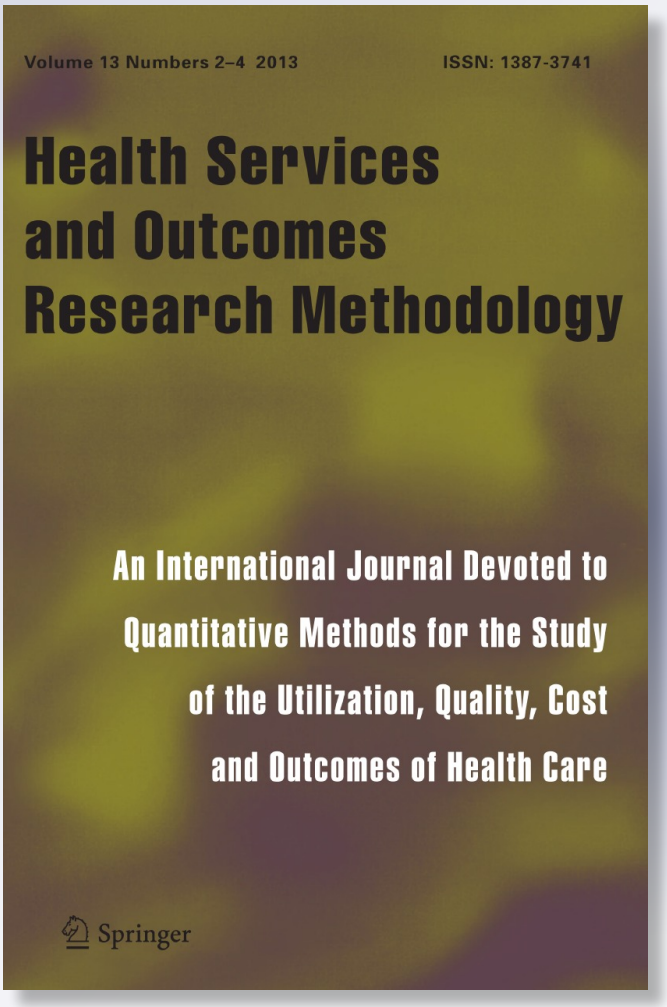
Modern psychometric methods for estimating physician performance on the Clinician and Group CAHPS® survey

Shubhabrata Mukherjee, Hector P. Rodriguez, Marc N. Elliott & Paul K. Crane

Health Services and Outcomes Research Methodology
An International Journal Devoted to Quantitative Methods for the Study of the Utilization, Quality, Cost and Outcomes of Health Care

ISSN 1387-3741
Volume 13
Combined 2-4

Health Serv Outcomes Res Method
(2013) 13:109-123
DOI 10.1007/s10742-013-0111-8



Springer

Health Serv Outcomes Res Method (2013) 13:109–123
DOI 10.1007/s10742-013-0111-8

Modern psychometric methods for estimating physician performance on the Clinician and Group CAHPS® survey

Shubhabrata Mukherjee · Hector P. Rodriguez · Marc N. Elliott · Paul K. Crane

Received: 23 February 2013 / Revised: 18 October 2013 / Accepted: 30 October 2013 /
Published online: 12 November 2013

Abstract Modern psychometric methods for scoring the Clinician & Group Consumer Assessment of Healthcare Providers and Systems (CG-CAHPS®) instrument can improve the precision of patient scores. The extent to which these methods can improve the reliable estimation and comparison of individual physician performance, however, remains unclear. Using CG-CAHPS® data from 12,244 unique patients of 448 primary care physicians in southern California, four methods were used to calculate composite scores: (1) standard scoring, (2) a single factor confirmatory factor analysis model, (3) a bifactor model, and (4) a correlated factor model. We extracted factor scores for physicians from each model and adjusted the scores for respondent characteristics, including age, education, self-rated physical health, and race/ethnicity. Physician-level reliability and physician rankings were examined across the four methods. The bifactor and correlated factor models achieved the best fit for the core CG-CAHPS® questions from the three core composite measures. Compared to standard adjusted scoring, the bifactor model scores resulted in a 25 % reduction in required sample sizes per physician. The correlation of physician rankings between scoring methods ranged from 0.58 to 0.86. The discordance of physician rankings across scoring methods was most pronounced in the middle of the performance distribution. Using modern psychometric methods to score physician performance on the core CG-CAHPS® questions may improve the reliability of physician performance estimates on patient experience measures, thereby reducing the required respondent sample sizes per physician compared to standard scoring. To assess the predictive validity of the CG-CAHPS® scores generated by modern psychometric methods, future research should

S. Mukherjee · P. K. Crane
Division of Internal Medicine, School of Medicine, University of Washington, Box 359780, Seattle, WA 98112, USA

H. P. Rodriguez (✉)
Department of Health Policy and Management, UCLA Fielding School of Public Health, University of California, Los Angeles, Box 951772, Los Angeles, CA 90095-1772, USA
e-mail: hrod@berkeley.edu

M. N. Elliott
RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, USA

examine the relative association of different scoring methods and important patient-centered outcomes of care.

Keywords Performance measurement · CAHPS® · Primary care physicians · Psychometric analyses · Ranking

1 Background

The Clinician & Group Consumer Assessment of Healthcare Providers and Systems (CG-CAHPS®) survey is increasingly being used in public reporting and pay-for-performance initiatives in the United States (Browne et al. 2010; Rodriguez et al. 2009b) and to examine the impact of physician communication training (Beach et al. 2005; Rao et al. 2007) and primary care delivery system interventions (Browne et al. 2010; Campbell et al. 2010; Sequist et al. 2009). Further, physician-specific patient experience data are also being used to facilitate patient selection of primary care physicians and primary care practice sites in some settings (Fanjiang et al. 2007). Given the utility of physician-specific CG-CAHPS® information in providing actionable information to ambulatory care stakeholders (Browne et al. 2010), valid and reliable data collection are important considerations. Collecting physician-specific information from patients, however, can come at a substantial cost to physician organizations, payers, and quality improvement initiatives (Rodriguez et al. 2006). To improve measurement precision and reduce patient survey data collection costs in quality improvement (QI) initiatives, it is important that analytic methods maximize the efficiency of making physician and practice comparisons on patient experience measures.

