

Rating communication in GP consultations: the association between ratings made by patients and trained clinical raters

AUTHOR(S)

Jenni Burt, Gary Abel, Natasha Elmore, Jenny Newbould, Antoinette Davey, Nadia Llanwarne, Inocencio Maramba, Charlotte Paddison, John Benson, Jonathan Silverman, Marc N Elliott, John Campbell, Martin Roland

PUBLICATION DATE

01-04-2018

HANDLE

10536/DRO/DU:30119839

Downloaded from Deakin University's Figshare repository

Deakin University CRICOS Provider Code: 00113B

Medical Care Research and Review
2018, Vol. 75(2) 201–218
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1077558716671217
journals.sagepub.com/home/mcr



Rating Communication in GP Consultations: The Association Between Ratings Made by Patients and Trained Clinical Raters

Jenni Burt[1], Gary Abel[2], Natasha Elmore[1], Jenny Newbould[1], Antoinette Davey[2], Nadia Llanwarne[1], Inocencio Maramba[2], Charlotte Paddison[3], John Benson[1], Jonathan Silverman[1], Marc N. Elliott[4], John Campbell[2], and Martin Roland[1]

Abstract

Patient evaluations of physician communication are widely used, but we know little about how these relate to professionally agreed norms of communication quality. We report an investigation into the association between patient assessments of communication quality and an observer-rated measure of communication competence. Consent was obtained to video record consultations with Family Practitioners in England, following which patients rated the physician's communication skills. A sample of consultation videos was subsequently evaluated by trained clinical raters using an instrument derived from the Calgary-Cambridge guide to the medical interview. Consultations scored highly for communication by clinical raters were also scored highly by patients. However, when clinical raters judged communication to be of lower quality, patient scores ranged from "poor" to "very good." Some patients may be inhibited from rating poor communication negatively. Patient evaluations can be

This article, submitted to *Medical Care Research and Review* on March 3, 2016, was revised and accepted for publication on August 29, 2016.

[1]University of Cambridge School of Clinical Medicine, Cambridge, UK
[2]University of Exeter Medical School, Exeter, Devon, UK
[3]Anglia Ruskin University, Cambridge, UK
[4]RAND Corporation, Santa Monica, CA, USA

Corresponding Author:
Jenni Burt, Primary Care Unit, Institute of Public Health, Forvie Site, University of Cambridge School of Clinical Medicine, Box 113 Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.
Email: jab35@medschl.cam.ac.uk

useful for measuring relative performance of physicians' communication skills, but absolute scores should be interpreted with caution.

Keywords
patient experience, physician–patient communication, health care surveys, health care quality measurement, health care quality

Introduction

Good physician–patient communication is central to good patient experience, and a major driver of overall patient assessments of primary care in the United States and the United Kingdom (Paddison et al., 2013; Quigley et al., 2014). Communication skills are consequently core strands of medical training, postgraduate assessment, and ongoing professional development (Accreditation Council for Graduate Medical Education, 2015; General Medical Council, 2015). While communication is important in its own right, it may be associated with other dimensions of quality of care such as clinical effectiveness and patient safety, and evidence suggests that good communication skills tend to be found alongside good clinical skills (Doyle, Lennox, & Bell, 2013; Llanwarne et al., 2013; Price, Elliott, Cleary, Zaslavsky, & Hays, 2015). The quality of physician–patient communication has been associated with patient adherence to treatment (Zolnierek & Dimatteo, 2009), uptake of cancer screening (Carcaise-Edinboro & Bradley, 2008), improved blood pressure control in hypertensive patients (Orth, Stiles, Scherwitz, Hennrikus, & Vallbona, 1987), and reductions in the risk of serious medical error (Kuzel et al., 2004). However, concerns about the quality of physician–patient communication remain, and a significant proportion of malpractice claims are driven by poor communication (Tamblyn et al., 2007; Vincent et al., 2006).

The quality of physician–patient communication is assessed using approaches including observer rating and patient surveys (Duffy et al., 2004). Observer rating of consultation skills relies on either the use of simulated patients, such as in Objective Structured Clinical Exams (Turner & Dankoski, 2008), or in the observation or videotaping of actual consultations (encounters; Zill et al., 2014). Due to its complexity, observer rating is usually confined to the assessment of medical students and postgraduate examinations, or used for research purposes. Patient surveys, by contrast, are widely used to assess the standard of physician–patient communication, and national survey programs include the English GP Patient Survey and the U.S. CAHPS (Consumer Assessment of Healthcare Providers and Systems; Agency for Healthcare Research and Quality, 2015; Ipsos MORI, 2015). Findings from such surveys inform a variety of official metrics of care quality. In England, the Care Quality Commission (the regulator of health and social care) uses data from the GP Patient Survey as part of its monitoring of key performance indicators in its practice inspection regime (Care Quality Commission, 2015). Additionally, GP Patient Survey data are made publicly available both via a dedicated website (https://gp-patient.co.uk) and in the form of performance

scores attached to practice profiles on a public listing of NHS services (http://www.nhs.uk). In the United States, CAHPS scores influence payments to hospitals and Medicare plans (Centers for Medicare & Medicaid Services, 2012; Medicare, 2016).

New Contributions

Understanding the meaning of such patient assessments of care within the context of accepted professional standards is crucial to the expectation that patient feedback can and should act as a catalyst to change. Previous research has explored the relationship between patient and examiner ratings of trainee general practitioner (GP) communication skills, and has found either no evidence of an association (in an underpowered study, with a sample size of 19 [McKinstry, Walker, Blaney, Heaney, & Begg, 2004]) or weak-to-moderate association (Greco, Spike, Powell, & Brownlea, 2002). More recently, a study of observer-rated verbal and nonverbal elements of a consultation found aspects of these predicted patient satisfaction with communication and the doctor–patient relationship (Little et al., 2015). However, no study has yet explored the association between patient assessments of communication skills on items used in national survey programs (and consequent quality metrics) and observer assessment of the performance of practicing physicians.

This study investigated the association between patient assessments of the quality of communication in their consultations with their Family Physician, using items derived from the English GP Patient Survey (Ipsos MORI, 2015), and a recently developed observer-rated measure of communication competence, the Global Consultation Rating Scale (GCRS), derived from the Calgary-Cambridge guide to the medical interview (Burt et al., 2014; Kurtz & Silverman, 1996; Kurtz, Silverman, Benson, & Draper, 2003).

