Provider Attitudes toward Pay-for-Performance Programs: Development and Validation of a Measurement Instrument

Mark Meterko, Gary J. Young, Bert White, Barbara G. Bokhour, James F. Burgess, Jr., Dan Berlowitz, Matthew R. Guldin, and Marjorie Nealon Seibert

---

Objective. To develop an instrument for assessing physician attitudes toward quality incentive programs, and to assess its reliability and validity.

Data Sources. Study involved primary data collection. A 40-item paper and pencil survey of primary care physicians in Rochester, New York, and Massachusetts was conducted between May 2004 and December 2004. Seven-hundred and ninety-eight completed questionnaires were received, representing a response rate of 32 percent (798/2,497).

Study Design. Based on an extensive review of the literature and discussions with experts in the field, we developed a conceptual framework representing the features of pay-for-performance (P4P) programs hypothesized to affect physician behavior in that context. A draft questionnaire was developed based on that conceptual model and pilot tested in three groups of physicians. The questionnaire was modified based on the physician feedback, and the revised version was distributed to 2,497 primary care physicians affiliated with two of the seven sites participating in Rewarding Results, a national evaluation of quality target and financial incentive programs.

Data Collection. Respondents were randomly divided into a derivation and a validation sample. Exploratory factor analysis was applied to the responses of the derivation sample. Those results were used to create scales in the validation sample, and these were then subjected to multitrait analysis (MTA). One scale representing physicians' perception of the impact of P4P on their clinical practice was regressed on the other scales as a test of construct validity.

Principal Findings. Seven constructs were identified and demonstrated substantial convergent and discriminant validity in the MTA: awareness and understanding, clinical relevance, cooperation, unintended consequences, control, financial salience, and impact. Internal consistency reliabilities (Cronbach's α coefficients) ranged from 0.50 to 0.80. A statistically significant 25 percent of the variation in perceived impact was accounted for by physician perceptions of the other six characteristics of P4P programs.

Conclusions. It is possible to identify and measure the key salient features of P4P programs using a valid and reliable 26-item survey. This instrument may now be used in further studies to better understand the impact of P4P programs on physician behavior.

Key Words. Pay-for-performance, financial incentives, physician attitudes, psychometrics

---

During the past 5 years, an increasing number of health plans and self-insured employers have instituted financial incentive programs as a strategy for motivating health care providers to improve quality of care. Under these programs, known as pay-for-performance (P4P), providers receive money for achieving predefined quality targets that usually encompass primary care or preventive services (Young et al. 2005). A recently published paper reports that over 100 quality incentive programs now exist in the United States in both the private and public sector (Baker and Carter 2005).

As the number of P4P programs increases, efforts are underway to evaluate their impact. These evaluations will need to investigate whether and to what degree improvements in the relevant quality measures are attributable to the incentive programs. To understand what features of incentive programs are associated with positive or negative results, it will be important to assess the attitudes of providers, particularly those of the physicians who are the target recipients of most P4P programs. For example, providers' attitudes toward key program features, such as the amount of the incentive or the attainability of the quality targets, may help predict the extent to which they change their clinical behavior. Certain program features, such as the frequency and type of performance feedback, could potentially affect provider awareness and understanding of particular programs. Further, as many P4P programs are in the

---

Address correspondence to Mark Meterko, Ph.D., Chief Survey Methodologist, Center for Organization, Leadership and Management Research, Department of Veterans Affairs, VA Medical Center (152M), 150 South Huntington Avenue, Boston, MA 02130. Mark Meterko, Ph.D., Research Associate, Bert White, M.B.A., D.Min., Project Director, Barbara G. Bokhour, Ph.D., Assistant Professor, James F. Burgess, Jr., Ph.D., Associate Professor, Dan Berlowitz, M.D., Professor, and Matthew R. Guldin, M.P.H., Research Assistant, are with the Boston University School of Public Health, Boston, MA. Gary J. Young, Ph.D., J.D., Associate Director, James F. Burgess, Jr., Ph.D., Research Associate, and Marjorie Nealon Seibert, M.B.A., Survey Administrator, are with the Center for Organization, Leadership and Management Research, Department of Veterans Affairs, Boston, MA. Gary J. Young, Ph.D., J.D., is Professor and Chair, Health Services Department, Boston University School of Public Health, Boston, MA. Barbara G. Bokhour, Ph.D., Research Scientist and Dan Berlowitz, M.D., Director, are with the Center for Health Quality, Outcomes, and Economic Research, Department of Veterans Affairs, Bedford, MA. James F. Burgess, Jr., Ph.D., Economist, is with the Management Science Group, Department of Veterans Affairs, Bedford, MA.

early stages of their life cycle, program managers need a tool for assessing provider attitudes so that they can gauge problem areas early on and make mid-course corrections as needed during implementation. Such advance “intelligence” can be especially important given that a long lag may occur between the time a program is implemented and its actual impacts can be detected from either medical records or administrative data. At present, however, the health care community lacks a valid, reliable, and responsive measurement tool.