Statistical advances may improve measurement precision or physician-level reliability, thereby reducing patient sample size requirements and costs of measuring physician performance on patient experience measures (Holmboe et al. 2010; Kaplan et al. 2009; Lyratzopoulos et al. 2011). Previous analyses of the CG-CAHPS® data indicate that 30–45 patient responses per physician are required to achieve adequate physician-level reliability ($\alpha_{MD} = 0.70$) when unweighted averages (total or "standard" scores) are used to calculate composite measures (Hays et al. 2003a; Nelson et al. 2004; Rodriguez et al. 2009a, c; Roland et al. 2009; Safran et al. 2006b). Although these sample size requirements are modest, patients are increasingly reluctant to participate in surveys and response rates to patient surveys have been steadily declining, increasing costs per completed survey. Consequently, analytic approaches that enable equally valid and reliable estimates of individual physician performance on patient experience measures with less data can reduce overall survey data collection requirements and costs.

1.1 The CG-CAHPS® as a unidimensional or multidimensional survey

The composite scoring method delineated in the CG-CAHPS® analysis guidance recommends that users calculate three case-mix adjusted composite scores from the completed survey questions by taking the unweighted average of questions comprising each of the three core composite measures: physician communication, access to care, and office staff interactions (Dyer et al. 2012). Cognitive testing of CAHPS® surveys suggests that patients think of the patient experience domains as separate constructs and better understand the data presented in that form (Harris-Kojetin et al. 1999; Schnaier et al. 1999). However,

recent research suggests that the ambulatory care CAHPS® questions can be modeled as unidimensional and multidimensional constructs (Reise et al. 2007), underscoring that there is substantial covariation among CAHPS® core composite measures, i.e., communication, access to care, and office staff interaction. Because of the small number of items compromising the office staff interactions composite (2 items) and our interest in assessing the impact of using psychometric methods on the ranking of individual physicians, we focused our analyses on comparing overall physician composite scores across scoring methods rather than separate composite scores for each the three core measures.

1.2 Composite scoring methods and item difficulty

We were interested in exploring how different methods for scoring patient responses to the CG-CAHPS® surveys (Table 3) can influence the rankings of individual primary care physicians. Quantitative comparisons of different factor models were presented in Reise et al. 2007. The standard scoring approach (Model A) is the most common method utilized for ranking physicians based on patient experience measures (Anastario et al. 2010). Although standard scoring is simple to understand and implement, it ignores the distribution of item difficulties. For standard scoring to discriminate performance well across the whole performance range, the item difficulties must be evenly spaced across the continuum of the trait or attribute measured by the test. One way to assess this is by examining the test characteristic curve, a plot of the most likely standard score associated with each level of the trait or attribute, i.e. the patient's experience of care. The test characteristic curve for the CG-CAHPS® questions is curvilinear (Appendix B), indicating that a 1-unit difference in the standard score at the lower end of the scale (people with worse care experiences) does not provide the same level of performance discrimination as an identical 1-unit difference at the higher end of the scale (people with more favorable care experiences) (Rodriguez and Crane 2011). The CG-CAHPS® includes a diverse set of questions related to physician-level experiences (physician communication) and questions assessing primary care clinic experiences (access to care and office staff interactions). It is likely that the various questions do not contribute equally to patients' overall experiences of care. Failure to address curvilinearity may lead to biased estimates of the underlying trait, i.e., patient's experience of care, when standard scores are employed in linear regression models or other statistical approaches that assume a linear scaling metric.

Modern psychometric theory provides alternative scoring methods that address curvilinearity. Confirmatory factor analysis (CFA) is one modern psychometric approach. In CFA models, the covariance between items is explained with a single common latent variable (in our example, the overall quality of the patient's experience). In CFA scoring, no assumptions are made a priori about the weighting for each question compromising the composite measure. CFA is a data-driven approach that estimates difficulty and discrimination parameters of each item. Figure 1 depicts path diagrams of the four models we implemented.

The CG-CAHPS® includes items that assess the clinical interaction and organizational aspects of care. A score derived from 12 items (Models A and B) is conceptually broader than if subsets of the CG-CAHPS® questions associated with "physicians" or the "practice" are considered. Unlike the standard score derived from Model A, the factor scores extracted from Model B account for the difficulty and discrimination of the various CG-CAHPS® questionnaire items.

Another approach to account for the separate physician-level and practice-level constructs is the bifactor model. In a bifactor model (Model C), a general factor (overall quality of the patient's experience) is posited that explains covariance between all of the items, and secondary factors capture covariation across selected groups of items (questions

