all 38 physicians, with a range of 45.7–95.1%. Focusing on percent-excellent elucidates the variation between physicians.

4. Discussion and conclusion

4.1. Discussion

The 15-item Communication Assessment Tool (CAT) is a reliable and valid instrument for measuring patient perceptions of physician performance in the area of interpersonal and communication skills. Conceptually, it focuses on the achievement of communication tasks rather than prescribing

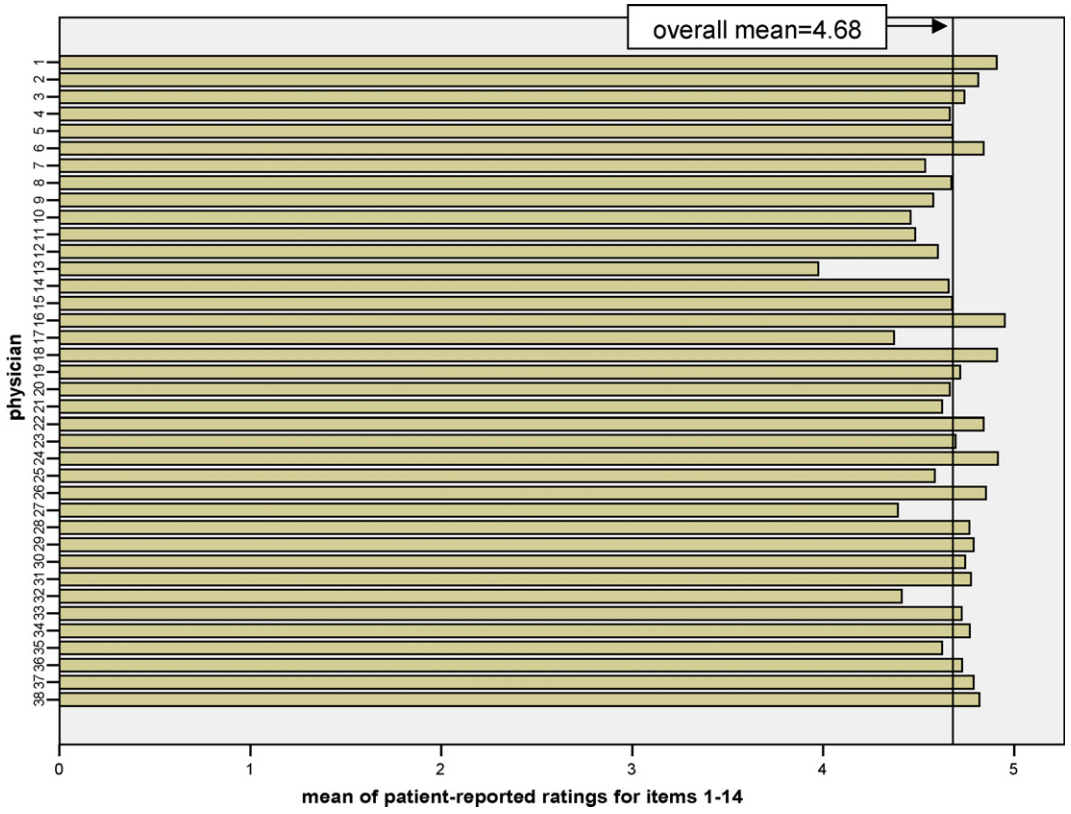

Fig. 3. Overall mean scores per physician.