In this paper, we describe the development of an instrument for assessing provider attitudes toward quality incentive programs and present evidence regarding its reliability and validity. We developed this instrument as part of our role as the national evaluator for the Rewarding Results program, which consists of seven demonstration projects that are testing different ways to incentivize providers to improve quality of care. The basic characteristics of these demonstrations have been described elsewhere (Young et al. 2005).

CONCEPTUAL MODEL

Based on an extensive review of the literature and discussions with experts in the field, we developed a conceptual model of the factors that may affect the impact of P4P programs (Young et al. 2005). Our literature review focused on research addressing the conditions under which providers have the motivation and capability to change their practice behavior (e.g., Dudley et al. 1998; Conrad and Christianson 2004), specifically studies on provider adoption of clinical/administrative innovation and adherence to clinical guidelines and best-evidence practice. These studies point to the importance of providers’ practice setting, including culture, resource capabilities, and market characteristics (e.g., Shortell et al. 1995; Banaszak-Holl, Zinn, and Mor 1996) as well as experience and demographics (e.g., Tamblyn 2003). Further, the literature on guideline adherence highlights the importance of assessing what physicians perceive as barriers to their ability to adopt guidelines or innovative practices (e.g., Pathman et al. 1996; Cabana et al. 1999). Indeed, P4P programs essentially are initiatives to address what has been cited in this literature as a key perceived barrier to guideline adherence, namely lack of reimbursement or financial incentives (Institute of Medicine 2001).

Our conceptual model consists of three broad domains: characteristics of the incentive program, characteristics of the practice environment, and

Figure 1. Conceptual Framework

provider-level characteristics. Within the provider characteristics domain, we distinguish between provider demographics and provider attitudes. This model is presented in Figure 1.

The goal of the present study was to develop and validate a measure of one of the provider characteristics in our model: provider attitudes. More specifically, based on our literature review we hypothesized seven critical dimensions of provider attitudes related to quality targets and incentives: (1) *awareness and understanding* of the incentive program, (2) *salience* of the financial incentives, (3) *clinical relevance* of the quality targets, (4) *control* over the resources needed to achieve the quality targets, (5) *fairness* in the administration of the incentive program, (6) frequency and nature of *performance feedback* provided, and (7) possible *unintended consequences* associated with the pursuit of the quality targets.

Regarding *awareness and understanding*, providers must have some degree of familiarity with the definitions of the quality targets used to evaluate their performance if they are going to actively participate in the incentive program. Providers' motivation to pursue quality targets may be affected by

their degree of understanding of the criteria and rules for distributing incentive money among participating providers. Assuming awareness and some understanding of the P4P program, we further hypothesize that providers' responses to incentives will be affected by *financial salience*; that is, the amount of the financial award compared with the costs in time and effort necessary to achieve the quality targets.

Regardless of the amount of the incentive involved, we also hypothesize that providers' behavior will depend on their perceptions of several clinical issues related to the quality targets. These include providers' judgments of the *clinical relevance* of the quality targets, including consideration of such issues as whether or not the targets are based on sound medical science, and whether reaching the targets will truly improve the health of their patients. Additionally, providers' estimates regarding the potential for negative *unintended consequences* are likely to be important; that is, whether they believe that their efforts to achieve the quality targets will detract in any way from attending to other important aspects of care.

We also hypothesize that providers' behavior relative to an incentive program will depend in part on whether they believe that they have adequate *control* over the activities and/or resources necessary to achieve the quality targets. If, for example, providers believe that achieving the quality targets depends more on patient behavior than their own efforts, or that they will not be able to secure the cooperation of other physicians or providers involved in the provision of program-required tests or services, then they may be less likely to be fully engaged in the pursuit of the incentives. Additionally, we posit that providers' perceptions of the *fairness* of the incentive program affect their motivation to pursue P4P quality targets. Fairness in this context refers to the appropriateness of the proposed quality measure, including relevant case-mix adjustment considerations. If providers believe that the characteristics of their patients—for example, age, educational attainment, health status, or comorbidities—make it especially difficult to achieve the quality targets, then they might be less inclined to pursue those targets. We also propose that providers' perceptions of the helpfulness of the *feedback* they receive regarding their progress toward achieving program quality targets are important. For example, a program in which providers only received performance feedback once a year, and then only a short time before the annual incentive checks were distributed, might engender a different level of participation than an incentive program that involved monthly or quarterly performance progress reports.

METHODS

Questionnaire Development

With these concepts to guide us, the study team, which included a physician, an economist, a psychologist, a former health plan administrator, and health services researchers, generated a pool of over 50 items to represent the range of content associated with each of the hypothesized dimensions, ensuring that we had at least five items for each. The item pool was constructed in an iterative fashion whereby individual members of the team generated potential items independently. These were subsequently reviewed, modified, and consolidated during team meetings. These items formed the core of the pilot questionnaire.