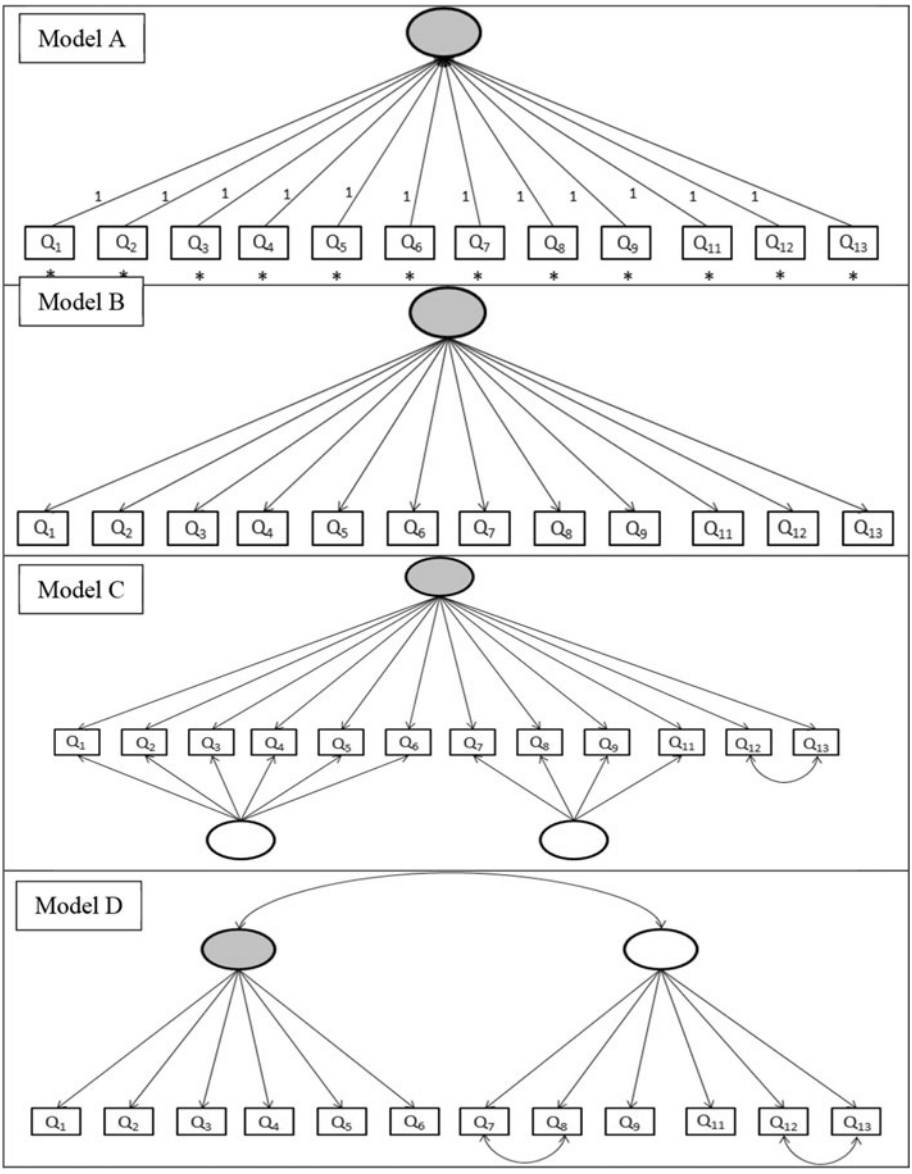


Fig. 1 Item response theory/confirmatory factor analysis models used to model the CG-CAHPS®. Note: Model A = 12 item standard/total score model, Model B = 12 item single factor model, Model C = 12 item bifactor model, Model D = physician score from 12 item correlated factor model. The standard score model can be thought of as a factor model where **a** each item has the same loading (*1*) and **b** thresholds (*asterix*) are identical at fixed values for all items. All loadings and thresholds were freely estimated for Models B, C and D. *Gray ellipses* indicate the factor scores that were extracted for analyses summarized above. Item Q_{10} was dropped from the analyses because of low coverage

primarily associated with the physician and questions primarily associated with the practice) independent of the covariation attributable to the general factor. Patient experience surveys like the CG-CAHPS® have been found to have better fit statistics for a bifactor

model(s) due to their underlying multidimensional construct (Reise et al. 2007; Rodriguez and Crane 2011).

Another way to model the core CG-CAHPS® items is to use a correlated factor model (Model D). Model D suggests that CG-CAHPS® indicators can be conceptualized as measuring two distinct but correlated underlying factors—a physician factor and a practice factor. Two secondary residual correlations capture additional covariation between pairs of indicators beyond their relationship with the overall practice factor.

Our interest was in ranking physicians, so we extracted physician factor scores (designated with a gray ellipse in Fig. 1) from each of the three psychometric models. To our knowledge, no previous research has compared standard scoring and modern psychometric scoring methods for evaluating the performance of individual physicians on patient experience measures. Here we compare the reliability of physician performance comparisons on patient experience measures (Hays et al. 2003b; Holmboe et al. 2010; Safran et al. 2006a; Sequist et al. 2010) under modern psychometric scoring methods and standard scoring methods. In addition, we examine the extent to which the various scoring methods result in similar or divergent rankings of individual physicians across the performance continuum.

2 Methods

The study analyzed 2008 CG-CAHPS® survey data from a primary care quality improvement project focused on improving patients experiences involving eight southern California physician organizations (6 independent practice associations and 2 integrated medical groups). Random samples of approximately 75 commercially insured patients per physician were mailed patient experience surveys. Patients were eligible if they had at least one visit with their primary care physician (named in the survey) during the 12 months prior to the date the survey was fielded and administrative data from the physician organization indicated that they were established members of one of the organization's primary care physicians. The patient survey administration achieved a 39 % response rate and includes 12,244 unique patients of 448 primary care physicians (average patients per physician = 27.3, SD = 11.0). Patients in the analytic sample all self-reported having an established relationship with the primary care physician named in the survey and endorsed having had at least one visit with the physician during the prior 12 months. The survey was fielded in English and included the core CG-CAHPS® composite measures: physician communication (6 items), access to care (5 items), and office staff interactions (2 items). The CG-CAHPS® survey included twelve month reference periods for all questions and all questions are experience-based reports. All core CG-CAHPS® questions employed the 6-point response option version that includes the following categories: "Always," "Almost Always," "Usually," "Sometimes," "Almost Never," and "Never." Responses were scored with values ranging from 0 to 5, where "Never" = 0 and "Always" = 5. Item descriptions appear in Table 3. Previous analyses highlight the absence of differential item functioning (DIF) from multiple sources for the CG-CAHPS® measures (Rodriguez and Crane 2011).

3 Analyses

We calculated composite scores using four different scoring approaches. Model A is the standard scoring approach; Models B-D are modern psychometric approaches to scoring. The item (Q_{10}) —…when you called this doctor's office after regular office hours, how often did you get the medical help or advice you needed?—was dropped from all models

when calculating the composite measures because the item had a high proportion of missing and "not applicable" responses (65.9 %). Other items had little missing data. We performed listwise deletion to facilitate head to head comparison among the different scoring techniques. The models for generating the patient-level composite scores are:

3.1 Model A

Standard scoring of the 12 CG-CAHPS® questions (communication, access to care, office staff interactions), calculating an average score for each patient.

The three modern psychometric models (Fig. 1, models B-D) included:

3.2 Model B

A single factor score using all 12 CG-CAHPS® questions for each patient.

3.3 Model C

A bifactor model (Reise et al. 2007) with the six physician communication questions modeled as one of the secondary factors; the four 'access to care' and two 'office staff interaction' questions modeled as the other secondary factor and a residual correlation respectively to generate a score for each patient. We considered several related bifactor models and found that Model C presented in the figure to have the best fit statistics of the bifactor models. Factor scores for the different, but related, models were extremely highly correlated, so we chose to use the best fitting of the bifactor models for the analyses.