Conceptual Model

In this study, we hypothesized that patient evaluations of physician communication, as expressed on patient experience instruments, may be influenced by a number of factors both internal and external to the consultation (see Figure 1). Some of these will be, and some will not be, visible to an outside observer. For example, both the patient and the physician they are consulting with bring various characteristics and experiences to each consultation which will determine the interaction. These come together as the overall "consultation experience," which is nested within both the previous physician–patient relationship (if there is one), and the relationship the patient has with the wider practice. These, together with the outcome of the consultation, will determine the patient's evaluation of physician communication. Thus, while an instrument may ask patients to rate their experience of being listened to or involved in decision making in a particular consultation, their choice of answer could be influenced by a host of other factors as well as these particular dimensions of care. For example, the same physician communication may produce different patient experiences in patients with different health literacy. An external observer, by contrast, is unlikely to know anything about the patient's past history with the practice or their characteristics and experiences,

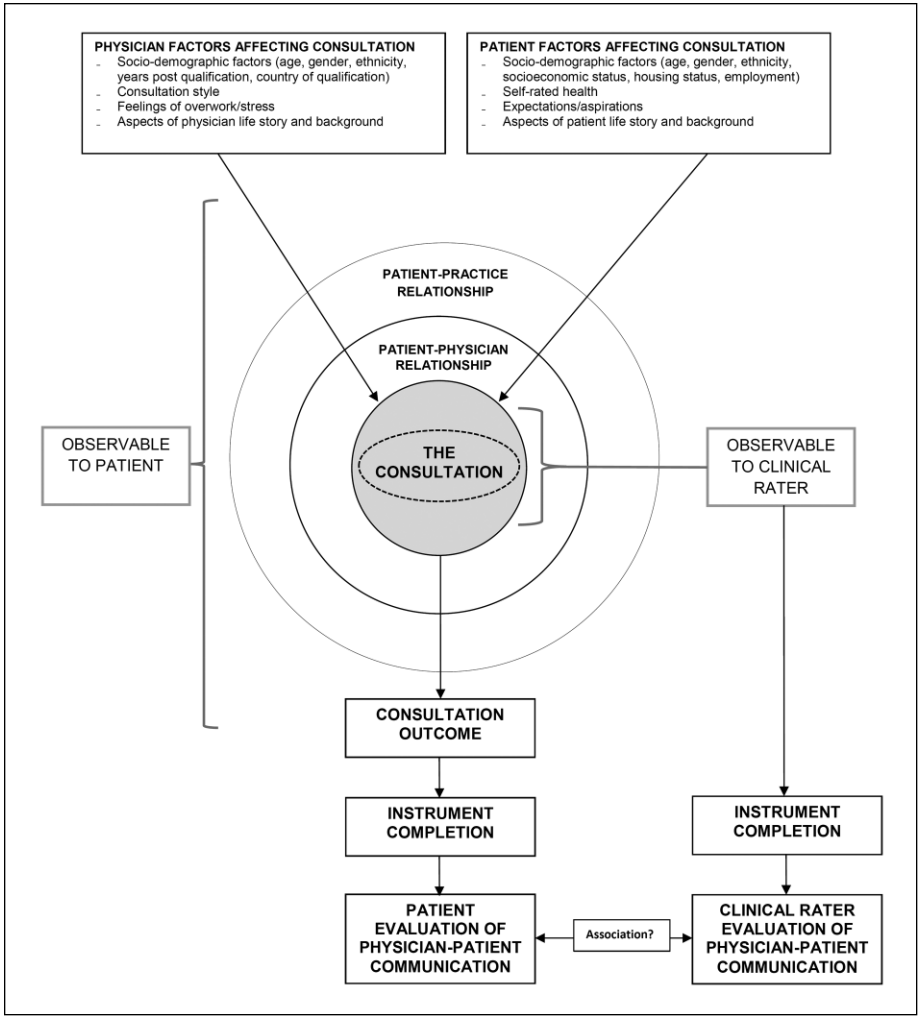


Figure 1. Conceptual model: mechanisms of patient and clinical rater assessment of physician-patient communication in a consultation.

other than that which is directly observable through the consultation. This limits the observer to rating to what can be seen and judged from the consultation itself. This is not to say that observers are free from influence, far from it, and variability in raters' assessments of the same clinical consultation is a well-known phenomenon (Burt et al., 2014). Through such mechanisms, we anticipated that it was possible that patient and clinical rater assessments may not always align, as patients and raters may be assessing related but distinct constructs. While raters assess the extent to which physicians' communication adheres to best practices, patients report on the effects of that communication on their health care experiences.

Method

The study was conducted in General Practices in England in two broad geographic areas (Devon, Cornwall, Bristol, Dorset, and Somerset; and Cambridgeshire, Bedford, Luton, and North London). Practices were eligible if they (a) had more than one Family Physician (hereafter just "physician") working a minimum of 2 days per week in direct clinical contact with patients and (b) had low scores on physician–patient communication items used in the national GP Patient Survey (defined as practices below the 25th percentile for mean communication score in the 2009/2010 survey, adjusted for patient case mix [Paddison et al., 2012]). Low-scoring practices were chosen to obtain more consultations with low patient ratings for communication than is typical: Nationally, 94% of patients score all questions addressing GP communication within consultations as good or very good (the two most positive options) in the GP Patient Survey.

Recruitment

In order to obtain the 28 required video recorded consultations that patients judged to have less than good communication (see below), many more consultations had to be video recorded than would be rated. The 28 required "good" consultations were more easily obtained, as they were more common. The research team approached adult patients on their arrival in participating practices and sought written informed consent to video record their consultation. Data collection continued until each required number was reached.

Patient Ratings

Immediately following the consultation the patient was asked to complete a short questionnaire. The questionnaire included a set of seven items taken from the national GP Patient Survey to assess physician–patient communication (Box 1), and basic sociodemographic questions. The seven patient communication items were previously derived to represent a single underlying construct; this has been confirmed using factor analysis (Campbell et al., 2009). As previously, we calculated a physician–patient communication score by linearly rescaling responses between 0 and 100 and taking the mean of all responses where four or more informative answers were given (Lyratzopoulos et al., 2012; Roberts et al., 2014).