Because many incentive programs have multiple quality targets, we also included a screening question to focus providers on a specific quality target and its associated financial incentive in their responses to the core questionnaire items. Accordingly, respondents were asked to review a customized list of medical conditions and procedures known to apply to the incentive program available to them, and to select from that list a condition/procedure that was high volume in their practice. Respondents were then instructed to have this condition/procedure in mind when they answered the subsequent core items regarding the impact of quality targets and incentives on their practice behavior. The number of items in this section varied during the development process and pilot testing, but ultimately 26 items appeared in the version that we fielded to obtain data for the psychometric analyses. All of these items utilized a five-point response scale that ranged from strongly disagree (1) to strongly agree (5) with a midpoint labeled "neutral." In addition, we also thought it important to measure providers' perceptions of the overall impact of the quality incentive program. Therefore, we asked respondents to judge the extent to which their practice behavior had changed in response to the quality targets and associated financial rewards.

Pilot Testing and Revision

We conducted pilot studies of the initial, longer versions of the instrument at three medical groups in the greater Boston area that had contracts with health plans offering quality-related financial incentives. One was a large group practice (59 physicians) located in a rural area; the second group was an urban physician primary care practice consisting of 17 providers affiliated with a major health care network; and the third was a suburban medical practice consisting of 19 primary care physicians.

At each group practice, the pilot study involved between 13 and 15 physicians and lasted 40–60 minutes. We followed a standard protocol in which we distributed copies of the questionnaire to physicians at a meeting and asked them to complete the survey, noting any ambiguous or otherwise confusing words or phrases as they proceeded. Once physicians had completed their questionnaires, project staff led a “think-aloud debriefing” (Aday 1989) that involved a page-by-page review of the questionnaire with specific probes for those items that the study team had decided in advance were potentially problematic. Physicians were also asked if they had left any of the questions unanswered on each page and why. Following the debriefing, items identified as problematic were either rephrased or dropped. The resulting revised instrument was then presented to the next physician group until, at the conclusion of the pilot testing, we had reduced the questionnaire to 26 core items with an overall completion time of 15 minutes.

Psychometric Evaluation

Sample. We tested the psychometric properties of the 26-item instrument using data obtained from surveying physicians in two of the seven Rewarding Results demonstration sites. One site consisted of a partnership between Excellus, Inc. and the Rochester Individual Practice Association (RIPA). Excellus, a single health maintenance organization (HMO), has enrolled more than 70 percent of the commercial (non-Medicare and non-Medicaid) population in the nine county region surrounding Rochester, New York. RIPA is a contracting entity comprised of more than 800 primary care physicians who collectively care for more than 420,000 commercial patients, approximately 80 percent of whom are members of Excellus. Excellus and RIPA have collaborated to share administrative data for the purpose of their P4P program. The sample for the present study consisted of physicians who qualified to receive a financial reward from RIPA for achieving quality targets related to care pathways for at least one of four conditions: asthma, diabetes, otitis media, or sinusitis. A total of 597 physicians (152 family practitioners, 290 internists, and 155 pediatricians) met this criterion; all were included in the survey sample in support of RIPA’s desire to obtain information on attitudes toward quality targets and incentives from its physicians.

The second site consisted of Massachusetts-based medical groups whose quality performance is being profiled by the Massachusetts Health Quality Partnership (MHQP). MHQP is a coalition of physicians, hospitals, health plans, purchasers, and government agencies with the broad goal of

improving the quality of health care services in Massachusetts. Coalition members are collaborating on the development and public release of a multipayer report card for 12 different clinical conditions or procedures, including antidepressant medication management, asthma medication use, breast and cervical cancer screening, and diabetes. Each of the five participating insurance plans negotiate independently with physician contracting entities in Massachusetts to establish quality targets and associated financial incentives for these and other medical conditions/procedures. The negotiations are plan-entity specific and outside the purview of the consortium itself. In general, financial incentives are directed to the medical group practice level, not at the level of individual physicians.

The universe of potential study participants in Massachusetts consisted of 5,297 primary care physicians affiliated with one of the 69 entities that contract with one or more of the health plans participating in the MHQP profiling program. Taking into account precision estimates, expected response rate, and design effects from the clustering of physicians within contracting entities and medical practices, we calculated that a sample size of approximately 1,600 physicians would be needed. To achieve this goal, we employed a two-stage probability sampling process. First, we grouped contracting entities into three geographic regions: east, central, and west. Contracting entities within each of these regional strata were further stratified by size (small, medium, large) based on the number of affiliated primary care physicians (less than 20, 20–69, and 70 or more physicians, respectively). We then selected contracting entities randomly from each of the nine region-by-size cells such that the number of physicians selected from each cell was proportionate to their frequency in the sampling frame. A total of 1,978 Massachusetts physicians were selected to participate in the survey.