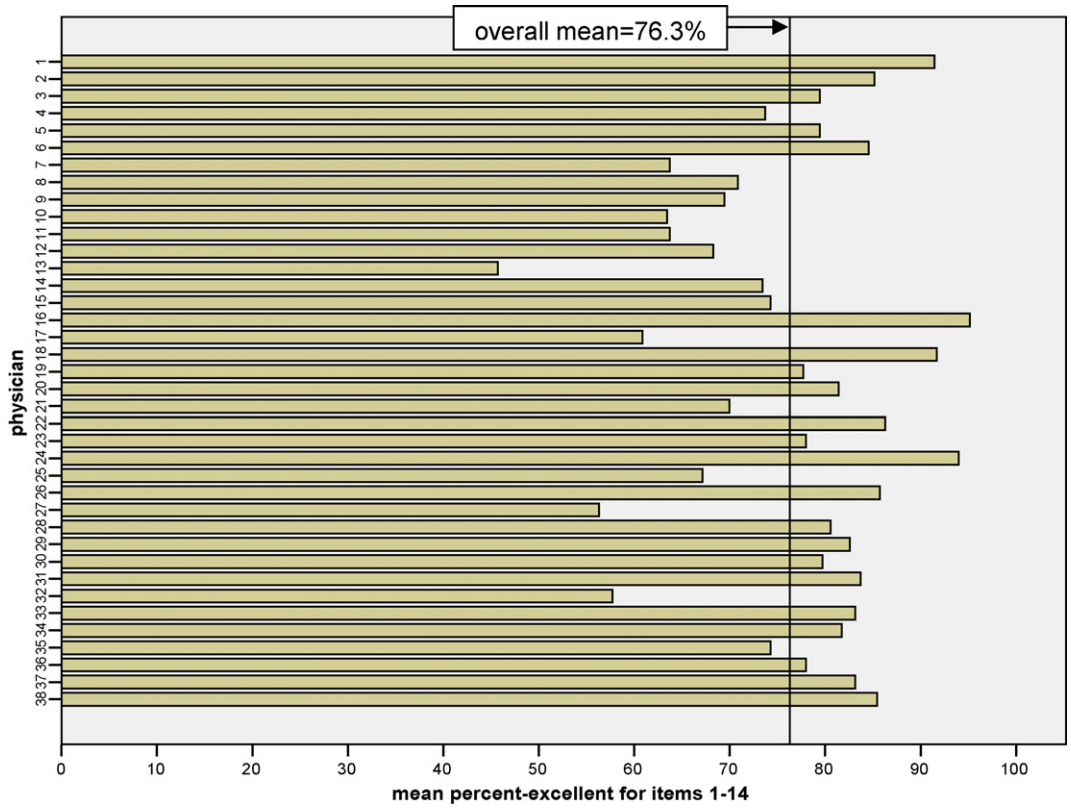

Fig. 4. Overall percent-excellent scores per physician.

particular ways of accomplishing them. At a very practical level, it is a simple and straightforward tool with discrete items that are accessible to patients across literacy levels. Moreover, the CAT can be completed via paper, an automated telephone system, or the Internet. A small percentage of patients reported having difficulty with entering their ID and password into the telephone system; it will be important to determine whether improvements can be made before implementing the telephone version on a larger scale. In addition, the possibility that a telephone version might result in slightly higher scores than an Internet version requires further study.

There are a number of other scales designed to measure patients' experiences of medical care. However, we believe that our systematic approach to scale development and testing yielded a combination of positive characteristics that exist in no other single instrument. More specifically, the final version of the CAT benefited from a careful review of prominent models to generate a list of communication tasks, focus groups to gather patient perspectives on items and response scales, a national survey to determine the importance attached to each item, expert review to ensure a comprehensive list of items, Lexile analysis to assess readability, and psychometric analyses to determine the most viable response scale. The plan and procedure of item generation ensured content and construct validity; scores also exhibited expected relationships with patient satisfaction data, establishing predictive validity [25].

The field test demonstrated that the 15-item CAT can be successfully completed by both physicians and patients. The scale reliability was high, both across doctors and for each doctor in the field test. Based on this study as well as previous research and practice, we recommend collecting 20–30 completed forms per physician [4,26–30]. This sample size is congruent with Rasch-based generalizability theory, which estimates that 12–30 ratings per examinee are required when seeking a reliability of 0.96 for data collected on a five-point scale [31]. The high scale reliability suggests that the CAT could be streamlined by dropping some items. However, we advocate keeping the full set, as it provides specific information for physicians without placing undue burden on patients (i.e., the CAT takes only 1–2 min to complete).

As evidenced by the national survey, the communication tasks that comprise the CAT are tangible to and valued by American adults. Views toward some of the communication tasks might differ in other countries; this is an empirical question that can certainly be addressed in future studies. Generic, specialty-specific, and country-specific norms can be established in standard-setting studies as well as large-scale comparative studies. These norms and standards may differ with stage-of-training (e.g., medical students versus practicing physicians).

