The World Health Organization Health and Work Performance Questionnaire (HPQ)

Ronald C. Kessler, PhD
Catherine Barber, MPA
Arne Beck, PhD
Patricia Berglund, M.BA
Paul D. Cleary, PhD
David McKenas, MD
Nico Pronk, PhD
Gregory Simon, MD
Paul Stang, PhD
T. Bedirhan Ustun, MD
Phillip Wang, MD, ScD

*This report describes the World Health Organization Health and Work Performance Questionnaire (HPQ), a self-report instrument designed to estimate the workplace costs of health problems in terms of reduced job performance, sickness absence, and work-related accidents-injuries. Calibration data are presented on the relationship between individual-level HPQ reports and archival measures of work performance and absenteeism obtained from employer archives in four groups: airline reservation agents (*n = *441), customer service representatives (*n = *505), automobile company executives (*n = *554), and railroad engineers (*n = *850). Good concordance is found between the HPQ and the archival measures in all four occupations. The paper closes with a brief discussion of the calibration methodology used to monetize HPQ reports and of future directions in substantive research based on the HPQ.* (J Occup Environ Med. *2003;45:156–174*)

Interest in the social consequences of illness has broadened in the past decade as epidemiologists have joined health economists and health services researchers to devise methods that rationalize the allocation of health care resources.[1,2] Research showing that untreated and undertreated health problems exact substantial personal costs from the individuals who experience them as well as from their families, employers, and communities has been a central part of this work.[3,4] Among the most important of these results have been those concerning the workplace costs of illness from the perspective of the employer.[5,6] These costs have enormous implications for the economy. For example, a recent analysis estimated that depression causes an annual loss of $33 billion in work absenteeism in the U.S.[7] Given the low rate of depression treatment[8] and the fact that treatment substantially improves role functioning among people with depression,[9,10] such data suggest that it might be cost-effective for employers to increase the proportion of depressed workers who receive treatment.[11] Similar arguments have been made for a number of other illnesses,[12–14] the notion being that targeted expansion of employee health care benefits, including an outreach component, can represent an investment opportunity for employers.

Only a small minority of employers has as yet been convinced of the business case for targeted investment in employee health care. This is partly a result of the absence of data to evaluate the indirect costs of un-

From Harvard Medical School, Boston, Massachusetts (Dr Kessler, Dr Cleary, Dr Wang); Harvard School of Public Health, Boston, Massachusetts (Ms Barber); Kaiser-Permanente, Denver, CO (Dr Beck); University of Michigan, Ann Arbor, MI (Ms Berglund); American Airlines, Fort Worth, TX (Dr McKenas); HealthPartners, Minneapolis, MN (Dr Pronk); Center for Health Studies, Group Health Cooperative, Seattle, WA (Dr Simon); Galt Associates, Blue Bell, PA (Dr Stang); and World Health Organization, Geneva, Switzerland (Dr Ustun).

Address correspondence to: R. C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; e-mail: kessler@hcp.med.harvard.edu.

DOI: 10.1097/01.jom.0000052967.43131.51

treated and under-treated health problems in the workplace. Employers who have access to integrated databases on medical expenditures, pharmacy expenditures, workplace injuries, and disability can go part way to resolve this problem, as such databases can be used to evaluate the effects of changes in health plan benefits over time on a number of important employer costs.[15] However, even completely integrated databases typically lack two critical types of information that are needed to make a strong business case for expanded investment in employee health care. First, few companies have accurate individual-level job performance data for most of their employees. Even basic data on sickness absence are generally available only for blue-collar and pink-collar workers, but not for white-collar workers, whereas data on job performance are usually either nonexistent, superficial, or very difficult to obtain in machine-readable form. Second, medical data, when available, typically focus on treated health problems. No information is generally available on untreated health problems other than in companies that perform routine physical examinations on all their employees and link these data with information about job performance. In the absence of such linked data files, it is impossible either to estimate the number of workers with unmet need for treatment or the effects of untreated health problems on work performance.

Recognizing the need for such data, a number of health services researchers have developed self-report measurement tools to collect data in employee surveys on untreated health problems and work performance. Although inferior to objective performance-based measures, self-report work performance measures can be extremely useful when objective measures are unavailable. This is especially true when the self-report measures are calibrated against objective measures in such a way that scores on the self-report measures can be meaningfully interpreted.

Lynch and Riedel[16] recently reviewed the most widely used work performance measures in the literature. Their review showed that these measures generally have good internal consistency reliability and good face validity, but have not been compared to objective data on work performance either to demonstrate their validity or to generate calibration rules. The current report presents data of this sort for one such self-report work performance measure, the World Health Organization's (WHO) Health and Work Performance Questionnaire (HPQ). Data are presented from four HPQ calibration surveys, each carried out in a separate corporation and focused on a single type of worker for whom archival data were available either on sickness absence, work performance, or both. The four samples include reservation agents working for a major airline, customer service representatives working for a large telecommunications company, executives working for a major automobile manufacturer, and railroad engineers working for a large railroad company. Data are presented on the relationships of individual-level HPQ job performance measures with archival measures used by the companies to monitor worker performance. Data are also presented on comparisons between individual-level HPQ absenteeism data and employer payroll records. The paper closes with a brief discussion of calibration methodology and future directions in substantive research based on the HPQ.