Survey Administration. Cover letters and questionnaire booklets were distributed in envelopes addressed to each physician by name. Each packet also contained a prepaid business reply envelope that respondents were instructed to use to return their completed surveys directly to a third-party data entry service. In Massachusetts, most surveys were sent to a liaison at each of the 32 selected contracting entities, who then distributed the surveys either at a medical group meeting or by placing the packets into office mailboxes. At RIPA, where most physicians were members of solo or small group practices, questionnaires were distributed by mail or by RIPA managers at group events such as grand rounds at affiliated hospitals.

Analysis Strategy. Our analysis proceeded in two broad phases: identification of scales, followed by the assessment of their psychometric properties. As a preliminary step, we randomly divided the respondents into two groups, a derivation sample and a validation sample. To determine whether any of the questionnaire items formed coherent, relatively independent subsets, we applied exploratory factor analysis (initial extraction of factors using principal components analysis, followed by varimax rotation) to the data from the derivation sample (Tabachnick and Fidell 1983). We used three decision rules—Kaiser’s (1960) “eigenvalue greater than one” criterion, Cattell’s (1966) scree test, and Guertin and Bailey’s (1970) minimum percent of variance accounted for (5 percent) guideline—to decide on the number of factors to extract. We then computed scale scores by averaging those items with factor loading coefficients greater than or equal to 0.40 on the factor in question; all items were weighted equally. Three items (Q20, Q26, Q33) with loadings greater than or equal to 0.40 on two different factors were assigned to the factor with the highest loading.

To test the stability of the scales identified in the derivation sample, we applied multitrait analysis (MTA) to the data in the validation sample. MTA, based on the multitrait/multimethod strategy originally described by Campbell and Fiske (1959), assesses both the reliability and validity of proposed multiitem scales. The latter is accomplished by comparing the strength of the correlation of each item with its assigned scale (convergent validity) as compared with its correlation with all other proposed scales (discriminant validity; Hays and Hayashi 1990).

Additionally, we assessed the construct validity of the attitudinal measures. Specifically, our model posits that the impact of P4P programs on provider behavior will be related to the provider attitudes previously described. We examined this proposed relationship—a strategy for construct validation (American Educational Research Association 1985)—by estimating a regression model in which the attitudinal measures were used to predict the perceived impact of quality targets and incentives on clinical practice behavior.

Finally, because the instrument potentially may be used to compare medical groups or other groups of providers regarding attitudes toward quality incentive programs, we estimated the reliability of our final scales at the medical group level. We conducted a one-way analysis of variance (ANOVA) using medical group as the between factor, and then computed an intraclass correlation as $(MS_{\text{between}} - MS_{\text{within}})/MS_{\text{between}}$ (Hays et al. 1999). The number of responses needed to obtain adequate reliability for group

comparisons, defined as 0.70 (Nunnally 1978), was estimated using the Spearman–Brown prophecy formula (Solomon et al. 2005).

RESULTS

We obtained a total of 798 completed questionnaires from physicians. After accounting for physicians who could not be contacted ($n = 78$), this represented a 32 percent overall response rate: 43 and 29 percent for RIPA and Massachusetts physicians, respectively. The percent of missing responses was very low, averaging 2.3 percent (median 1.9 percent) with an observed maximum of 5.4 percent after taking into account legitimate missing data because of skip instructions.

Table 1 presents background characteristics of the respondents (self-report from the questionnaire). Of the 798 respondents, 632 (79 percent) answered the screener question affirmatively by selecting a particular clinical condition/procedure from those listed as the focus for their responses to the core survey items. These respondents were then randomly divided into the derivation sample ($n = 316$) and validation sample ($n = 316$). The success of the randomization was checked by comparing these two groups on the four background characteristics presented in Table 1; no significant differences were found.

Table 1:   Respondent Demographics (Percentages)

<i>Demographic Characteristic</i>	<i>Sample</i>		
	<i>MA (n = 552)</i>	<i>RIPA (n = 246)</i>	<i>Total (n = 798)</i>
Years since completing residency			
10 years or less (%)	31.7	25.9	29.9
11–20 years	29.5	34.6	31.1
21 or more years	38.8	39.5	39.0
Physician type			
Internist	55.8	41.7	51.3
Family medicine	20.0	26.7	22.1
Pediatrician	24.2	31.6	26.6
Faculty appointment			
Yes	48.9	67.6	54.6
N of current active patients			
1,500 or less	39.8	23.6	34.8
1,501–2,500	36.4	48.1	40.0
Over 2,500	23.8	28.3	25.2

Regarding the exploratory factor analysis, six factors accounting for 57 percent of item variance were retained. This rotated simple structure generally supported the proposed key dimensions with the exception of fairness, which did not appear to be represented by an independent factor. The item-to-scale assignments for these six factors were used to create the summated rating scales for the validation sample. We then conducted a series of MTAs. As recommended by Ware et al. (1997) for the development of new scales, these analyses were performed using the subset of 296 respondents (93 percent) with complete data on all 26 of the core survey items. After each round of MTA, items were dropped from a given scale if doing so improved the reliability of that scale, or if the item demonstrated a higher correlation with another scale. After two iterations of MTA, no further improvements in scale reliabilities could be achieved, and five items had been identified as potential drops. Two of those items (Q21, Q29) were originally composed to represent financial salience (Q21, Q29), two were fairness items (Q25, Q44) and one was a control item (Q33). To test the viability of retaining scales for salience and fairness, the MTA was repeated again including two-item scales for those dimensions. Both scales demonstrated internal consistency reliabilities around 0.50. However, one of the fairness items (Q44) failed the discriminant validity criterion by being significantly more highly correlated with three other scales. Neither of the salience items exhibited significantly higher correlations with other scales, and we therefore elected to include the salience scale in the final model.