Ratings by Trained Clinical Raters

We sampled 56 consultations for rating by experienced trained clinical raters. Raters scored each of the selected consultations using the GCRS (Burt et al., 2014). The GCRS is designed to assess the effectiveness of communication across an entire consultation, and is based on the widely used Calgary-Cambridge guide to the medical interview (Kurtz et al., 2003; Kurtz & Silverman, 1996). Raters score each consultation in 12 domains (including gathering information, building the relationship, providing structure, and achieving a shared understanding), resulting in a final score of between 0 and 10 (see supplementary material for full instrument [all supplementary

Box 1. GP-Patient Communication Items Used in the Patient Experience Survey.

Thinking about the consultation which took place *today*

How good was the doctor at each of the following?

Please put an × in one box for *each* row.

	Very good	Good	Neither good nor poor	Poor	Very poor	Doesn't apply*
Giving you enough time . . .	☐	☐	☐	☐	☐	☐
Asking about your symptoms . . .	☐	☐	☐	☐	☐	☐
Listening to you . . .	☐	☐	☐	☐	☐	☐
Explaining tests and treatments . . .	☐	☐	☐	☐	☐	☐
Involving you in decisions about your care	☐	☐	☐	☐	☐	☐
Treating you with care and concern . . .	☐	☐	☐	☐	☐	☐
Taking your problems seriously . . .	☐	☐	☐	☐	☐	☐

*Considered to be uninformative for the purposes of our analysis.

materials are available online at http://mcr.sagepub.com/content/by/supplemental-data]). Raters were physicians experienced in the teaching of communication skills; all attended a 2-hour training session on GCRS delivered by one of the original authors of the Calgary-Cambridge guide (JS). We used four raters for each consultation to increase reliability. Each rater scored consultations in a different random order to minimize the consequences of any order effects, and the same raters were used for all consultations. A simple mean of the four raters was calculated for each consultation.

From the rating of 56 consultations, we expected 80% power (.05 significance level) to detect a correlation coefficient of .37. To best measure this correlation our a priori sampling strategy included consultations with a wide range of scores: 28 (half) from those where all patient responses to the seven communication items were either *good* or *very good*, and 28 (half) where at least one rating was less than *good*. For the 28 "less-than-good" consultations, we selected the 28 consultations with the lowest patient communication scores. The 28 "good" consultations were selected at random. We further barred the inclusion of more than two consultations involving the same physician.

Statistical Analyses

First, we assessed the rater-adjusted consultation-level reliability of the GCRS scores by fitting a mixed-effects linear regression model to the 224 individual ratings (four

ratings of 56 consultations). Following the standard approach for adjusted unit-level reliability (e.g., Elliott et al., 2010), we included a random consultation effect for between-consultation variance (σ_b^2) and rater fixed effects, with the residual capturing within-consultation, between-rater, variance (σ_w^2) in ratings. The reliability (λ_{GCRS}) of the mean GCRS rating is as follows:

$$\lambda_{GCRS} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_w^2}{4}} \quad (1)$$

We explored the association between individual patient ratings and the mean ratings obtained by four trained raters using a simple correlation coefficient and scatter plot. Because adjusting for patient sociodemographic characteristics did not meaningfully reduce standard errors, unadjusted results are shown. Bootstrapping with 1,000 replicates, clustered within physicians, accounted for some physicians being included twice and for possible deviations from normality. Further illustration was provided by dichotomizing patient ratings into below 75 (requiring at least one less-than-good response) versus higher and cross-tabulating this with tertiles of GCRS ratings. The resulting 2 × 3 association was tested using logistic regression with a sandwich estimator to account for clustering by physician (Rogers, 1993).

The data analyzed relate to patient ratings of a particular consultation, rather than many patients' ratings of a particular physician. Even when the association between patient scores and rater scores is weak, it may be that by aggregating scores from many patients, reliable physician scores may be obtained. We illustrate this concept by simulating scores for 100 hypothetical physicians with a range of communication skills as measured by GCRS. The patient ratings for a given GCRS score are drawn from a distribution informed by the findings of the observational work. For each physician, mean patient scores are calculated for 1, 10, 30, and 100 patients. In this illustration physicians are assumed to score consistently on GCRS for all consultations.

All analysis was carried out using Stata V13.1 (StataCorp, 2015, Stata Statistical Software: Release 13, College Station, TX).

Results

Consultations with 45 participating physicians from 13 general practices were video recorded. Of 741 eligible patients, 529 (71.4%) consented to participate and completed a questionnaire (see supplementary material for a recruitment flowchart). Reasons patients gave for declining participation commonly related to the nature of their clinical problem. The videos selected for rating using GCRS came from all 13 general practices and included 37 physicians. Table 1 shows the self-reported demographics of patients who completed a questionnaire, along with those whose consultation was selected for rating by trained raters. Men, 18- to 24-year-olds, and Asian patients were somewhat more likely to have been selected to have their consultations rated. The distribution of patient scores and GCRS ratings is shown in Figure 2. Patient scores were highly skewed: The most common score (found for 21/56 consultations) was 100 out of a

Table 1. Self-Reported Demographics for Patients Who Completed a Questionnaire and Those Selected for Consultations to be Rated by Trained Raters.

	Completed questionnaire, *n* (%)	Rated consultations, *n* (%)
Sex		
Male	212 (40.15)	26 (46.43)
Female	316 (59.85)	30 (53.57)
Age, years		
18-24	39 (7.41)	10 (18.18)
25-34	78 (14.83)	7 (12.73)
35-44	64 (12.17)	7 (12.73)
45-54	82 (15.59)	4 (7.27)
55-64	85 (16.16)	8 (14.55)
65-74	103 (19.58)	7 (12.73)
75-84	60 (11.41)	8 (14.55)
85+	15 (2.85)	4 (7.27)
Self-rated health		
Excellent	50 (9.51)	3 (5.36)
Very good	173 (32.89)	14 (25)
Good	182 (34.60)	23 (41.07)
Fair	83 (15.78)	13 (23.21)
Poor	38 (7.22)	3 (5.36)
Ethnicity		
White	474 (90.98)	44 (81.48)
Mixed	5 (0.96)	1 (1.85)
Asian or Asian British	15 (2.88)	6 (11.11)
Black or Black British	22 (4.22)	1 (1.85)
Chinese	4 (0.77)	1 (1.85)
Other	1 (0.19)	1 (1.85)

possible 100 (i.e., *very good* for all reported communication items). The median score was 91 (interquartile range 71-100) and the lowest reported score 31. In contrast, the GCRS ratings are reasonably symmetrical: The median GCRS score was 4.3 of 10 (interquartile range 3.6-5.5) and scores ranged from 2.2 to 6.8. The estimated variance components of the GCRS ratings were 1.01 between consultations and 1.18 within consultations (between rater). Reliability for the mean of four ratings was 0.77.