Development of the HPQ

Background

The HPQ was developed as an expansion of the work role module in the WHO Disability Assessment Schedule (WHO-DAS).[17] The WHO-DAS is a self-report measure of role functioning that was created by WHO for use in community surveys as well as in intervention studies aimed at reducing the role impairments associated with untreated or under-treated health problems. The full WHO-DAS includes scales of role functioning in each of the core domains of the newly revised International Classification of Functioning,[18] whereas the HPQ focuses exclusively on the work role domain.

HPQ development began with a review of other existing scales, followed by pilot interviews, development of preliminary questions, systematic evaluation and refinement of these questions by experts in survey question wording using the methods described by Converse and Presser,[19] and additional pilot testing with cognitive debriefing interviews aimed at detecting and removing ambiguities in question wording.[20] Full-scale pilot surveys were then carried out in three managed care samples and one large corporation sample in order to study the psychometric properties of the scales and to examine the effects of various chronic conditions on HPQ measures of work performance.[21] The final HPQ, based on all these earlier studies, was then administered to the four calibration samples described in the current report.

The complete text of the HPQ is posted at: http://www.hcp.med.harvard.edu/hpq. Benchmark survey scores, information on using the HPQ, and updates of ongoing HPQ evaluations will be posted on this site as they become available.

Work Performance

Three outcomes are traditionally measured in studies conducted by organizational researchers on the effects of various workplace productivity enhancement interventions: absenteeism, work performance, and job-related accidents.[22] We decided to measure all three of these outcomes in the HPQ. Work performance is obviously the most difficult of these three to assess. Indeed, the decision to develop the HPQ was based largely on our failure to find an existing self-report measure of

work performance that met our needs.

The ideal way to assess work performance, of course, would be by means of objective performance-based assessment rather than self-report. Many employers have developed assessments of this sort for at least some of their workers.[23,24] However, these systems vary enormously in coverage as well as in sophistication, making them impossible to use in broad-based studies of health and work performance. Another possibility is to use special performance-based tests, many of which have been developed in conjunction with the Department of Labor Occupational Information Network (O*NET) system of job classification.[25] However, these tests assess ability rather than actual performance on the job, making it impossible to evaluate under-performance by workers with high native ability who fail to perform up to their ability on the job.[26] Based on these considerations, we concluded that self-report measures are the most feasible tools for our purposes.

A comprehensive review of the literature found a number of useful self-report measures of work performance.[27–29] The most compelling of these, however, focused on single occupations and included questions that were tailored to the unique demands of those occupations.[30–32] The measure we needed, in comparison, had to be appropriate across the full range of the occupational spectrum. While we found scales of this sort in our review, they all suffered from one of three serious problems: unequal relevance across the full range of the occupational spectrum; incomplete coverage of important performance domains; or lack of translation rules to link domain-specific performance measures with an overall assessment of work performance. These problems are briefly reviewed in the next four paragraphs.

The problem of unequal relevance stems from the difficulty of developing concrete self-report work performance questions that are equally relevant to workers across the full range of the occupational spectrum. A number of work performance scales can be faulted along these lines. The Work Productivity Scale,[33] for example, includes questions about putting off business phone calls and failing to attend business meetings, whereas the Stanford Presenteeism Scale[34] includes questions about being cranky with work subordinates and failing to find new-creative solutions to work problems. These questions are much more relevant to white-collar workers than to blue-collar or pink-collar workers, introducing a bias into these scales that can lead to an overestimation of the prevalence of impaired work performance among white-collar workers compared to other workers. This bias, in turn, can lead to biased results suggesting that health problems have greater effects on white-collar workers than on other workers and that the health problems most relevant to the performance of white-collar workers have greater adverse effects on work functioning than the health problems most relevant to the performance of other workers.

Other measures of work performance have been designed explicitly to overcome the problem of equal relevance across the occupational spectrum[35,36] by including brief assessments of health-related impairments in each of a wide number of basic domains of role performance (eg, mobility, vision and hearing, fine motor coordination, concentration, communication). The hope is that this heterogeneous coverage will tap the main job demands of workers in all occupations. However, no systematic attempt is made in these scales to assess all the important domains of work performance that need to be covered. As a result, although these scales cover a number of domains, there is no reason to believe that the coverage is either comprehensive or comparable across all occupations. The depth of this problem can be seen by examining the Dictionary of Occupational Titles (DOT),[37] a document prepared by the Department of Labor (DOL) to describe the skill sets needed in the over 22,000 occupations in the U.S. labor force. Systematic observation of day-to-day work performance in each of these occupations by DOL employees documented over 50 different work performance domains. No existing work performance scale either assesses all these domains or attempts systematically to sample across these domains.

This problem could be overcome if a small number of global work performance domains was isolated empirically. An early psychometric analysis designed to search for such global domains in the DOT yielded promising results.[38] More current work along the same lines might evolve from DOL's new on-line O*NET system of job classification (www.onetcenter.org). Indeed, one goal of the O*NET system is to assemble a set of objective performance-based tests that will cover all the many different dimensions of work performance in the O*NET classification. Self-report measures already exist for some of the O*NET dimensions.[39,40] It is conceivable that self-report measures of the other O*NET dimensions could be developed. However, a comprehensive set of such scales, if they were ever developed, would likely take hours or even days to administer.