Table 2 presents the final results of the MTA. Seven dimensions were represented: clinical relevance, awareness and understanding, cooperation, concern for unintended consequences, control, financial salience, and impact on clinical behavior. The pattern of convergent and divergent correlations constitutes strong evidence of the reliability and validity of these scales. The item-to-scale correlations were substantial in magnitude, ranging from 0.34 to 0.73 across the seven proposed dimensions (median 0.54). Correlations of 0.40 or higher between an item and its overall scale score (adjusted for overlap), conventionally regarded as indicative of adequate item internal consistency (Kerlinger 1973; Ware et al. 1997), were observed for 83 percent of all items. Further, the correlations between items and their hypothesized scales were significantly higher than correlations with any other scale in 121 out of 138 comparisons (88 percent), and were higher, though not significantly, in an additional 11 comparisons. Thus, appropriate discriminant validity was observed for 96 percent of item-to-scale correlations.

Table 3 reports basic descriptive statistics for the seven final attitude scales. The unintended consequences items were reverse-coded so that a

Table 2: Multitrait Analysis Results: Item Descriptive Statistics and Item-to-Scale Correlations

Scale/Items	Mean	SD	Impact	Relevance	Aware	Coop.	UC	Control	Salience
Impact									
Q36	2.54	1.08	0.41	0.09	0.12	0.15	-0.13	-0.09	0.05
Q39	2.45	0.99	0.49	0.14	0.32	0.27	-0.06	0.13	0.2
Q41	2.74	1.07	0.59	0.47	0.36	0.44	0.28	0.30	0.3
Q32	2.00	0.96	0.35	0.02	0.13	0.12	0.07	0.16	0.21
Q37	2.03	0.84	0.36	0.21	0.22	0.18	-0.00	0.14	0.35
Q45	2.72	0.93	0.47	0.51	0.34	0.39	0.36	0.44	0.38
Clinical relevance									
Q38	4.04	0.82	0.25	0.55	0.25	0.21	0.46	0.27	0.09
Q30	3.49	0.98	0.38	0.73	0.43	0.32	0.44	0.42	0.22
Q22	3.64	1.04	0.32	0.65	0.40	0.30	0.46	0.44	0.21
Awareness and understanding									
Q27	2.61	1.06	0.20	0.22	0.62	0.30	0.17	0.33	0.26
Q28	2.18	1.04	0.23	0.14	0.45	0.19	0.12	0.23	0.36
Q40	2.66	0.91	0.40	0.39	0.54	0.45	0.31	0.43	0.37
Q35	2.88	1.09	0.40	0.30	0.53	0.31	0.17	0.39	0.38
Q20	3.73	1.07	0.16	0.50	0.43	0.27	0.36	0.35	0.18

Cooperation	Q42 I am able to get the cooperation of other physicians as needed to obtain this financial incentive	2.80	0.79	0.39	0.27	0.33	0.64	0.18	0.32	0.29
		3.12	0.88	0.34	0.33	0.43	0.64	0.23	0.33	0.28
	Unintended consequences (UC)									
Q31	The effort required to obtain this financial incentive has an adverse impact on other types of patients in my practice	3.72	0.87	0.10	0.45	0.31	0.21	0.67	0.41	0.22
	Efforts to obtain this financial incentive hinder me from providing other essential medical services to this group of patients	3.86	0.90	0.15	0.53	0.27	0.21	0.67	0.40	0.19
	Control									
Q34	Physicians are on a level playing field for obtaining this incentive	2.93	1.07	0.23	0.37	0.37	0.34	0.37	0.49	0.26
	The actions necessary to obtain this financial incentive are largely within my control	3.08	1.14	0.21	0.37	0.38	0.24	0.33	0.49	0.34
	The data used to assess achievement of the quality target are accurate	2.68	1.13	0.20	0.34	0.40	0.28	0.35	0.54	0.33
Financial salience	Q21 This financial incentive represents an opportunity for me to appreciably increase my income	2.23	1.03	0.34	0.14	0.35	0.23	0.1	0.25	0.34
	Q29 This financial incentive is sufficiently large to compensate for expenditures that might be necessary in order to meet the quality target	2.46	0.94	0.27	0.2	0.36	0.28	0.28	0.4	0.34

Boldface entries are item-to-hypothesized scale (convergent validity) Pearson correlation coefficients.