Figure 3 shows patient scores plotted against average GCRS ratings for each consultation. There is weak evidence ($p = .054$) of an association between patient scores and GCRS ratings, with an unadjusted correlation coefficient of .29 (reliability-adjusted $r = .33$; Muchinsky, 1996). When trained raters assessed communication within a consultation to be of a high standard (highest tertile), patients tended to do the same (with the exception of a single outlying low patient score). However, when trained raters judged communication within a consultation to be poor (lowest tertile), patients reported communication ranging from poor to very good. This is illustrated in

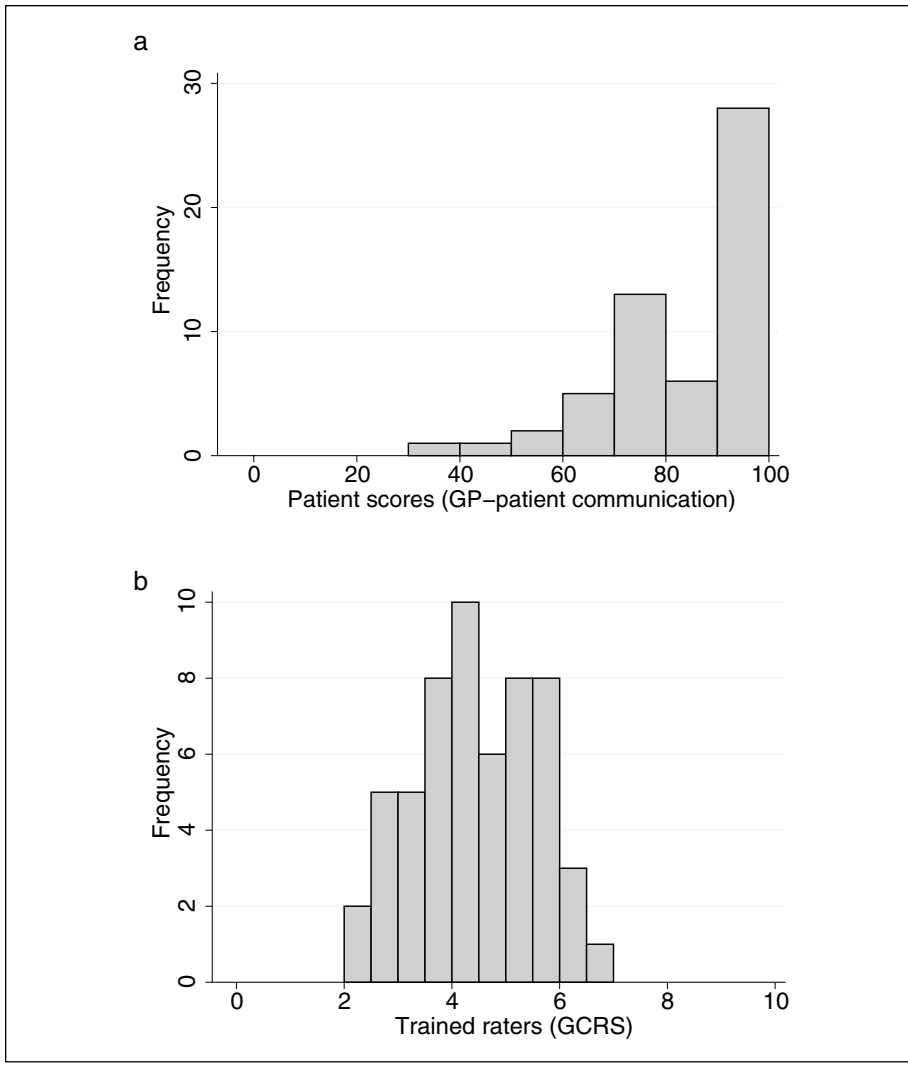


Figure 2. Distribution of patient scores based on GP patient survey items and ratings by trained raters on the GCRS.
Note. GCRS = Global Consultation Rating Scale.

Table 2, which shows that in the consultations in the lowest third of rater scores 58% of patient scores were under 75 (out of 100) compared with 17% in the highest third.

Figure 4 shows the results of the simulation study, which is based on a hypothetical set of consultations with a range of trained rater scores (GCRS). For each GCRS score, we defined a range of possible simulated patient scores, shown by the shaded grey area in Figure 4. The lower limit of these simulated patient scores increased as GCRS score

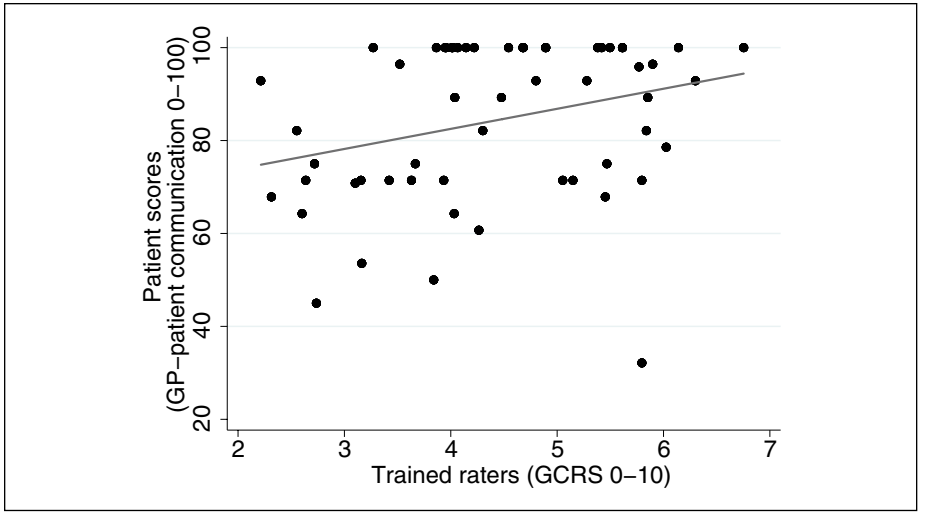


Figure 3. Scatterplot comparing patient scores based on GP patient survey items and ratings by trained raters on the GCRS.
Note. GCRS = Global Consultation Rating Scale.