Furthermore, even assuming that comprehensive O*NET scales could be developed and used, a final problem would remain: that no rules exist to combine the separate domain scores into an overall measure of work performance that is valid across all occupations. Such combination rules would, at a minimum, require different weights to be applied across domains for different occupations. Health-related difficulties in the domain of unskilled manual labor (eg, digging, lifting, carrying), for example, are presumably much more impairing to a manual

laborer than to a lawyer. This difference would have to be taken into account in combining domain performance scores into an overall work performance score that applies equally well to laborers, lawyers, and to workers in all the thousands of other occupations in the labor force. In addition, the correct combination rules are likely to be quite complex, involving different domain weights, nonlinearities, and nonadditivities for particular occupations or classes of occupations. It might be possible to develop such rules by analyzing extremely large databases containing the appropriate variables. However, in the absence of such databases and such rules, neither of which currently exists, it is unclear how one could arrive at a principled basis for combining domain-specific work performance scores into a valid overall assessment of work performance.

Based on these considerations about the current intractability of the above problems, we decided to use a simple self-report global rating scale to assess work performance in the HPQ. In this approach, respondents are asked to rate their overall work performance during the past four weeks on a 0-to-10 self-anchoring scale in which 0 is defined as the "worst possible work performance" a person could have on this job and 10 is defined as "top work performance" on this job. Our reasoning in selecting this simple approach was that workers are in a better position than researchers to recognize the work performance domains that are most relevant to their particular occupations, to evaluate their recent performance in these domains, and to arrive at a rating of their overall work performance based on this evaluation.

At the same time, we know from the methodological literature that responses to 0-to-10 global rating scales can be improved by two refinements, both of which we used in the HPQ: decomposition[41] and internal anchoring.[20] Decomposition is one of several strategies that have been developed by survey methodologists to facilitate active memory search in response to complex survey questions. Research on the cognitive processes used to arrive at accurate answers to complex survey questions shows that active memory search and review of component experiences substantially improve response accuracy.[42] This same research shows, though, that many respondents give superficial answers based on general semantic memories or other response heuristics because they are unwilling to engage in serious memory search.[43] Decomposition addresses this problem by asking preliminary component questions that force respondents to engage in active memory search before being asked the complex question.

Decomposition is used in the HPQ by asking respondents a series of questions that require them to review critical aspects of their work performance before assigning themselves a rating on the global 0-to-10 scale. Specifically, we ask component questions about quantity of work (how often during the recall period: the respondent's speed/productivity of work was lower than expected, the respondent did no work at times he/she was expected to be working), quality of work (how often during the recall period: the respondent did not work as carefully as he/she should, his/her work quality was lower than expected, he/she was daydreaming and not concentrating on work), interpersonal aspects of work (how often during the recall period: the respondent had trouble getting along with others at work, had difficulty controlling his/her emotions at work, and avoided interacting with others at work), special work successes, special work failures, and accidents-injuries. All of these questions were explicitly designed to be sufficiently general that they apply to all occupations, but sufficiently focused that they facilitate relevant memory search and review. The global 0-to-10 scale is administered only after these memory-priming decomposition questions are asked.

Internal anchoring is an especially important strategy to improve the accuracy of responses to questions that use self-anchoring response scales. The issue here is that most self-anchoring scales define only the ends of the distribution (eg, 0 defines the "worst possible performance" whereas 10 defines the "best possible performance"), but not intermediate values, while the vast majority of respondents rate themselves as having values between these extremes and have no guidelines for selecting among intermediate values.

Schwarz[44] has shown that this problem can be addressed by rescaling 0-to-10 scales to range from -5 to $+5$ with a clear 0 point in the middle. This rescaling substantially improves the accuracy of response to self-anchoring scales in the middle part of the scale distribution by highlighting the midpoint. The difficulty with this approach in the case of rating work performance, however, is that we have no reason to believe that the performance of the average worker is at a level halfway between the theoretical extremes of worst and best performance. Indeed, our pilot research found that most workers report average performance in their occupation to be substantially above the midpoint of this range. Based on this evidence, we designed the HPQ rating so that respondents could generate their own internal anchors before responding. This is done by asking each respondent to give separate ratings for the average worker on their job and for their own usual performance before rating their recent performance. In addition to providing internal anchors, these additional rating questions provide information that allows us to calculate ipsative scores of recent performance in comparison to usual performance and in comparison to the performance of other workers. In order to obtain multiple indicators for self-other comparisons, the HPQ also includes a separate question that

asks respondents explicitly to compare their own recent performance with that of the average worker on the same job using a standard seven-point better-worse unfolding scale (ie, better, worse, or about the same and, if either better or worse, a three-category rating of personal performance as either a lot, some, or only a little better/worse than the average worker).