3.4 Model D

A correlated factor model where the two factors are "physicians" and "practice". Two 'access to care' and two 'office staff interaction' questions were modeled as residual correlations under the "practice" factor. We considered several similar correlated factor models, but found Model D to have the best fit statistics of similar models examined. Similar to the bifactor model selection, we selected the best fitting of the correlated factor models for the analyses. It should be noted that only fit criteria—not associations with outcomes—were used to select the best of the best bifactor (Model C) and correlated factor (Model D) models.

We extracted factor scores for physicians from each model for use in subsequent analyses. All four scores were transformed to z-scores for ease of comparison. We case-mix adjusted each score for patient characteristics, as is necessary when comparing CG-CAHPS® measures across organizations or conducting individual physicians comparisons (Zaslavsky et al. 2001) because some patient factors that are associated with CG-CAHPS® scores are not amenable to intervention by health delivery organizations or physicians. The CG-CAHPS® instructions (Dyer et al. 2012) for analyzing the data recommend that users calculate case-mix adjusted scores for each of the survey composite measures (communication, access, office staff interactions) using patient age (18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and 85 years or older), education (< 8th grade, some high school, high school graduate, some college, college graduate, and graduate school), and self-reported health (excellent, very good, good, fair, and poor). We used these same case-mix adjusters. Although not specifically delineated as a case-mix adjuster in the CG-CAHPS® guidance, considerable evidence indicates that patient race/ethnicity meets the criteria for

case-mix adjusters (Eselius et al. 2008; Goldstein et al. 2010; Johnson et al. 2010; O'Malley et al. 2005). As a result, we also adjust for patient race and ethnicity (Hispanic, non-Hispanic White, African-American, and Asian). Ethnicity was non-missing for 90.5 % of the sample. Respondents with missing ethnicity information were omitted from the analyses.

Next, we compared the four composite scores on two important criteria for measures used in ongoing physicians' performance assessment and improvement activities. First, we estimated the physician-level reliability using one-way analysis of variance models of each of the four scores. Physician-level reliability (range 0.1–1.0) indicates the proportion of the variance in physician-level scores attributable to true differences between physician performance (as opposed to within-physician sampling error (Lyratzopoulos et al. 2011). We also calculated the sample sizes required to achieve adequate ($\alpha_{MD} = 0.70$) and good ($\alpha_{MD} = 0.85$) physician-level reliability using the intra-physician correlation and the Spearman–Brown prophecy formula (McHorney et al. 1994):

$$\text{Reliability} = (n \times ICC)/\{1 + (n-1) \times ICC\}, \quad (1)$$

where

$$\text{ICC} = \text{Variance between physicians}/\text{Variance between and within physicians in patient-level scores} \quad (2)$$

Second, we used Kendall's τ correlation coefficient to examine the consistency of physician rankings based on each of the four scoring methods. A Kendall's τ of 1 would indicate that physicians were ranked in exactly the same order for two scoring methods being compared. A Kendall's τ correlation coefficient of 0.80 would mean that 10 % ((1–0.8)/2) of all possible physician pairings have different orders for the two scoring approaches being compared. We examined Kendall's τ to see how well each version approximates the most efficient scoring method, as determined from our reliability and sample size criteria. Finally, we plotted the physician rankings for each method to clarify the extent to which concordance of physician rankings was similar or different across the performance continuum.

4 Results

Overall, 62 % of the commercially-insured survey respondents were female, 13 % reported fair or poor self-rated health, 5 % did not complete high school, 31 % were Latino, and 4 % were African-American. The factor loadings for the psychometric models (Models B–D) are detailed in Table 4. The CFA analyses indicate that the single factor (Model B) model fit was somewhat inconsistent across fit indices, with acceptable comparative fit index (CFI) = 0.92 and Tucker–Lewis index (TLI) = 0.96, but a poor root mean square error of approximation (RMSEA) = 0.23. CFI and TLI around 0.95 and RMSEA around 0.05 are generally considered to indicate very good fit (Reeve et al. 2007). In contrast, the bifactor model (Model C) had better fit across indices, with CFI = 0.98, TLI = 0.99, and RMSEA = 0.09. The correlated-factor model (Model D) had the best overall fit with CFI = 0.99, TLI = 0.99, RMSEA = 0.06. The resulting factor scores were highly correlated with the standard score (Pearson correlation ranged between 0.90 and 0.93).

Adequate physician-level reliability ($\alpha_{MD} = 0.70$) was achieved with 12 patients per physician when the items were scored using the bifactor model (Table 1; Model C), fewer than the 16 patients required when using the standard total score (Table 1; Model A). To

achieve a higher standard of reliability ($\alpha_{MD} = 0.85$), 30 patients per physician were required using the bifactor model (Model C), while 39 responses required to achieve the same level of physician-level reliability using the standard scoring approach (Model A). The results indicate that bifactor scoring reduces sample size requirements for physician performance comparisons by approximately 25 % compared to the standard scoring method.

The Kendall τ findings indicate that the psychometric scoring methods (Table 2; Models B-D) result in somewhat different ranking of individual physicians compared to standard scoring method (Table 2; Model A). The single factor CFA model (Model B) resulted in physician rankings that were most consistent with scores generated using standard scoring approaches (Kendall's $\tau = 0.86$). Physician rankings were moderately-to-highly correlated across the four scoring methods, with the exception for the adjusted bifactor score (Model C) and the adjusted correlated factor score (Model D), where physician rankings diverged (Kendall's $\tau = 0.58$). Figure 2 depicts the consistency and divergence in physician rankings by scoring method. Physician rankings generated by the different models are more consistent at the two extremes of performance (for example, the top and bottom 5 % of the 448 physicians). Physician rankings across the scoring methods are highly divergent in the middle of the performance continuum and most consistent on the extremes of the performance continuum.