Table 2. Comparison of Lower (<75) and Higher (≥75) Patient Scores for Consultations in Each Third of the Distribution of Rater Scores.

	Patient score (GP–patient communication, 0-100 scale)		
	<75, *n* (%)	≥75, *n* (%)	Total, *n* (%)
Trained raters (GCRS 0-10 scale)			
Lowest third	11 (57.9)	8 (42.1)	19 (100)
Middle third	4 (21.1)	15 (79.0)	19 (100)
Highest third	3 (16.7)	15 (83.3)	18 (100)
Total	18 (32.1)	38 (67.9)	56 (100)

Note. GCRS = Global Consultation Rating Scale. Test of association from logistic regression accounting for clustering by physician, $p = .049$.

increased. However, the upper limit of simulated patient scores was set at 100 for all possible GCRS scores in the simulation. For any given GCRS score, we allowed patient scores to take any value in this range, with equal probability. The simulation is designed for illustrative purposes only and is not intended to directly reflect our current findings. Panel A, designed to be reminiscent of Figure 3, shows what would be observed with just a single patient score per physician, that is, a weak correlation between patient rating and communication skill. The remaining panels illustrate the effect of combining scores (taking the mean) from multiple consultations, rather than using a single rating. As the number of patient ratings taken increases, the correlation

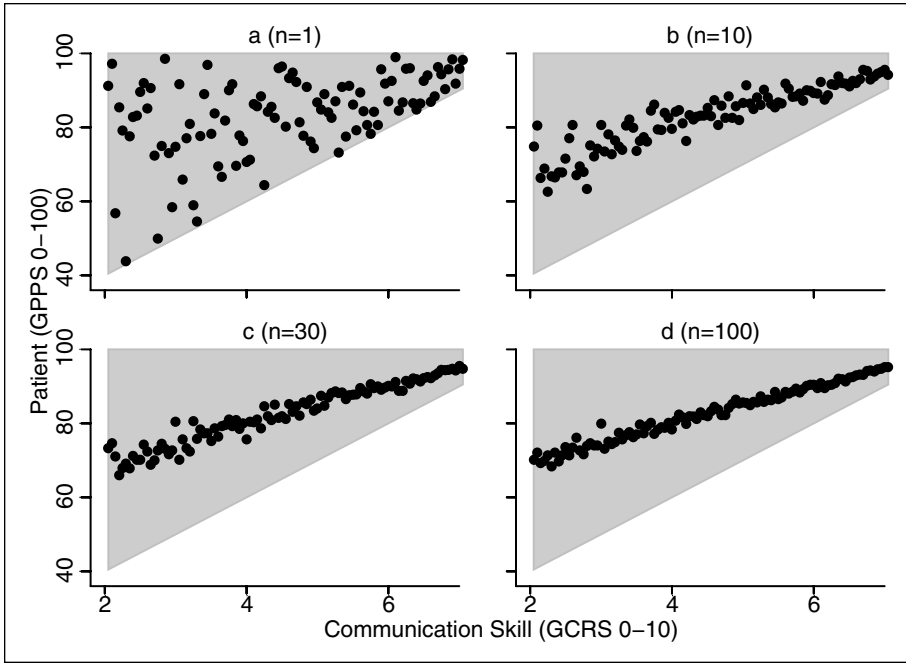


Figure 4. Results of simulation illustrating the effect of estimating physician communication scores based on different numbers of patient ratings. In the simulation the range of individual patient sores we allowed to be taken for any one value of communication skill are shown by the grey areas. The black points show the scores for individual simulated physicians when averaging over 1 (a), 10 (b), 30 (c) and 100 (d) patient scores for each physician.
Note. GCRS = Global Consultation Rating Scale.

between trained rater scores and patient scores gets stronger. When the number of consultations are 30, this correlation becomes very strong ($\rho = .97$), becoming stronger still when $n = 100$.

Discussion

This study aimed to compare patient assessments of physician communication with measures of adherence to professionally agreed standards and norms of physician communication. We found a weak positive correlation between patients' and trained clinical raters' assessments of the quality of communication in physician–patient consultations. This suggests that there is an association between patient ratings of communication and professionally defined standards of care. Importantly, when trained clinical raters identified communication as good, patients tended to agree with this. However, when trained clinical raters identified communication as poor, patients ranged in their assessments of communication from poor to very good.

National survey programs commonly feedback patients' evaluations to physicians, assuming that mechanisms such as reputational concern will drive performance improvements (Contandriopoulos, Champagne, & Denis, 2014; Fung, Lim, Mattke, Damberg, & Shekelle, 2008). However, recent evidence suggests that physicians struggle to make sense of patient experience feedback and may dismiss findings rather than initiate change (Asprey et al., 2013; Boiko et al., 2014). The persistence of concerns about the reliability and validity of the questionnaires used are key factors contributing toward disregard for patient survey results (Boiko et al., 2014). These concerns are complex, and driven in part by limited information regarding how much patient ratings: (a) confined to perceive physician communication quality alone, without being influenced by other considerations or aspects of care or (b) match accepted professional standards of communication. Our study aimed to address the latter question, and our findings suggest that, while trained raters' and patient's tend to agree what good communication looks like in a consultation, clinical raters are more likely than patients to judge communication as poor.