Absenteeism

Most health and work performance instruments assess absenteeism with a single question about the number of days in the past month (or other recall period) the respondent missed a day of work because of illness. Previous research has shown good agreement between these self-reports and employer records of absenteeism.[45] However, the results of cognitive interviews led us to use a more detailed set of questions about absenteeism in the HPQ. Four refinements were involved. First, we decided to focus on hours rather than on days of work during the past four weeks based on the fact that workers differ substantially in the number of hours they work as well as in whether they work the same number of hours each day. Second, in addition to asking about hours missed on sickness absence days, we ask about hours missed on workdays (ie, coming in late or going home early) due to the fact that a substantial proportion of missed work time occurs on days when people come to work. Third, we ask about extra hours of work (ie, coming in early, going home late, working on days off) because of the fact that many workers put in extra hours to make up for sickness absence. Fourth, although we distinguish between sickness absence and other types of absence (eg, vacation, absence due to a family emergency etc.), we also create a combined measure of total hours absent because workers who have used up their allotted sick days often use accrued personal days or vacation days when they are too ill to come to work. In addition, many employers consolidate the number of paid absence days they allow their workers to take into a single category that combines vacation and personal days and sickness absence days, making the distinction among these categories artificial.

The question sequence in the HPQ absenteeism series makes use of the same decomposition strategy described above in the discussion of the work performance measure. Specifically, the series begins by asking separately about number of days missed in the past four weeks for vacation and sickness absence, followed by number of partial workdays, and about days of extra hours worked. The aim is to focus memory search by simplifying the task of calculating total lost work hours in response to a single question. It is noteworthy that the decomposition questions ask about days rather than hours, even though hours are the unit of ultimate interest, because cognitive interviews show that the vast majority of working people reconstruct work schedules more naturally in terms of days than hours. At the end of this sequence, we ask about overall hours worked. It is noteworthy that we ask about hours worked rather than hours missed because cognitive interviews show that most workers think more naturally in terms of the former than the latter. A final question is asked about the number of hours each week the respondent is normally expected to work in order to have a denominator for calculating a percentage measure of work loss.

Job-Related Accidents

Although job-related accidents are uncommon, they are important because of their potential high cost. We explored a number of options for asking fully structured questions about accidents. In the end, though, the rarity and great variety of accidents led us to include a single open-ended question about job-related accidents-injuries in the final HPQ. This question is worded in such a way that respondents are explicitly asked to include incidents that led either to 1) breakage or other loss of property; 2) delays in production or other decreases in work performance of the respondents or other workers; 3) physical injury of the respondent or others; and 4) serious risk of loss, delay, or injury. The textual responses to these questions are converted into general anonymous vignettes and presented to supervisors for scoring in terms of their monetary cost to the company. Open-ended reports are also obtained for responses to questions about special successes (eg, making a big sale, getting a bonus or a promotion, being selected as the employee of the month) and special failures (eg, failing to meet a production quota, reprimand from a supervisor, failing to get an expected bonus or promotion). As with accidents-injuries, special successes and failures, although comparatively uncommon, are very important components of the overall indirect costs of illness and the cost-savings associated with treatment. As with accidents, open-ended responses to the questions about successes and failures are converted into general anonymous vignettes and presented to supervisors to obtain estimates of the costs to employers of failures and the values of successes.

The HPQ Calibration Survey

Samples

Calibration surveys were performed in four occupations to compare HPQ work performance and absenteeism measures with archival data obtained from employer records. No attempt was made to validate the HPQ question about accidents-injuries because of the rarity of these events. The four occupations were reservation agents working for a major airline, customer service representatives working for a large telecommunications company, executives working for a major automobile

TABLE 1
Sample Dispositions

	Reservation Agents	Customer Service Representatives	Executives	Railroad Engineers
	%	%	%	%
Invalid Numbers[1]	29.3	37.6	15.0	15.8
Answering Machines	19.9	18.0	27.8	10.3
Refusals	12.2	14.1	8.2	16.8
Cooperation Rate	75.9	64.0	85.6	77.1
Initial Sample	(1143)	(1713)	(1131)	(1491)
Completed Interviews[2]	(441)	(505)	(554)	(850)

[1] Invalid numbers are defined as disconnected numbers with no forwarding number, incorrect numbers (e.g. businesses, fax machines, respondent unknown), Good numbers to respondents who report that they are no longer working for the company, and no contact after 20 call attempts.

[2] Cooperation rate is defined as completed interviews divided by completed interviews plus refusals.

manufacturer, and railroad engineers working for a large railroad company. Names, home addresses, and home telephone numbers were obtained for initial samples of between 1131 and 1491 workers in each occupation. An advance letter (or, in the case of the executives, an e-mail) was sent to each predesignated respondent by the medical director of their company. The letter informed recipients that the company was collaborating with researchers from Harvard Medical School (HMS) in a survey of employee health and work performance. The letter went on to say that an HMS telephone interviewer would contact them in the next week to carry out a telephone interview. The letter made it clear that participation was completely voluntary and anonymous. An 800 number to the HMS study manager was included in the letter for recipients who had questions or who wanted to opt out of the survey. A prestamped and pre-addressed return postcard was included in the mailing for recipients who wanted either to report good times to be reached or to opt out by mail. Professional telephone interviewers made 20 attempts to contact each of those who did not initially opt out. Verbal informed consent was obtained before administering the survey. These recruitment and consent procedures were approved by the Human subjects Committee of Harvard Medical School.