Table 3: Physician Attitude Scale Descriptive Statistics

<i>Scale (k)</i>	<i>Mean</i>	<i>SD</i>	<i>Percent of Respondents</i>	
			<i>At Floor</i>	<i>At Ceiling</i>
Impact (6)	2.39	0.66	1.7	0.0
Relevance (3)	3.69	0.82	0.3	8.1
Awareness (3)	2.81	0.74	0.0	0.0
Cooperation (2)	2.97	0.75	4.7	0.7
Unintended consequences (2)	3.77	0.84	0.7	13.9
Control (3)	2.89	0.86	3.0	1.0
Financial salience (2)	2.35	0.81	11.2	0.0

higher score would indicate more favorable attitudes toward quality targets and incentives on all scales. Table 4 reports correlations among the scales (off-diagonal entries) and the scale internal reliability consistency estimates (diagonal entries). The pattern of relationships observed here are consistent with what we would expect for valid measures of related yet distinct attitudes toward different aspects of quality targets and incentives. Specifically, the correlations range from 0.03 to 0.44 with a median of 0.35. However, the α coefficients (diagonal entries) are considerably higher than the interscale correlations (off-diagonal entries), which adds additional support for the discriminant validity of the scales. Additionally, we repeated the MTA separately for the RIPA and Massachusetts physicians to check the robustness of the scale structure. The results were very similar in both groups and supportive of the seven-scale model.

To assess construct validity, we conducted an ordinary least squares (OLS) regression analysis of the relationship between the provider attitude measures and the perceived impact of incentives. The analysis was run in a

Table 4: Correlations among Physician Attitudes toward Incentives Scales (Internal Consistency Reliability Estimates in Diagonal)

<i>Scale</i>	<i>Impact</i>	<i>Relevance</i>	<i>Aware</i>	<i>Coop.</i>	<i>UC</i>	<i>Control</i>	<i>Salience</i>
Impact	(0.71)						
Relevance	0.29	(0.80)					
Awareness	0.35	0.41	(0.75)				
Cooperation	0.36	0.33	0.40	(0.78)			
Unintended consequences	0.03	0.44	0.23	0.20	(0.80)		
Control	0.31	0.43	0.44	0.35	0.38	(0.69)	
Financial salience	0.38	0.28	0.43	0.34	0.10	0.38	(0.50)

stepwise manner. Six control variables were forced into the equation first: years since residency, number of active patients, and dummy variables for medical school faculty appointment and medical specialty (family practice, pediatrics, or general internal medicine; the latter was the hold-out group). The sixth control variable was a single-item measure of a provider's overall satisfaction with his/her practice that was included to account for any general positive or negative bias that might be associated with respondents' current general career situation. Collectively, the control variables accounted for just over 2 percent of the variance in perceived impact.

The attitude measures were then allowed into the equation using a forward selection process. Financial salience accounted for about 13 percent of the remaining variance in perceived impact, followed by cooperation (5 percent) and relevance (2 percent). Unintended consequences, control, and awareness each accounted for about 1 percent of the remaining variance. Because most of the physicians in the study sample were nested or clustered within medical group practices and thus statistically not entirely independent observations, we repeated the analysis using random effects models that incorporated the clustering of physicians into medical groups. The results from the random-effects analysis were not significantly different from those obtained from OLS.

Finally, we computed the observed group-level reliabilities for each scale as well as the number of providers needed to achieve a group reliability of 0.70. Group-level reliabilities ranged from 0.29 (financial salience) to 0.73 (awareness and understanding) with a median of 0.60. The number of respondents per medical group needed to achieve a group-level reliability of 0.70 correspondingly ranged from 22.3 (awareness and understanding) to 148.4 (financial salience), with the larger samples required by those scales for which the observed variability between groups was lower relative to the variability within groups.

DISCUSSION

In this paper, we reported results from our efforts to develop and test a self-report measure of providers' attitudes toward the key features of P4P programs. To examine the reliability and validity of the questionnaire-based instrument, we obtained data from over 600 physicians who were participating in programs in which they received financial incentives to achieve quality targets.

In general, the empirical results supported the scientific integrity of the instrument. Specifically, the psychometric procedures we applied to the data, factor analysis, and MTA, produced scales that corresponded to our a priori conceptual model. These scales relate to provider attitudes about different features of P4P programs, namely awareness, financial salience, clinical relevance, cooperation, concern for unintended consequences, and control. As hypothesized, each of these attitudinal measures was also a statistically significant predictor of providers' perceived impact of quality-based financial incentives. In certain instances where the empirical methods did not group items with their a priori scales, a logical relationship existed between the hypothesized and observed scale structures. For example, we originally hypothesized separate awareness and feedback scales. The empirical analyses, however, combined these items into a single scale. In retrospect, it seems reasonable that the timeliness and quality of the feedback that physicians receive about their performance regarding the quality targets will be related to their degree of awareness and understanding of the incentive program.