We outline two possible mechanisms driving this divergence: In both cases, we are benchmarking the patient ratings against those of the raters (an issue we discuss further below). The first mechanism arises from the well-known phenomenon of skewed patients' ratings, or positive response tendency, with a large proportion of patients rating communication as "very good" (Campbell et al., 2009; Elliott, Haviland, Kanouse, Hambarsoomian, & Hays, 2009; Rodriguez & Crane, 2011). By contrast, GCRS ratings tend to cluster around the middle scores available to raters. It is therefore possible that the poor measurement of above-average experiences inherent in the patient question items may artificially limit the responses patients would like to give (thereby applying a ceiling effect), preventing them from being able to distinguish the very best consultations from those they judge to be simply good. This mechanism requires that patients differ from raters in their views of what good or poor communication in a consultation looks like. As a result, the more positive patient opinion is "held back" by only being able to endorse questionnaire options ranging from very poor to very good (and not, e.g., "excellent"), despite extensive instrument development (Ipsos MORI, 2015).

However, the second (and we argue more plausible) mechanism is that there are wider factors at play which inhibit some patients from assigning poor scores to consultations that they do perceive as involving poor communication. It is important to note that any such inhibition would have to apply unevenly between patients to explain the range of patient scores seen for consultations rated as poor by the trained raters: While some patients are easily able to choose "poor" as an option, others feel less able to do so. This is distinct from the ceiling effect described above in so far as consultations are not being underrated due to the maximum available rating, but that ratings are often higher than the consultation might merit.

While we are unable to determine the relative contribution of either mechanism from the methodology of this current study, there is existing evidence that patients may be inhibited in their judgments of care. This hypothesis is consistent with evidence that tendencies to avoid negative and extreme responses vary across patients (Elliott et al., 2009; Mayer, Elliott, Haas, Hays, & Weinick, 2016). In addition, qualitative research

has identified a number of psychological and social factors that suggest patients struggle to criticize physicians' performance in surveys. For example, an investigation into how patients evaluated community mental health services found that they frequently avoided giving negative scores on experience questionnaires: Instead, allowances for poor care were constructed by referencing their perceptions of the duties and culpabilities of health care providers (Williams, Coyle, & Healy, 1998). Similarly, patients undergoing elective orthopedic surgery reinterpreted their experiences in a positive light as a result of feelings of dependency on their health care providers, and a perceived need to maintain constructive relationships with physicians (Edwards, Staniszweska, & Crichton, 2004). A tendency to excuse rather than report poor care has also been identified in breast cancer patients (Davoll et al., 2013). In light of this, we are currently undertaking further qualitative research with our sample to determine whether such factors may be found within the assessment of patient experience in primary care.

Limitations

Our sampling strategy was informed by the need to locate consultations patients identified as less than good; the proportion of such consultations is small, so to increase study efficiency, we deliberately approached some practices who had received lower scores for communication in the national GP Patient Survey. Not all physicians in every practice took part, and it is possible that the physicians who participated were more confident in their ability to communicate with patients. If those physicians who had poorer communication skills did not participate, this may have reduced the variability of the communication quality in our sample, in turn reducing study power and, potentially, the strength of the observed correlation. Power was also limited by the number of consultations rated and, while the study was not powered to detect weak correlations, it did have sufficient power to detect moderate ones.

Our patient consent rate was 71.4% of eligible patients. The research team missed only a small number of patients (2.0% [15/741] of those eligible), so exclusions predominantly reflect those who did not consent to participate. Recorded consultations concerning some medical conditions may be underrepresented as participants may have been more likely to decline being video recorded: While we were not able to elicit detailed reasons from patients who declined to participate, our observations suggest that in some cases this was due to sensitive presenting complaints. However, participants' age, gender, self-rated health, and ethnicity were broadly representative of the population attending general practice.

We assessed communication using two well-validated instruments: the GP Patient Survey items for patients and the GCRS for trained raters (Burt et al., 2014; Ipsos MORI, 2015). The GCRS was derived from the Calgary-Cambridge guide, which is used widely for communication skills training, and represents agreed professional norms of high-quality communication (Gillard, Benson, & Silverman, 2009; Kurtz et al., 2003; Kurtz & Silverman, 1996). Recently, the question has arisen as to how and whether trained raters take account of contextual factors in assessing the communication skills of physicians, for example, by allowing variations from "accepted practice"

when scoring performance in particular situations (Essers et al., 2013; Essers et al., 2014). However, the GCRS has been explicitly designed to focus only on the consultation process, and contains no task-based items which may be context-specific. Additionally, it enables raters to choose "not applicable" where necessary: In fact, this was rarely endorsed by raters in this study.

As mentioned above, in drawing conclusions about the meaning of patients' ratings of communication quality, we compare them with assessments by trained clinical rater. This is not to suggest raters are more valued or competent assessors of communication than patients, but simply to use them as representative of professionally agreed norms of behavior against which to judge patient evaluations of communication. In doing so, we are able to provide evidence that to some extent patient assessments tap in to the same underlying construct of communication drawn on by trained raters, but also that patients are less likely to judge consultations as poor.

Conclusions

Patient experience surveys are widely used to assess the standard of care provision. While physicians rated poorly by patients are generally rated poorly by trained raters, our findings suggest that patients may be inhibited in criticizing doctors' performances. Mean patient survey scores are likely to overestimate adherence to best physician communication practices, and treating apparently high patient experience scores as indicating absolutely high physician or practice performance is inadvisable. However, the use of relative rankings to identify physicians who are better or poorer at communicating with patients may be an acceptable approach to benchmarking performance, as long as statistically reliable figures are obtained. Previous research has demonstrated that the GP Patient Survey communication questions can differentiate between the performance of physicians and practices, as long as an adequate sample size is used to achieve acceptable statistical reliability (Lyratzopoulos et al., 2011; Roberts et al., 2014). This was confirmed by our simulation: With sufficient patient scores, a strong correlation between patient rating and rater evaluations will be observed. In the use of patient experience scores as quality indicators, our findings suggest that it is therefore possible to (a) trust aggregated patients scores that meet traditional standards of reliability as valid measures of comparative performance with respect to communication and (b) trust relatively low mean patient ratings as indicating poor performance. However, crucially, we cannot necessarily assume that an apparently high mean patient rating means all is well. Thus, lower patient experience scores should spur improvement efforts and higher scores should not breed complacency.

Acknowledgments

We would like to thank the patients, practice managers, GPs, and other staff of the general practices who kindly agreed to participate in this study and without whom the study would not have been possible. Particular acknowledgement goes to our four trained clinical raters for their contribution to this work, and to James Brimicombe, our data manager, who developed the online

rating system. We would also like to thank the Improve Advisory Group for their input and support throughout this study.