5 Discussion

As the CG-CAHPS® is increasingly used in high stakes initiatives like public reporting and pay for performance programs, it is important to employ data collection protocols and analytic methods that enable valid and reliable performance measurement and comparisons of individual physicians and practices. It is also important that the costs and respondent burden associated with such efforts not exceed what is needed for valid and reliable measurement. Our study compared four conceptually-driven scoring approaches for scoring patient experience measures, including adjusted scores generated using the standard CG-CAHPS® guidance (Model A) and the bifactor model (Model C), which has been found to have better model fit than single factor CFA models in previous studies (Reise et al. 2007) and in our analyses.

We found that the bifactor model appears to have important efficiency advantages for comparing physician performance on patient experience measures, equivalent to a 25 % reduction in required sample size at any specified level of reliability for measuring physician performance using CG-CAHPS®. In absence of an external validation criterion, improvements in reliability and sample size reduction could be helpful in choosing one scoring technique over another since a more reliable scale leads to a smaller sample size

Table 1 Intraclass correlations (ICCs) and sample size requirements, by scoring method

Scores from models	ICC estimate (95 % confidence interval)	Estimated sample size requirement	
		$\alpha_{MD} = 0.7$	$\alpha_{MD} = 0.85$
A	0.128 (0.11–0.15)	16	39
B	0.113 (0.10–0.13)	18	44
C	**0.160 (0.14–0.18)**	**12**	**30**
D	0.100 (0.08–0.12)	21	51

ICCs are estimated using an average sample of 27 patients per physician. Bold denotes the superior scoring approach based on the ICC criterion. Score A = 12 item total score; Score B = 12 item single factor score; Score C = 12 item bifactor score; Score D = physician score from 12 item correlated factor score

Table 2 Kendall correlation and pearson correlation of adjusted composite scores

	Score A	Score B	Score C	Score D
Score A	1.00			
Score B	$\tau = 0.86$ $\rho = 0.93$	1.00		
Score C	$\tau = 0.79$ $\rho = 0.91$	$\tau = 0.75$ $\rho = 0.92$	1.00	
Score D	$\tau = 0.74$ $\rho = 0.90$	$\tau = 0.82$ $\rho = 0.97$	$\tau = 0.58$ $\rho = 0.82$	1.00

Note: Kendall correlation (τ) was computed using the case-mix adjusted scores. The Pearson correlation (ρ) was calculated from patient-level scores obtained from the four models. A high τ indicates that the two scoring methods result in a very similar ranking of individual physicians. A lower τ indicates that the two scoring methods result in different rankings of individual physicians. Score A = 12 item total score; Score B = 12 item single factor score; Score C = 12 item bifactor score; Score D = physician score from 12 item correlated factor score

needed for analysis resulting in cost savings. To clarify the efficiency benefits of individual physician comparisons on the CG-CAHPS®, future studies should replicate our approach.

Our assessment of the impact of using different scoring methods on individual physician rankings underscores that the individual rankings on the CG-CAHPS® are somewhat sensitive to the choice of scoring method. Physician rankings were moderately to strongly correlated across scoring methods, with the exception of the bifactor model and the correlated factor models. Rankings were especially divergent in the middle of the performance continuum. Our results indicate that modeling decisions can affect the ranking of individual physician and this suggests that the differences in physician ranking are large enough to have significant impact on the allocation of performance-based incentive payments (Roland et al. 2009) among physicians, especially in the middle of the performance continuum. In the interest of equitable performance of individual physicians on patient experience measures, more research assessing the impact of scoring and case-mix adjustment on individual profiling is warranted.

Our results should be considered in light of important limitations. Our sample consisted of commercially-insured patients who tend to be of higher socioeconomic status and better health compared to other patients. The effect of modern psychometric scoring on the reliability of physician performance using the CG-CAHPS® may not generalize to patients with public insurance or to patients without health insurance. It will be important to compare standard scores and psychometrically sophisticated scores along similar criteria in samples of safety net patients, older adults, and non-English speaking patients (Setodji et al. 2011) to assess the robustness of our findings. In addition, we were not able to examine physician characteristics with the data available to us. As a result, we could not examine the role of physician-patient sex or race concordance. Female physicians, for example, tend to care for higher proportions of female patients and the modern psychometric scoring methods might have different effects on performance comparisons for female versus male physicians. Given that female patients tend to gravitate toward female primary care physicians for their care (Fang et al. 2004), it will be important to examine the differential effects of scoring methods, based on physician characteristics, including physician sex, in future research. The survey had a 39 % response rate and it is possible that this is due to differential non-response by physicians which in turn can lead to biased findings. We are unable to assess non-response bias with the data we have, although given previous research on the impact of patient non-response on performance comparison (Elliott et al. 2009), it is not likely that relationships between scoring approaches would differ based on the response rate.