The scales are also highly consistent with concepts that appear in the literature concerning physician adherence to clinical guidelines and evidence-based practice. As noted, this literature greatly influenced our initial selection of dimensions and related items for the questionnaire. For example, concepts such as awareness and control have also been cited as key factors influencing guideline adherence (e.g., Cabana et al. 1999). In addition, Berwick (2003) emphasized the perceived benefit of a change in clinical practice—the net balance between its potential risks and gains—as the most important of five perceived attributes affecting the diffusion of innovation in health care settings. Two of our scales, clinical relevance and unintended consequences, are directly related to this construct. Such consistency between the results of our psychometric analyses and the general literature on guideline adherence provides support for the construct validity of the present scales and also suggests that these scales have possible application beyond P4P to other topics regarding physician adherence to guidelines and evidence-based practice.

Our analyses also supported the use of our instrument for purposes of comparing provider attitudes among medical groups and other types of provider organizations. The observed group-level reliabilities were generally very good with a median value of 0.60, similar to the values observed on other major measures, such as Consumer Assessment of Healthcare Providers and Systems (CAHPS), used for aggregate-level comparisons (Hargraves, Hays, and Cleary 2003; Solomon et al. 2005). Given some recent research suggesting that medical groups play an important role in mediating quality-related

incentive arrangements from payers (Bokhour et al. 2005), group-level comparisons of provider attitudes may prove to be very informative for managing the implementation of P4P programs.

Certainly, our testing of the instrument has noteworthy limitations. Although the number of physicians who responded to the questionnaire was substantial, the overall adjusted response rate of 32 percent nonetheless raises questions about the representativeness of the respondents. It is not possible for us to know whether the respondents differed in any systematic way from the nonrespondents in their attitudes toward quality targets and incentives. However, for RIPA physicians we were able to compare respondents and nonrespondents with regard to medical specialty (family practice, internal medicine, pediatrics), practice type (multispecialty or single specialty), and practice size. We found significant differences with regard to medical specialty and practice size, but these were of very modest magnitude, and neither variable was significantly related to attitudes among the respondents. For Massachusetts, we were able to compare the age and medical specialty distributions of our respondents to those of the almost 8,500 licensed primary care providers in the state. We found differences on both variables, but as with RIPA these were of modest magnitude and unrelated to attitudes. Additionally, scale means differed substantially, suggesting that physicians' opinions differed from dimension to dimension and that, overall, responses to the questionnaire were not uniformly positive or negative as would be expected if they were primarily determined by a general bias for or against incentives.

Another limitation is that we tested the validity of the measures by using data collected concurrently from the same instrument to regress the physicians' perceived impact of incentives on other measures of their attitudes toward P4P programs. As such, the strength of the findings we report may be somewhat inflated by common-method bias. An important direction for future research would be an empirical examination of the relationship between physician attitudes on one hand and the actual degree of success of P4P incentives on the other, with the latter measured objectively by the quality target adherence rates of the same physicians who responded to the attitude questionnaire.

Important opportunities also exist to improve the instrument itself. In particular, the current two-item version of the financial salience scale had relatively low internal consistency in that a sample of 148 would be required to achieve the degree of reliability generally regarded as adequate for group comparisons. Thus, one opportunity would be to improve the reliability of the financial salience scale through the addition of one or two items. The same is

true for the three-item control scale, which demonstrated adequate but not strong internal consistency reliability.

In summary, using separate derivation and validation samples, we were able to create replicable and psychometrically sound scales representing providers' attitudes toward seven key features of P4P programs. As the number and variety of such programs increase in the future, an assessment of their successes and failures will depend in large part on an understanding of relevant provider attitudes. We believe that the instrument described here will allow for the reliable and valid measurement of those attitudes and thereby further the development of effective P4P programs.

ACKNOWLEDGMENTS

Financial support for this article was provided by the Agency for Healthcare Research and Quality and the Robert Wood Johnson Foundation.

The authors wish to thank Anthony L. Suchman, M.D., M.A., for his thoughtful comments regarding physician attitudes toward health plans and incentives, and two anonymous reviewers for their insightful and constructive critique of this manuscript. We also wish to thank the primary care physicians of the three medical groups who were so generous with their time and candid in their feedback during the early stages of instrument development.

Disclosures: The authors have no affiliation with or financial interest in any product mentioned in this article. The authors' research was not supported by any commercial or corporate entity.

Disclaimers: Drs. Meterko, Young, Bokhour, Burgess, Berlowitz, and Ms. Nealon Seibert are affiliated with the Department of Veterans Affairs. However, the views expressed herein are solely the responsibility of the authors and do not necessarily reflect the views of the Department of Veterans Affairs.

REFERENCES

- Aday, L. A. 1989. *Designing and Conducting Health Surveys*. San Francisco: Jossey-Bass.
- American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. 1985. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.
- Baker, G., and B. Carter. 2005. *Provider Pay-for-Performance Incentive Programs: 2004 National Study Results*. San Francisco: Med-Vantage.