4.2. Conclusion

It is important to differentiate between collecting the data and reporting the data. In terms of administering the survey, the five-point *poor–excellent* scale emerged as best in the pilot tests. However, we found that essentially dichotomizing scores by reporting the proportion of ''excellent'' ratings given by patients is more useful than summarizing scores via patient-reported means, which are highly skewed (see Figs. 3 and 4). Indeed, a mean score in the ''very good'' range, whether on an

individual item or as a summary score, might not motivate physicians to address the need for improvement to the same extent as learning that only 55% of patients thought they did an excellent job in terms of communication skills. That said, the results must be put in context and resources (e.g., workshops, online modules) should be made available to facilitate quality improvement efforts. The following language may be a useful model for CAT reports to individual physicians, and could be tailored for use with trainees:

> Communication with patients is a very important part of quality medical care. Accordingly, interpersonal and communication skills are considered a core area of competency. As you know, a sample of your patients were asked to complete the Communication Assessment Tool (CAT), a reliable and valid instrument that patients can use to provide feedback on your interpersonal and communication skills, based on their most recent medical encounter with you.
>
> The CAT consists of 15 items and uses a five-point scale: 1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent. Your *overall score* presents an average of the first 14 items and offers a general sense of how patients view your interpersonal and communication skills; the overall score for all physicians in your group is provided for your information. In addition, your report provides individualized feedback by displaying *item scores* that represent the proportion of patients who assigned a score of ''excellent'' for each item.
>
> This type of specific, systematic feedback from patients is both rare and valuable. While the report is properly considered a ''snapshot'' of patient perceptions, it offers a solid opportunity for reflection on your interpersonal and communication skills with the goal of reinforcing strengths and identifying areas that merit more attention for improvement.

While having patients complete the CAT appears to be a promising mechanism for periodically assessing interpersonal and communication skills, this approach should be considered part of a toolbox that includes self-assessment, observation of communication during real and/or simulated medical encounters, examinations or interactive computer modules that capture knowledge and attitudes regarding communication, and additional surveys that focus on other aspects of communication (e.g., teamwork) [3,32,33]. In short, high-quality assessment requires more than one high-quality measurement tool.

4.3. Practice implications

We believe that the CAT can be productively used across the continuum of medical education in both inpatient and outpatient contexts. While the version described in this article was developed to be used by patients of practicing physicians, educators and researchers at Northwestern have also constructed versions tailored for medical students and residents. The first 14 items are identical; the versions for physicians-in-training have different introductions and eliminate the item regarding the physician's staff. Specialty boards, residency programs, medical schools, and practice plans may find the CAT valuable for both collecting information and providing coherent feedback about communication in everyday clinical practice.

Acknowledgements

This study was supported by the American Board of Medical Specialties Research and Education Foundation, Evanston, Illinois, USA. Results of this research were presented at the International Conference on Communication in Healthcare (Basel, Switzerland, 2006). We thank all of the physicians, patients, and staff who participated in the development and testing process.

Appendix A

A.1. Communication Assessment Tool

Communication with patients is a very important part of quality medical care. We would like to know how you feel about the way your doctor communicated with you. Your answers are completely confidential, so please be as open and honest as you can. Thank you very much.

1	2	3	4	5
poor	fair	good	very good	excellent

Please use this scale to rate the way the doctor communicated with you. Circle your answer for each item below.

The doctor	poor				excellent
1. Greeted me in a way that made me feel comfortable	1	2	3	4	5
2. Treated me with respect	1	2	3	4	5
3. Showed interest in my ideas about my health	1	2	3	4	5
4. Understood my main health concerns	1	2	3	4	5
5. Paid attention to me (looked at me, listened carefully)	1	2	3	4	5
6. Let me talk without interruptions	1	2	3	4	5
7. Gave me as much information as I wanted	1	2	3	4	5
8. Talked in terms I could understand	1	2	3	4	5
9. Checked to be sure I understood everything	1	2	3	4	5
10. Encouraged me to ask questions	1	2	3	4	5
11. Involved me in decisions as much as I wanted	1	2	3	4	5
12. Discussed next steps, including any follow-up plans	1	2	3	4	5
13. Showed care and concern	1	2	3	4	5
14. Spent the right amount of time with me	1	2	3	4	5
The doctor's staff	poor				excellent
15. Treated me with respect	1	2	3	4	5

References

[1] Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Affair (Millwood) 2002;21:103–11.

[2] Horowitz SD. Evaluation of clinical competencies: basic certification, subspecialty certification, and recertification. Am J Phys Med Rehab 2000;79:478–80.

[3] Duffy FD, Gordon GH, Whelan G, Cole-Kelly K, Frankel R, all participants in the American Academy on Physician and Patient's Conference on Education and Evaluation of Competence in Communication and Interpersonal Skills. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. Acad Med 2004;79:495–507.