As shown in Table 1, the telephone lists had substantial proportions of invalid numbers (15.0–37.6% across samples) and high proportions of answering machines (10.3–27.8% across samples). The refusal rate (including initial opt-outs) was in the range 8.2–16.8% across samples, while the cooperation rate (the percent of completed interviews among people who were contacted) was in the range 64.0 to 85.6% across samples. Calibration interviews were completed with 441 reservation agents, 505 customer service representatives, 554 executives, and 850 railroad engineers. The demographic distribution of the samples is presented in Table 2. Reservation agents were largely women, while executives and railroad engineers were largely men. The modal age range of reservation agents and customer service representatives was 30 to 44, whereas executives and railroad engineers were generally older (the mode being in the age range 45 to 59). Railroad engineers had the lowest education (50.6% had no more than a high school education), while executives had the highest educations (98.7% were college graduates).

Once the calibration survey was completed, probability subsamples of 105 reservation agents and 181 customer service representatives who participated in the calibration survey were recruited into a 1-week follow-up Experience Sample Method (ESM) evaluation[46] of moment-to-moment work experience. The ESM design involved giving participants a beeper and an ESM diary to keep with them at all times during the seven-day study week. The beeper was programmed by an auto-dialer to be called at five random times each day, with randomization beginning at the start of the workday (or, on regularly scheduled days off work, 1 hour after the respondent reported typically awakening) and ending two hours before the respondent reported typically going to bed on that day of the week. A constraint was imposed on the randomization that no call could be made less than 90 minutes after the preceding call. The respondent was asked to fill out the diary as soon as the beeper went off. The diary asked structured questions about whether the respondent was at work and, if so, about quantity and quality of work at the moment-in-time when the beeper went off. The last entry of each day asked additional questions about the day overall. A separate diary book was provided for each of the seven days. Respondents were asked to mail back each day's completed book the following morning in a pre-stamped, pre-addressed mailer

TABLE 2
Demographic Distributions of the Samples

	Reservation Agents		Customer Service Representatives		Executives		Railroad Engineers	
	%	(se)	%	(se)	%	(se)	%	(se)
Sex								
Female	80.3	(1.9)	47.2	(2.2)	19.3	(1.7)	2.4	(0.5)
Male	19.7	(1.9)	52.7	(2.2)	80.7	(1.7)	97.6	(0.5)
Age								
18–29	6.6	(1.2)	23.9	(1.9)	0.0	(0.0)	4.5	(0.7)
30–44	46.7	(2.4)	49.3	(2.2)	22.7	(1.8)	37.3	(1.7)
45–59	40.5	(2.4)	25.2	(1.9)	69.5	(2.0)	52.9	(1.7)
≥ 60	6.2	(1.2)	1.6	(0.6)	7.6	(1.1)	5.3	(0.8)
Education								
<High school	0.2	(0.2)	0.8	(0.4)	0.0	(0.0)	1.9	(0.5)
High school	21.9	(2.0)	15.7	(1.6)	0.5	(0.3)	48.7	(1.7)
Some college	36.4	(2.4)	43.3	(2.2)	0.7	(0.4)	37.1	(1.7)
≥College	41.4	(2.4)	40.3	(2.2)	98.7	(0.5)	12.4	(1.1)
(*n*)	(441)		(505)		(554)		(850)	

in order to avoid the problem of retrospective completion that sometimes occurs when diaries are sent back only at the end of the study.[47] Reminder phone calls from data collection staff were made on the evening of the first, third, and fifth diary days to encourage respondents to stick with the task.

With 35 possible diary entries for each respondent (5 per day × 7 days), there were 3675 (105 × 35) logically possible completed ESM diary entries for the reservation agents and 6335 (181 × 35) for the customer service representatives. The response rates for the entries in the two samples were, respectively, 61.3% (n = 2253) and 68.7% (n = 4353). Among reservation agents, 44.8% of valid entries were made while the respondent was at work (n = 1010) and 80.8% of the latter were made while the respondent was working as opposed to on break or at lunch (n = 816). Among customer service representatives, 56.3% of valid entries were made while the respondent was at work (n = 2450) and 78.5% of the latter were made while the respondent was working (n = 1926). A 1-week version of the HPQ work performance and absenteeism questions was administered as part of the debriefing telephone interview that was administered the day after the end of the diary week. This allowed us to calibrate HPQ ratings against aggregated ESM reports in an effort to evaluate the effects of recall bias on HPQ reports. The debriefing interviews were completed with 91 of the 105 reservation agents (86.7%) and 172 of the 181 customer service representatives (95.0%).