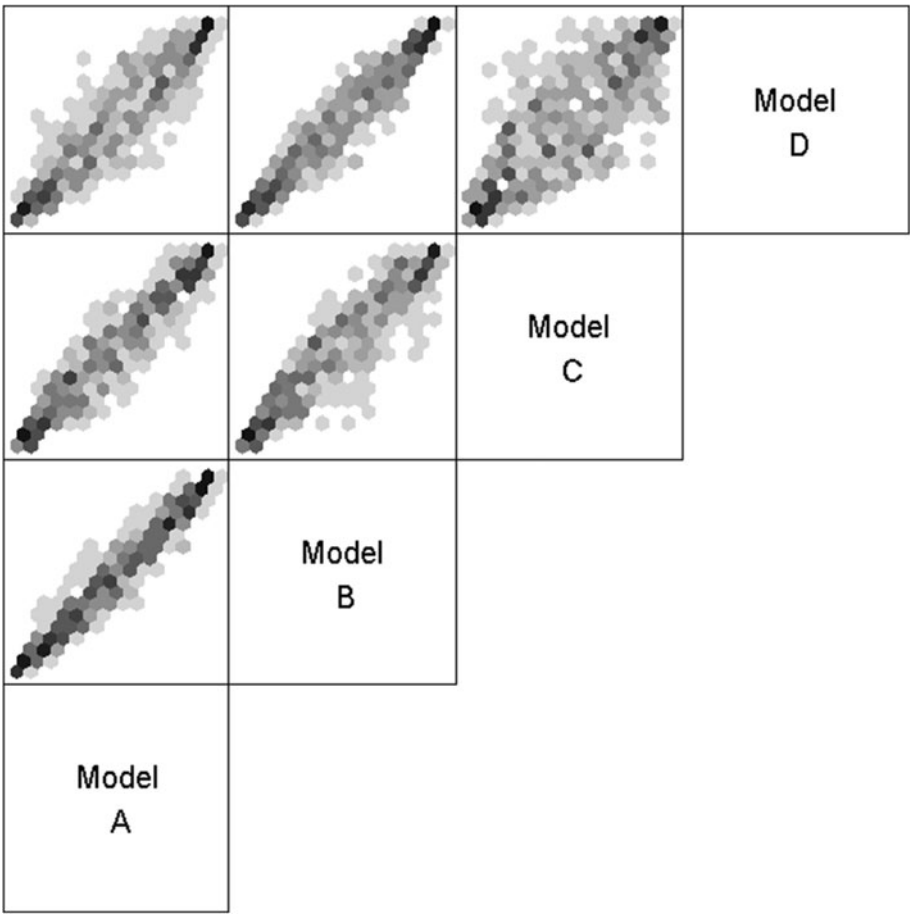


Fig. 2 Comparison of individual physician rankings on CG-CAHPS®, by scoring method. The scatterplot matrix shows the correlation between the rankings generated by using the four different models; A = 12 item total score model; B = 12 item single factor score model; C = 12 item bifactor score; D = physician score from 12 item correlated factor score. The hexagonal binning plot divides each screen on a hexagonal grid and shows the density of points falling in each hexagon (*darker* more agreement and *whiter* less agreement). The *plots* show that the four different scores agree more at the two extremes but the agreement diminishes as we move away from the two ends. The figure was produced in R using the *hexbin* package. Code for generating the plot can be obtained from the authors on request

In conclusion, our comparison of standard scoring and modern psychometric scoring methods for patient experience measures suggests that the bifactor model can improve the precision of individual physician performance comparisons based on ambulatory care experience measures. Model fit is only one criterion by which one might choose a scoring approach. In this instance, the bifactor model fit nearly as well, and smaller patient sample sizes per physician are needed to reliably discern differences among individual physicians. This is an important advantage over the correlated factor model, which had better model fit. Many health care delivery system stakeholders would consider a 25 % reduction in sample size requirements to be a meaningful efficiency gain for patient experience data collection efforts. The complexity of psychometric scoring methods (used to implement the bifactor model) and perceived difficulty of implementing and communicating the scoring

approach, however, may be less appealing to some stakeholders. We provide detailed information in our appendices to aid future research and the implementation of scores for quality improvement initiatives (Appendix D). Because physician scores and rankings are sensitive to modeling decisions, future research should compare the predictive validity of standard and modern psychometric scoring methods for the CG-CAHPS® core questions, including the relative association of scores with important and consequential outcomes of patient care (Hickson et al. 2007).

Appendix 1

See Table 3.

Table 3 Clinician & group CAHPS® questions

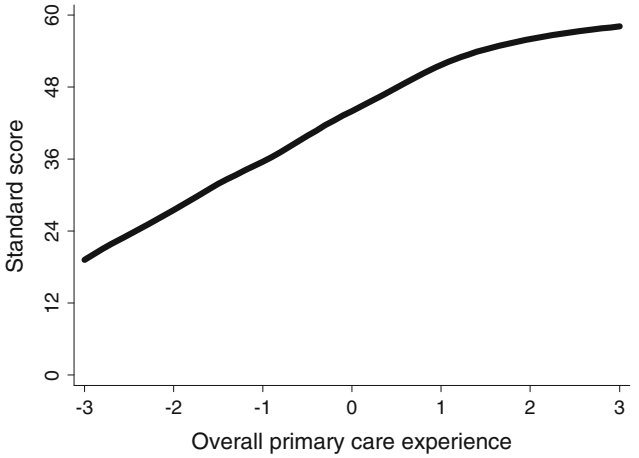
Item	Question	CG CAHPS composite
Q1	…how often did this doctor *explain things* in a way that was easy to understand?	Communication
Q2	…how often did this doctor *listen carefully* to you?	Communication
Q3	…how often did this doctor give you easy-to-understand instructions about *what to do to take care of the health problems or concerns* that were bothering you?[a]	Communication
Q4	…how often did this doctor seem to know the important information about your medical history?	Communication
Q5	…how often did this doctor spend enough time with you?	Communication
Q6	…how often did this doctor show respect for what you had to say?	Communication
Q7	…when you called this doctor's office to get an appointment for care *you needed right away*, how often did you get an appointment as soon as you thought you needed it?[a]	Access to care
Q8	…when you made an appointment for a *check-up or routine care* with this doctor, how often did you get an appointment as soon as you thought you needed it?[a]	Access to care
Q9	…when you called this doctor's office with a medical question *during regular office hours*, how often did you get an answer to your question that same day?[a]	Access to care
Q10	…when you called this doctor's office *after regular office hours*, how often did you get the medical help or advice you needed?	Access to care
Q11	Wait time includes times spent in the waiting room and exam room. In the last 12 months, how often did your visits at this doctor's office start within 15 minutes of your appointment?	Access to care
Q12	…how often were clerks and receptionists at this doctor's office as helpful as you thought they should be?	Office staff
Q13	…how often did clerks and receptionists at this doctor's office treat you with courtesy and respect?	Office staff

The 6-point response option version was used that includes the following categories: "Always", "Almost Always", "Usually", "Sometimes", "Almost Never", and "Never

[a] Denotes that the question had a screener question (Yes/No) to exclude respondents that were not qualified to respond to the question

Appendix 2: Test characteristic curve (TCC) based on the 12 CAHPS® survey items

The *test characteristic curve* (TCC) is a plot of the most likely score associated with each level of cognitive functioning. It is useful for assessing whether the relationship between standard scores and the underlying level of cognitive functioning is linear. One can obtain the test characteristic curve by evaluating the probability of a given response at each ability level for all the items in the test using a given item characteristic curve model. Once these probabilities are obtained, they are summed at each ability level to produce the TCC.