- Banaszak-Holl, J., J. S. Zinn, and V. Mor. 1996. "The Impact of Market and Organizational Characteristics on Nursing Care Facility Service Innovation: A Resource Dependency Perspective." *Health Services Research* 31 (1): 97-117.
- Berwick, D. M. 2003. "Disseminating Innovations in Health Care." *Journal of the American Medical Association* 289 (15): 1969-75.
- Bokhour, B. G., J. F. Burgess, J. M. Hook, B. White, D. Berlowitz, M. R. Guldin, M. Meterko, and G. J. Young. 2006. "Incentive Implementation in Physician Practices: A Qualitative Study of Practice Executive Perspectives on Pay for Performance." *Medical Care Research and Review* 63 (1, Suppl): 73S-95S.
- Cabana, M. D., C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P. A. Abboud, and H. R. Rubin. 1999. "Why Don't Physicians Follow Clinical Practice Guidelines?" *Journal of the American Medical Association* 282 (15): 1458-65.
- Campbell, D. T., and D. W. Fiske. 1959. "Convergent and Discriminant Validity by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56: 81-105.
- Cattell, R. 1966. "The Meaning and Strategic Use of Factor-Analysis." In *Handbook of Multivariate Experimental Psychology*, edited by R. B. Cattell, pp. 174-243. Chicago: Rand McNally.
- Conrad, D. A., and J. B. Christianson. 2004. "Penetrating the 'Black Box': Financial Incentives for Enhancing the Quality of Physician Services." *Medical Care Research and Review* 61 (3, suppl): 37S-68S.
- Dudley, R. A., R. H. Miller, T. Y. Korenbrot, and H. S. Luft. 1998. "The Impact of Financial Incentives on Quality of Health Care." *Millbank Quarterly* 76 (4): 649-86.
- Guertin, W. H., and J. P. Bailey. 1970. *Introduction to Modern Factor Analysis*. Ann Arbor, MI: Edwards.
- Hargraves, J. L., R. D. Hays, and P. D. Cleary. 2003. "Psychometric Properties of the Consumer Assessment of Health Plans Study (CAHPS) 2." *Health Services Research* 38 (6, part 1): 1509-27.
- Hays, R. D., and T. Hayashi. 1990. "Beyond Internal Consistency Reliability: Rationale and User's Guide for Multitrait Analysis Program on the Microcomputer." *Behavior Research Methods, Instruments and Computers* 22 (2): 167-75.
- Hays, R. D., J. A. Shaul, V. S. L. Williams, J. S. Lubalin, J. D. Harris-Kojetin, S. F. Sweeny, and P. D. Cleary. 1999. "Psychometric Properties of the CAHPS 1." *Medical Care* 37 (3, suppl): 22-31.
- Institute of Medicine. Committee on Quality of Health Care in America. 2001. *Crossing the Quality Chasm: A New Health System for the Twenty-First Century*. Washington, DC: National Academy Press.
- Kaiser, H. F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20: 141-51.
- Kerlinger, F. N. 1973. *Foundations of Behavioral Research*, 2d Edition. New York: Holt, Rinehart and Winston Inc.
- Nunnally, J. 1978. *Psychometric Theory*, 2d Edition. New York: McGraw-Hill.
- Pathman, D. E., T. R. Konrad, G. L. Freed, V. A. Freeman, and G. G. Koch. 1996. "The Awareness-to-Adherence Model of the Steps to Clinical Guideline Compliance: The Case of Pediatric Vaccine Recommendations." *Medical Care* 34 (9): 873-89.

- Shortell, S. M., J. L. O'Brien, J. M. Carman, R. W. Foster, E. F. Hughes, H. Boerstler, and E. J. O'Connor. 1995. "Assessing the Impact of Continuous Quality Improvement/Total Quality Management: Concept Versus Implementation." *Health Services Research* 30 (2): 377-401.
- Solomon, L. S., R. D. Hays, A. M. Zaslavsky, L. Ding, and P. D. Cleary. 2005. "Psychometric Properties of a Group-Level Consumer Assessment of Health Plans Study (CAHPS) Instrument." *Medical Care* 43 (1): 53-60.
- Tabachnick, B. G., and L. S. Fidell. 1983. *Using Multivariate Statistics*. New York: Harper & Row.
- Tamblyn, R., P. McLeod, J. A. Hanley, N. Girard, and J. Hurley. 2003. "Physician and Practice Characteristics Associated with the Early Utilization of New Prescription Drugs." *Medical Care* 41 (8): 895-908.
- Ware, J. E., W. J. Harris, B. Gandek, B. W. Rogers, and P. R. Reese. 1997. *Multitrait/Multi-Method Analysis Program—Revised*. Boston: Health Assessment Lab.
- Young, G. J., B. White, J. F. Burgess Jr., D. Berlowitz, M. Meterko, M. R. Guldin, and B. G. Bokhour. 2005. "Conceptual Issues in the Design and Implementation of Pay-for-Quality Programs." *American Journal of Medical Quality* 20 (3): 144-50.