The Archival Work Performance Measures

The four samples considered here were selected because the performance of the workers in these occupations is evaluated using standardized assessments. Customer service representatives and airline reservation agents receive monthly supervisor performance ratings based on a combined score for quantity of work (number of cases resolved, number of tickets sold) and quality of work (based on supervisor review and coding of audio-taped customer interactions). The executives all have 360-peer evaluations of overall leadership based on the Campbell Leadership Index.[48] The railroad engineers receive performance ratings for each trip they make that combines information about speed tracking (ie, arriving at benchmark points on the route as close as possible to target times), brake wear, fuel efficiency, and a number of other safety and efficiency indicators. In addition, the two ESM samples generated moment-in-time data on work performance that avoid the recall bias inherent in more conventional self-report measures. As a result, these measures were treated as additional outcomes in the calibration analysis. This was done by combining ESM ratings into a scale with four items derived from exploratory factor analysis of moment-in-time performance reports (speed of work, quality of work, concentration on work, and perceived success at current work task), each of which was rated on a 1 to 7 self-anchoring scale of either "low" to "high" quality and speed or "not at all" to "very much" concentrating and succeeding. The Cronbach's alpha for this scale was 0.74 for the reservation agents and 0.81 for the customer service representatives.

We were required by our Institutional Review Board (IRB) to obtain informed consent from respondents before we could access their performance records. This consent was obtained in conjunction with the baseline telephone HPQ interviews. As a

TABLE 3
Distributions of the Work Performance Outcome Measures[1]

	Range		Low			High			
	Lower	Upper	%	(se)	Score	%	(se)	Score	(n)
Reservation agent supervisor ratings	79	100	21.1	(2.0)	95	25.6	(2.2)	100	(441)
Reservation agent ESM	0	100	19.7	(1.4)	50	22.2	(1.5)	92	(105)
Customer service representative ESM	0	100	18.7	(0.9)	46	18.8	(0.9)	83	(181)
Executive leadership scores	32	80	20.4	(2.5)	55	20.4	(2.5)	67	(269)
Railroad performance actions	0	1	0.8	(0.3)	0	—	—	—	(847)

[1] The Railroad Performance Action measure is a dichotomy (yes/no). All other measures are scales that have been transformed to a theoretical range between 0 and 100, with higher scores indicating better performance.

result, we were not able to evaluate the completeness of the archival information before selecting and interviewing the respondents. This led to considerable loss of objective data. The most extreme loss was for the customer service representatives, whose archival work performance and absenteeism data were unusable because of a corruption of the identification number link in the employer records. Archival performance data were obtained, though, for 441 reservation agents, 269 executives, and 847 railroad engineers. The reservation agent data were the most precise with regard to time in that they were based on supervisor ratings for the month prior to the HPQ survey. The executive data, in comparison, were based on peer evaluations made at the end of a leadership-training program that respondents participated in for one week at variable times up to two years before the survey. The ratings for railroad engineer were the least precise with regard to time in that they were aggregated at the end of each trip into a summary score that represented the engineer's cumulative performance over many years. It was impossible to recover disaggregated summary ratings from this file. However, we were able to obtain information about serious engineer performance problems in the month before the interview from a separate performance action file. As a result, we focused on this measure in the evaluation of the HPQ among railroad engineers. The EMS ratings, finally, were evaluated for the 816 moments-in-time when the 91 reservation agents who completed the post-ESM debriefing interview were at work and working (ie, not on break or at lunch) and for the 1926 moments-in-time when the 172 customer service representatives who completed the post-ESM debriefing interview were at work and working.

Four of the five outcome work performance measures, the exception being the dichotomous railroad engineer performance action measure, were transformed to 0–100 scales from their original metrics. These transformations were then used to divide workers into top performers, low performers, and average performers. The decision to make this three-part division was based on the results of focus groups with managers, who reported that they use work performance measures largely to target high performers for reward and low performers for remediation and that they generally do not make distinctions within the middle part of the range. Top performance was defined as the top 20th percentile of the range of each objective performance measure, while low performance was defined as the bottom twenty percentile of each objective performance measure. As shown in Table 3, all the outcome performance measures were refined enough at both tails of the distribution to make distinctions very near these target proportions. It is also noteworthy that the ESM performance measures have a wider range than the archival measures, suggesting that there is some subjective truncation of supervisor ratings. This is especially clear in the case of the supervisor ratings of reservation agents, where the lowest score is 79 on the 0-to-100 scale. The empirical distribution of the ESM scale for these same workers, in comparison, spans the entire scale range, showing that the workers themselves make more subtle distinctions about their work performance than do supervisors.

This observation raises the question as to how objective the archival data are. Although these data are treated as objective for purposes of calibrating the HPQ, we are aware that the archival data, especially those based on supervisor ratings (in the case of the reservation agents and customer service representatives) and peer ratings (in the case of the executives) are not without error. However, these measures are the actual measures used by employers to monitor the performance of workers and are, in this sense, "real" in an operational sense. Yet we would not expect perfect consistency between HPQ measures and these archival measures because we recognize that the latter are imperfect. Indeed, the HPQ ratings might be more accurate than the archival data in some respects. However, it is nonetheless important for us to demonstrate that the HPQ ratings are meaningfully related to the archival measures to assure that the HPQ is tapping the same aspects of workplace performance as those measured by the

work performance measures actually used by employers.