The CG-CAHPS® questionnaire produced a non-linear test characteristic curve, with steeper slopes at the lower end and shallow slopes at upper end. Curvilinear scaling metrics make the use of traditional scores somewhat problematic, as they suggest a non-linear relationship between total scores and the underlying construct measured by the test. Regression and change score analyses assume that the scaling metric is linear—that is, a change of a few points at the top end of the scale has the same implication as a change of the same few points at the bottom end of the scale. The linear assumption was not met here (Crane et al. 2008).

Appendix 3

See Table 4.

Table 4 Factor loadings for the different psychometric models

	Model B	Model C		Model D	
	Loading on primary factor	Loading on primary factor	Loading on secondary factor or residual correlation	Loading on primary factor	Loading on secondary factor or residual correlation
Q1	0.92	0.69	0.63	0.93	
Q2	0.95	0.69	0.67	0.96	
Q3	0.92	0.71	0.61	0.93	
Q4	0.87	0.69	0.56	0.89	
Q5	0.91	0.72	0.58	0.93	
Q6	0.93	0.69	0.64	0.94	
Q7	0.82	0.76	0.52	0.76	0.63[a]
Q8	0.83	0.78	0.50	0.79	
Q9	0.55	0.62	0.08	0.63	
Q11	0.71	0.84	−0.01	0.83	
Q12	0.83	0.77	0.71[a]	0.77	0.71[a]
Q13	0.82	0.78		0.78	

Model A = 12 item standard/total score model; Model B = 12 item single factor model; Model C = 12 item bifactor model; Model D = physician score from 12 item correlated factor model (the correlation between the physician and clinic factors was 0.75)

[a] Indicate residual correlation

Appendix 4: Code used to generate the factor scores from the confirmation factor analysis models:

We ran Mplus models from inside Stata statistical software using the runmplus.ado script written by one of our colleagues Dr. Rich Jones.

Model B) Estimating single factor score with 12 items

```
runmplus id Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q11 Q12 Q13,///
idvariable(id) categorical (Q1–Q13)///
model(factor by Q1–Q13*; factor @ 1;) output (standardized)///
```

Model C) Estimating bi-factor score with 12 items

```
runmplus id Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q11 Q12 Q13,///
idvariable(id) categorical (Q1–Q13)///
model (factor by Q1–Q13*; factor@1;///
f1 by Q1–Q6*; f1@1;///
f2 by Q7-Q11*; f2@1;///
Q12 with Q13;///
factor with f1-f2@0; f1 with f2@0;) output (standardized)///
```

Model D) Estimating physician score from a correlated factor structure with 12 items

```
runmplus id Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q11 Q12 Q13,///
idvariable(id) categorical (Q1–Q13)///
model(doctor by Q1–Q6;///
clinic by Q7–Q13; Q9 with Q11;///
Q12 with Q13;) output (standardized)///
```

References

Anastario, M.P., Rodriguez, H.P., Gallagher, P.M., Cleary, P.D., Shaller, D., Rogers, W.H., Bogen, K., Safran, D.G.: A randomized trial comparing mail versus in-office distribution of the CAHPS Clinician and Group Survey. Health Serv. Res. **45**(5 Pt 1), 1345–1359 (2010)

Beach, M.C., Price, E.G., Gary, T.L., Robinson, K.A., Gozu, A., Palacio, A., Smarth, C., Jenckes, M.W., Feuerstein, C., Bass, E.B., Powe, N.R., Cooper, L.A.: Cultural competence: a systematic review of health care provider educational interventions. Med. Care **43**(4), 356–373 (2005)

Browne, K., Roseman, D., Shaller, D., Edgman-Levitan, S.: Measuring patient experience as a strategy for improving primary care. Health Aff. (Millwood) **29**(5), 921–925 (2010)

Campbell, S.M., Kontopantelis, E., Reeves, D., Valderas, J.M., Gaehl, E., Small, N., Roland, M.O.: Changes in patient experiences of primary care during health service reforms in England between 2003 and 2007. Ann. Fam. Med. **8**(6), 499–506 (2010)

Crane, P.K., Narasimhalu, K., Gibbons, L.E., Mungas, D.M., Haneuse, S., Larson, E.B., Kuller, L., Hall, K., van Belle, G.: Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J. Clin. Epidemiol. **61**(10), 1018-27 e9. (2008)

Dyer, N., Sorra, J.S., Smith, S.A., Cleary, P.D., Hays, R.D.: Psychometric properties of the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Clinician and Group Adult Visit Survey. Med. Care (50 Suppl), S28–S34 (2012)

Elliott, M.N., Zaslavsky, A.M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M.K., Giordano, L.: Effects of survey mode, patient mix, and nonresponse on CAHPS hospital survey scores. Health Serv. Res. **44**(2 Pt 1), 501–518 (2009)

Eselius, L.L., Cleary, P.D., Zaslavsky, A.M., Huskamp, H.A., Busch, S.H.: Case-mix adjustment of consumer reports about managed behavioral health care and health plans. Health Serv. Res. **43**(6), 2014–2032 (2008)

Fang, M.C., McCarthy, E.P., Singer, D.E.: Are patients more likely to see physicians of the same sex? Recent national trends in primary care medicine. Am. J. Med. **117**(8), 575–581 (2004)

Fanjiang, G., von Glahn, T., Chang, H., Rogers, W.H., Safran, D.G.: Providing patients web-based data to inform physician choice: if you build it, will they come? J. Gen. Intern. Med. **22**(10), 1463–1467 (2007)

Goldstein, E., Elliott, M.N., Lehrman, W.G., Hambarsoomian, K., Giordano, L.A.: Racial/ethnic differences in patients' perceptions of inpatient care using the HCAHPS survey. Med. Care Res. Rev. **67**(1), 74–92 (2010)

Harris-Kojetin, L.D., Fowler Jr, F.J., Brown, J.A., Schnaier, J.A., Sweeny, S.F.: The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Consumer Assessment of Health Plans Study. Med. Care **37**(3 Suppl), MS10–MS21 (1999)

Hays, R.D., Chong, K., Brown, J., Spritzer, K.L., Horne, K.: Patient reports and ratings of individual physicians: an evaluation of the DoctorGuide and Consumer Assessment of Health Plans Study provider-level surveys. Am. J. Med. Qual. **18**(5), 190–196 (2003a)