Authors' Note

The views expressed are those of the author and not necessarily those of the NHS, the NIHR, or the Department of Health. Approval for the study was obtained from the NRES Committee East of England—Hertfordshire on October 11, 2011 (ref: 11/EE/0353).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by a National Institute for Health Research Programme Grant for Applied Research (NIHR PGfAR) program (RP-PG-0608-10050).

References

Accreditation Council for Graduate Medical Education. (2015). *Accreditation Council for Graduate Medical Education (ACGME)*. Retrieved from http://www.acgme.org/acgmeweb/

Agency for Healthcare Research and Quality. (2015). *CAHPS Clinician & Group Surveys*. Retrieved from https://cahps.ahrq.gov/surveys-guidance/cg/index.html

Asprey, A., Campbell, J. L., Newbould, J., Cohn, S., Carter, M., Davey, A., & Roland, M. (2013). Challenges to the credibility of patient feedback in primary healthcare settings: A qualitative study. *British Journal of General Practice*, *63*, e200-e208. doi:10.3399/bjgp13X664252

Boiko, O., Campbell, J. L., Elmore, N., Davey, A. F., Roland, M., & Burt, J. (2014). The role of patient experience surveys in quality assurance and improvement: A focus group study in English general practice. *Health Expectations*, *18*, 1982-1994. doi:10.1111/hex.12298

Burt, J., Abel, G., Elmore, N., Campbell, J., Roland, M., Benson, J., & Silverman, J. (2014). Assessing the communication quality of consultations in primary care: Initial reliability assessment of the Global Consultation Rating Scale, based on the Calgary-Cambridge guide to the medical interview. *BMJ Open*, *4*, e004339. doi:10.1136/bmjopen-2013-004339

Campbell, J., Smith, P., Nissen, S., Bower, P., Elliott, M., & Roland, M. (2009). The GP Patient Survey for use in primary care in the National Health Service in the UK—Development and psychometric characteristics. *BMC Family Practice*, *10*, 57. doi:10.1186/1471-2296-10-57

Carcaise-Edinboro, P., & Bradley, C. J. (2008). Influence of patient-provider communication on colorectal cancer screening. *Medical Care*, *46*, 738-745. doi:10.1097/MLR.0b013e318178935a

Care Quality Commission. (2015). *How CQC regulates: NHS GP practices and GP out-of-hours services*. London, England: Author.

Centers for Medicare & Medicaid Services. (2012). *2014 Quality bonus payment determinations and administrative review process for quality bonus payments and rebate*

retention allowances. Retrieved from https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/2014QualityBonusPaymentDeterminations-.pdf

Contandriopoulos, D., Champagne, F., & Denis, J. L. (2014). The multiple causal pathways between performance measures' use and effects. *Medical Care Research and Review*, *71*, 3-20. doi:10.1177/1077558713496320

Davoll, S., Kowalski, C., Kuhr, K., Ommen, O., Ernstmann, N., & Pfaff, H. (2013). "Tendency to excuse" and patient satisfaction of those suffering with breast cancer. *International Journal of Public Health*, *58*, 385-393. doi:10.1007/s00038-012-0405-6

Doyle, C., Lennox, L., & Bell, D. (2013). A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*, *3*, e001570. doi:10.1136/bmjopen-2012-001570

Duffy, F. D., Gordon, G. H., Whelan, G., Cole-Kelly, K., Frankel, R., Buffone, N., . . . Langdon, L. (2004). Assessing competence in communication and interpersonal skills: The Kalamazoo II report. *Academic Medicine*, *79*, 495-507.

Edwards, C., Staniszweska, S., & Crichton, N. (2004). Investigation of the ways in which patients' reports of their satisfaction with healthcare are constructed. *Sociology of Health & Illness*, *26*, 159-183. doi:10.1111/j.1467-9566.2004.00385.x

Elliott, M. N., Haviland, A. M., Kanouse, D. E., Hambarsoomian, K., & Hays, R. D. (2009). Adjusting for subgroup differences in extreme response tendency in ratings of health care: Impact on disparity estimates. *Health Services Research*, *44*(2 Pt 1), 542-561. doi:10.1111/j.1475-6773.2008.00922.x

Elliott, M. N., Lehrman, W. G., Goldstein, E., Hambarsoomian, K., Beckett, M. K., & Giordano, L. A. (2010). Do hospitals rank differently on HCAHPS for different patient subgroups? *Medical Care Research and Review*, *67*, 56-73. doi:10.1177/1077558709339066

Essers, G., Dielissen, P., van Weel, C., van der Vleuten, C., van Dulmen, S., & Kramer, A. (2014). How do trained raters take context factors into account when assessing GP trainee communication performance? An exploratory, qualitative study. *Advances in Health Sciences Education*, *20*, 1-17. doi:10.1007/s10459-014-9511-y

Essers, G., van Dulmen, S., van Es, J., van Weel, C., van der Vleuten, C., & Kramer, A. (2013). Context factors in consultations of general practitioner trainees and their impact on communication assessment in the authentic setting. *Patient Education & Counseling*, *93*, 567-572. doi:10.1016/j.pec.2013.08.024

Fung, C. H., Lim, Y. W., Mattke, S., Damberg, C., & Shekelle, P. G. (2008). Systematic review: The evidence that publishing patient care performance data improves quality of care. *Annals of Internal Medicine*, *148*, 111-123.

General Medical Council. (2015). *An introduction to revalidation*. Retrieved from http://www.gmc-uk.org/doctors/revalidation/9627.asp

Gillard, S., Benson, J., & Silverman, J. (2009). Teaching and assessment of explanation and planning in medical schools in the United Kingdom: Cross sectional questionnaire survey. *Medical Teacher*, *31*, 328-331. doi:10.1080/01421590801953018

Greco, M., Spike, N., Powell, R., & Brownlea, A. (2002). Assessing communication skills of GP registrars: A comparison of patient and GP examiner ratings. *Medical Education*, *36*, 366-376. doi:10.1046/j.1365-2923.2002.01175.x

Ipsos MORI. (2015). *GP Patient Survey*. Retrieved from https://gp-patient.co.uk/

Kurtz, S., Silverman, J., Benson, J., & Draper, J. (2003). Marrying content and process in clinical method teaching: Enhancing the Calgary-Cambridge guides. *Academic Medicine*, *78*, 802-809.