The Payroll Record Measures of Absenteeism

Payroll record measures of absenteeism were available for three of the four occupations, the exception being executives. In the case of reservation agents, data were obtained on hours scheduled and payroll record data were obtained on hours actually worked each day during the 4 weeks leading up to the HPQ interview. These data were aggregated into summary measures of hours worked and hours missed in the 1 week and 4 weeks before the interview. In the case of railroad engineers, who have an erratic work schedule, data were available on the days they were scheduled to work when they either called in sick or were absent for some other reason. These records were aggregated into summary measures of days of work missed in the one week and four weeks before the interview. In addition, the ESM data for the reservation agents and customer services representatives were also used to derive indirect measures of work absence. This was possible because one of the first questions asked in these diaries was whether the respondent was at work, at home, in transit between work and home, or elsewhere at the time of the beep. Because of the fact that the ESM data points were sampled at random moments-in-time throughout the week, these reports should provide representative data of the proportion of time respondents were at work over the week. Therefore, moment-to-moment reports of whether or not the respondent was at work at the time of a random beep were compared with HPQ reports about hours worked and days of work missed to provide an indirect validation of the HPQ reports.

As with the archival work performance measures, we recognize that employer payroll records can be imprecise because of workers or supervisors making erroneous reports about time spent at work. However, for the occupations considered here payroll records are likely to have a high degree of accuracy.

Analysis Methods: Work Performance

Calibration of the HPQ 0-to-10 global work performance rating scale against the archival measures and ESM measures of high and low work performance was carried out using logistic regression analysis in which dichotomous measures of either high or low performance based on either the archival measures or the ESM measures were the outcome variables and dummy variables defining ranges on the HPQ rating scale were the predictors. Chi-square tests were used to evaluate the global significance of the HPQ rating scale in these analyses. Received Operator Characteristic (ROC) curves were used to judge the strength of association between the HPQ ratings and either the archival measures or the ESM measures. Areas under the ROC curves and their 95% confidence intervals were calculated by the nonparametric method.[49]

Analysis methods: absenteeism

Analysis of HPQ absenteeism reports was carried out in two ways. First, linear correlations were calculated between HPQ self-reports and employer payroll records of absenteeism in the samples of reservation agents and railroad engineers. In the case of the reservation agents, this was done for hours worked and hours missed. In the case of the railroad engineers, it was done for days missed. Both one-week and four-week recall periods were examined. We also compared means for all these outcomes in the HPQ self-reports and the employer payroll records. Second, logistic regression analysis was used in the ESM person-time samples of reservation agents and customer service representatives to make an indirect evaluation of the HPQ self-reports about absenteeism. The dependent variable was a dichotomy for whether the respondent was at work on not at each random moment-in-time, while the predictors were the HPQ self-reports of hours worked during the ESM week and days missed during the week obtained in the post-ESM debriefing interview. The equations were estimated using a two-level random-effects model that included both person-level controls (age, sex), and within-person controls (number of days in the ESM study as of the time of the beep, sequence of the beep within the day) in order to improve the precision of estimates.

Results

The Distribution of the Global HPQ Work Performance Ratings

The distributions of the HPQ 0-to-10 global work performance ratings across the five different samples used in the calibration analysis as well as in the full customer service representative sample are presented in Table 4. Three patterns are noteworthy. First, the lower end of the scale is truncated at 0–7 because only a small minority of respondents rated themselves less than 7 in any of the samples. This truncation improves the precision of the calibration procedures described below.[50] Second, there is a clear tendency for the majority of respondents to rate themselves in the high-but-not-perfect range[8,9] much more so than at the very top of the range.[10] Between 61.7% (railroad engineers) and 80.0% (executives) of respondents across samples rated themselves 8 to 9 compared to between 11.9% (executives) and 25.9% (railroad engineers) who rated themselves 10. The distribution of the full reservation agent sample is included in the table even though we have no archival performance measures for that sample in order to present a comparison with the distribution in

TABLE 4
Distributions of the HPQ Global Work Performance Ratings

	Reservation Agents		Reservation Agent ESM		Customer Service ESM		Executives		Railroad Engineers	
HPQ Ratings	%	(se)	%	(se)	%	(se)	%	(se)	%	(se)
0–7	14.7	(1.7)	17.0	(4.0)	45.6	(3.8)	8.2	(1.7)	12.4	(1.2)
8	31.6	(2.7)	38.6	(5.2)	28.4	(3.5)	43.9	(3.0)	31.4	(1.6)
9	34.0	(2.3)	29.5	(4.9)	21.9	(3.2)	36.1	(2.9)	30.3	(1.6)
10	19.7	(1.9)	14.8	(3.8)	4.1	(1.5)	11.9	(2.0)	25.9	(1.5)
(*n*)	(441)		(91)		(172)		(269)		(847)	

TABLE 5
Associations of HPQ Global Ratings with Lowest 20 Percent of Archival and ESM Work Performance Outcome Measures