Hays, R.D., Chong, K., Brown, J., Spritzer, K.L., Horne, K.: Patient reports and ratings of individual physicians: an evaluation of the DoctorGuide and Consumer Assessment of Health Plans Study provider-level surveys. Am. J. Med. Qual. **18**(5), 190–196 (2003b)

Hickson, G.B., Federspiel, C.F., Blackford, J., Pichert, J.W., Gaska, W., Merrigan, M.W., Miller, C.S.: Patient complaints and malpractice risk in a regional healthcare center. South. Med. J. **100**(8), 791–796 (2007)

Holmboe, E.S., Weng, W., Arnold, G.K., Kaplan, S.H., Normand, S.L., Greenfield, S., Hood, S., Lipner, R.S.: The comprehensive care project: measuring physician performance in ambulatory practice. Health Serv. Res. **45**(6 Pt 2), 1912–1933 (2010)

Johnson, M.L., Rodriguez, H.P., Solorio, M.R.: Case-mix adjustment and the comparison of community health center performance on patient experience measures. Health Serv. Res. **45**(3), 670–690 (2010)

Kaplan, S.H., Griffith, J.L., Price, L.L., Pawlson, L.G., Greenfield, S.: Improving the reliability of physician performance assessment: identifying the "physician effect" on quality and creating composite measures. Med. Care **47**(4), 378–387 (2009)

Lyratzopoulos, G., Elliott, M.N., Barbiere, J.M., Staetsky, L., Paddison, C.A., Campbell, J., Roland, M.: How can health care organizations be reliably compared?: lessons from a national survey of patient experience. Med. Care **49**(8), 724–733 (2011)

McHorney, C.A., Ware Jr, J.E., Lu, J.F., Sherbourne, C.D.: The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med. Care **32**(1), 40–66 (1994)

Nelson, E.C., Gentry, M.A., Mook, K.H., Spritzer, K.L., Higgins, J.H., Hays, R.D.: How many patients are needed to provide reliable evaluations of individual clinicians? Med. Care **42**(3), 259–266 (2004)

O'Malley, A.J., Zaslavsky, A.M., Elliott, M.N., Zaborski, L., Cleary, P.D.: Case-mix adjustment of the CAHPS Hospital Survey. Health Serv. Res. **40**(6 Pt 2), 2162–2181 (2005)

Rao, J.K., Anderson, L.A., Inui, T.S., Frankel, R.M.: Communication interventions make a difference in conversations between physicians and patients: a systematic review of the evidence. Med. Care **45**(4), 340–349 (2007)

Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., Thissen, D., Revicki, D.A., Weiss, D.J., Hambleton, R.K., Liu, H., Gershon, R., Reise, S.P., Lai, J.S., Cella, D.: Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med. Care **45**(5 Suppl 1), S22–S31 (2007)

Reise, S.P., Morizot, J., Hays, R.D.: The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual. Life Res. **16**(Suppl 1), 19–31 (2007)

Rodriguez, H.P., Crane, P.K.: Examining multiple sources of differential item functioning on the Clinician & Group CAHPS(R) survey. Health Serv. Res. **46**(6pt1), 1778–1802 (2011)

Rodriguez, H.P., von Glahn, T., Chang, H., Rogers, W.H., Safran, D.G.: Measuring patients' experiences with individual specialist physicians and their practices. Am. J. Med. Qual. **24**(1), 35–44 (2009a)

Rodriguez, H.P., von Glahn, T., Elliott, M.N., Rogers, W.H., Safran, D.G.: The effect of performance-based financial incentives on improving patient care experiences: a statewide evaluation. J. Gen. Intern. Med. **24**(12), 1281–1288 (2009b)

Rodriguez, H.P., von Glahn, T., Li, A., Rogers, W.H., Safran, D.G.: The effect of item screeners on the quality of patient survey data: a randomized experiment of ambulatory care experience measures. Patient **2**(2), 135–141 (2009c)

Rodriguez, H.P., von Glahn, T., Rogers, W.H., Chang, H., Fanjiang, G., Safran, D.G.: Evaluating patients' experiences with individual physicians: a randomized trial of mail, internet, and interactive voice response telephone administration of surveys. Med. Care **44**(2), 167–174 (2006)

Roland, M., Elliott, M., Lyratzopoulos, G., Barbiere, J., Parker, R.A., Smith, P., Bower, P., Campbell, J.: Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England. BMJ **339**, b3851 (2009)

Safran, D.G., Karp, M., Coltin, K., Chang, H., Li, A., Ogren, J., Rogers, W.H.: Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. J. Gen. Intern. Med. **21**(1), 13–21 (2006a)

Safran, D.G., Karp, M., Coltin, K., Chang, H., Li, A., Ogren, J., Rogers, W.H.: Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. J. Gen. Intern. Med. **21**(1), 13–21 (2006b)

Schnaier, J.A., Sweeny, S.F., Williams, V.S., Kosiak, B., Lubalin, J.S., Hays, R.D., Harris-Kojetin, L.D.: Special issues addressed in the CAHPS survey of Medicare managed care beneficiaries. Consumer Assessment of Health Plans Study. Med. Care **37**(3 Suppl), MS69–MS78 (1999)

Sequist, T.D., Schneider, E.C., Li, A., Rogers, W.H., Safran, D.G.: Reliability of medical group and physician performance measurement in the primary care setting. Med. Care **49**(2), 126–131 (2010)

Sequist, T.D., von Glahn, T., Li, A., Rogers, W.H., Safran, D.G.: Statewide evaluation of measuring physician delivery of self-management support in chronic disease care. J. Gen. Intern. Med. **24**(8), 939–945 (2009)

Setodji, C.M., Reise, S.P., Morales, L.S., Fongwa, M.N., Hays, R.D.: Differential item functioning by survey language among older Hispanics enrolled in Medicare managed care: a new method for anchor item selection. Med. Care **49**(5), 461–468 (2011)

Zaslavsky, A.M., Zaborski, L.B., Ding, L., Shaul, J.A., Cioffi, M.J., Cleary, P.D.: Adjusting performance measures to ensure equitable plan comparisons. Health Care Financ. Rev. **22**(3), 109–126 (2001)