Kurtz, S. M., & Silverman, J. D. (1996). The Calgary-Cambridge Referenced Observation Guides: An aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical Education*, *30*, 83-89.

Kuzel, A. J., Woolf, S. H., Gilchrist, V. J., Engel, J. D., LaVeist, T. A., Vincent, C., & Frankel, R. M. (2004). Patient reports of preventable problems and harms in primary health care. *Annals of Family Medicine*, *2*, 333-340.

Little, P., White, P., Kelly, J., Everitt, H., Gashi, S., Bikker, A., & Mercer, S. (2015). Verbal and non-verbal behaviour and patient perception of communication in primary care: An observational study. *British Journal of General Practice*, *65*, e357-e365. doi:10.3399/bjgp15X685249

Llanwarne, N. R., Abel, G. A., Elliott, M. N., Paddison, C. A., Lyratzopoulos, G., Campbell, J. L., & Roland, M. (2013). Relationship between clinical quality and patient experience: Analysis of data from the English quality and outcomes framework and the National GP Patient Survey. *Annals of Family Medicine*, *11*, 467-472. doi:10.1370/afm.1514

Lyratzopoulos, G., Elliott, M., Barbiere, J. M., Henderson, A., Staetsky, L., Paddison, C., . . . Roland, M. (2012). Understanding ethnic and other socio-demographic differences in patient experience of primary care: Evidence from the English General Practice Patient Survey. *BMJ Quality & Safety*, *21*, 21-29. doi:10.1136/bmjqs-2011-000088

Lyratzopoulos, G., Elliott, M. N., Barbiere, J. M., Staetsky, L., Paddison, C. A., Campbell, J., & Roland, M. (2011). How can health care organizations be reliably compared? Lessons from a national survey of patient experience. *Medical Care*, *49*, 724-733. doi:10.1097/MLR.0b013e31821b3482

Mayer, L. A., Elliott, M. N., Haas, A., Hays, R. D., & Weinick, R. M. (2016). Less use of extreme response options by Asians to standardized care scenarios may explain some racial/ethnic differences in CAHPS scores. *Medical Care*, *54*, 38-44. doi:10.1097/MLR.0000000000000453

McKinstry, B., Walker, J., Blaney, D., Heaney, D., & Begg, D. (2004). Do patients and expert doctors agree on the assessment of consultation skills? A comparison of two patient consultation assessment scales with the video component of the MRCGP. *Family Practice*, *21*, 75-80.

Medicare. (2016). *Hospital Value-Based Purchasing*. Retrieved from https://www.medicare.gov/hospitalcompare/Data/hospital-vbp.html

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63-75. doi:10.1177/0013164496056001004

Orth, J. E., Stiles, W. B., Scherwitz, L., Hennrikus, D., & Vallbona, C. (1987). Patient exposition and provider explanation in routine interviews and hypertensive patients' blood pressure control. *Health Psychology*, *6*, 29-42.

Paddison, C., Elliott, M., Parker, R., Staetsky, L., Lyratzopoulos, G., Campbell, J. L., & Roland, M. (2012). Should measures of patient experience in primary care be adjusted for case mix? Evidence from the English General Practice Patient Survey. *BMJ Quality & Safety*, *21*, 634-640. doi:10.1136/bmjqs-2011-000737

Paddison, C. A., Abel, G. A., Roland, M. O., Elliott, M. N., Lyratzopoulos, G., & Campbell, J. L. (2013). Drivers of overall satisfaction with primary care: Evidence from the English General Practice Patient Survey. *Health Expectations*, *18*, 1081-1092. doi:10.1111/hex.12081

Price, R. A., Elliott, M. N., Cleary, P. D., Zaslavsky, A. M., & Hays, R. D. (2015). Should health care providers be accountable for patients' care experiences? *Journal of General Internal Medicine*, *30*, 253-256. doi:10.1007/s11606-014-3111-7

Quigley, D. D., Elliott, M. N., Farley, D. O., Burkhart, Q., Skootsky, S. A., & Hays, R. D. (2014). Specialties differ in which aspects of doctor communication predict overall physician ratings. *Journal of General Internal Medicine*, *29*, 447-454. doi:10.1007/s11606-013-2663-2

Roberts, M. J., Campbell, J. L., Abel, G. A., Davey, A. F., Elmore, N. L., Maramba, I., . . . Burt, J. A. (2014). Understanding high and low patient experience scores in primary care: Analysis of patients' survey data for general practices and individual doctors. *British Medical Journal*, *349*, g6034. doi:10.1136/bmj.g6034

Rodriguez, H. P., & Crane, P. K. (2011). Examining multiple sources of differential item functioning on the Clinician & Group CAHPS® survey. *Health Services Research*, *46*(6 Pt 1), 1778-1802. doi:10.1111/j.1475-6773.2011.01299.x

Rogers, W. H. (1993). sg17: Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13*, 19-23.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., . . . Hanley, J. A. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *Journal of the American Medical Association*, *298*, 993-1001. doi:10.1001/jama.298.9.993

Turner, J. L., & Dankoski, M. E. (2008). Objective structured clinical exams: A critical review. *Family Medicine*, *40*, 574-578.

Vincent, C., Davy, C., Esmail, A., Neale, G., Elstein, M., Cozens, J. F., & Walshe, K. (2006). Learning from litigation: The role of claims analysis in patient safety. *Journal of Evaluation in Clinical Practice*, *12*, 665-674. doi:10.1111/j.1365-2753.2006.00634.x

Williams, B., Coyle, J., & Healy, D. (1998). The meaning of patient satisfaction: An explanation of high reported levels. *Social Science & Medicine*, *47*, 1351-1359.

Zill, J. M., Christalle, E., Müller, E., Härter, M., Dirmaier, J., & Scholl, I. (2014). Measurement of physician-patient communication—A systematic review. *PLoS ONE*, *9*, e112637.

Zolnierek, K. B., & Dimatteo, M. R. (2009). Physician communication and patient adherence to treatment: A meta-analysis. *Medical Care*, *47*, 826-834. doi:10.1097/MLR.0b013e31819a5acc