[4] Makoul G. The SEGUE Framework for teaching and assessing communication skills. Patient Educ Couns 2001;45:23–34.

[5] Frankel RM, Stein T. Getting the most out of the clinical encounter: the four habits model. J Med Pract Manage 2001;16:184–91.

[6] Krupat E, Frankel R, Stein T, Irish J. The Four Habits Coding Scheme: validation of an instrument to assess clinicians' communication behavior. Patient Educ Couns 2006;62:38–45.

[7] PSQ Project Co-Investigators. Final report on the patient satisfaction questionnaire. Philadelphia: American Board of Internal Medicine; 1989.

[8] Kurtz SM, Silverman J, Draper J. Teaching and learning communication skills in medicine, 2nd ed., Oxford: Radcliffe Publishing; 2005.

[9] Silverman J, Kurtz SM, Draper J. Skills for communicating with patients, 2nd ed., Oxford: Radcliffe Publishing; 2005.

[10] Agency for Healthcare Research and Quality. CAHPS Overview. Available from: https://www.cahps.ahrq.gov [accessed December 20, 2006].

[11] Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Acad Med 2001;76:390–3.

[12] Stewart M, Brown JB, Weston WW, McWhinney IR, McWilliam CL, Freeman TR. Patient-centered medicine: transforming the clinical method. Thousand Oaks: Sage Publications; 1995.

[13] Campion P, Foulkes J, Neighbour R, Tate P. Patient centredness in the MRCGP video examination: analysis of large cohort. Membership of the Royal College of General Practitioners. Brit Med J 2002;325:691–2.

[14] Tate P, Foulkes J, Neighbour R, Campion P, Field S. Assessing physicians' interpersonal skills via videotaped encounters: a new approach for the Royal College of General Practitioners Membership Examination. J Health Commun 1999;4:143–52.

[15] Makoul G, Schofield T. Communication teaching and assessment in medical education: an international consensus statement. Patient Educ Couns 1999;37:191–5.

[16] Potthoff RF. Telephone sampling in epidemiologic research: to reap the benefits, avoid the pitfalls. Am J Epidemiol 1994;139:967–78.

[17] Waksberg J. Sampling methods for random digit dialing. J Am Stat Assoc 1978;73:40–6.

[18] Institute of Medicine. Health literacy: a prescription to end confusion. Washington, DC: National Academies Press; 2004.

[19] The Lexile Framework for Reading: Matching Readers to Text. Available from: http://www.lexile.com [accessed December 20, 2006].

[20] Stenner AJ, Horabin I, Smith DR, Smith M. The Lexile Framework. Durham, NC: MetaMetrics; 1988.

[21] Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. J Appl Meas 2002;3:325–59.

[22] Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. J Appl Meas 2003;4:205–21.

[23] Linacre JM, Wright BD. WINSTEPS: Rasch analysis for all two-facet models. Chicago: MESA Press; 2001.

[24] Holland PW, Wainer H, editors. Differential item functioning. Mahwah, NJ: Lawrence Erlbaum Associates; 1993.

[25] Nunnally JC. Psychometric theory. New York: McGraw Hill; 1978.

[26] American Board of Internal Medicine. Surveys of Peers, Patients, and Practice Systems. Available from: http://www.abim.org/moc/sempbpi.shtm#4 [accessed April 20, 2007].

[27] Campbell C, Lockyer J, Laidlaw T, Macleod H. Assessment of a matched-pair instrument to examine doctor–patient communication skills in practising doctors. Med Educ 2007;41:123–9.

[28] Makoul G, Arntson P, Schofield T. Health promotion in primary care: physician–patient communication and decision making about prescription medications. Soc Sci Med 1995;41:1241–54.

[29] Makoul G, Dhurandhar A, Goel MS, Scholtens D, Rubin AS. Communication about behavioral health risks: a study of videotaped encounters in 2 internal medicine practices. J Gen Intern Med 2006;21:698–703.

[30] Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003;15:270–92.

[31] Linacre JM. Rasch-based generalizability theory. Rasch Meas Trans 1993;7:283–4.

[32] Holmboe E, Kim N, Cohen S, Curry M, Elwell A, Petrillo MK, Meehan TP. Primary care physicians, office-based practice, and the meaning of quality improvement. Am J Med 2005;118:917–22.

[33] Holmboe ES, Rodak W, Mills G, McFarlane MJ, Schultz HJ. Outcomes-based evaluation in resident education: creating systems and structured portfolios. Am J Med 2006;119:708–14.