	Reservation Agent Supervisor Ratings[1]		Reservation Agent ESM		Customer Service Representative ESM		Executive Leadership Scores[1]		Railroad Engineer Performance Actions	
Objective ratings Low work performance[2]	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
0–7	3.2*	(1.3–7.5)	6.4*	(1.7–24.0)	7.3*	(1.6–33.0)	7.0*	(1.3–37.9)	12.3*	(1.3–112.3)
8	2.4*	(1.1–5.2)	1.6	(0.4–6.1)	2.8	(0.6–13.2)	5.4*	(1.2–24.2)	1.0	—
9	1.0	(0.4–2.3)	2.2	(0.6–8.2)	1.6	(0.3–8.0)	2.7	(0.6–12.6)	1.0	—
10	1.0	—	1.0	—	1.0	—	1.0	—	1.0	—
χ^2_3	14.3*		21.7*		57.5*		8.9*		4.91*	
(*n*)	(441)		(816)		(1,926)		(269)		(847)	

* Significant at the .05 level, two-sided test
[1] Based on 30-day job performance rating
[2] Model includes sex, age, day of study, and beep number as controls

the ESM subsample of reservation agents. The latter purposefully oversampled respondents from the full sample who had low ratings in order to increase statistical power in that part of the distribution. This was done based on evidence from the reservation agent sample, which was the first group we surveyed, that low objective work performance is concentrated among respondents with ratings in the 0–7 range on the scale. Third, the distribution in the reservation agent sample is not dramatically different from the distributions in other samples despite the fact that the supervisor ratings of work performance in the reservation agent sample are much more truncated than the other archival measures. This lack of difference in Table 4 is consistent with the suggestion mentioned earlier in the article that reservation agent supervisor ratings might be truncated due to rater bias.

The Associations of HPQ Ratings with Archival and ESM Performance Measures

The results of logistic regression analyses linking HPQ ratings with the five outcome measures of low work performance are presented in Table 5. Logistic regression coefficients have been exponentiated and are reported in the table in the form of odds-ratios (ORs). There is a consistently monotonic and statistically significant relationship between HPQ ratings and the odds of low archival/ESM work performance in all five equations. It is important to note that this association is not caused exclusively by the difference between respondents with HPQ ratings of 0 to 9 versus 10, although that distinction is important in predicting low performance in all five equations. The association is also partly due to the fact that respondents with HPQ ratings of 0 to 7 have higher odds of low performance than those with ratings of 8 and that respondents with ratings of 8 (with the exception of customer service representatives) have higher odds of low performance than those with ratings of 9.

The outcome with the widest range of odds between workers with high and low HPQ ratings (12.3:1) is the measure of railroad engineer performance action. The prediction equation for this outcome could only be estimated if we constrained the odds to be the same among respondents with HPQ ratings in the range 8 to 10 as a result of the fact that this is a rare outcome that is largely confined to engineers who rate themselves at the low end of the HPQ scale. The outcome with the narrowest range of odds, in comparison, is reservation agent performance (3.2:1). As noted above in the description

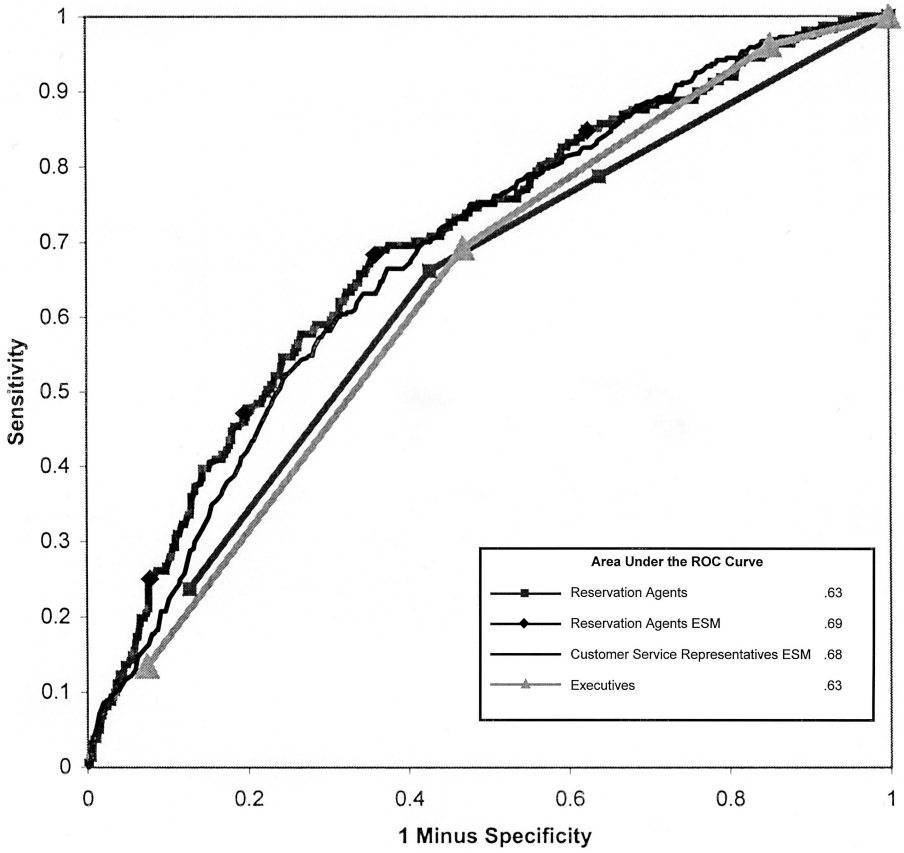

Fig. 1. Receiver operator characteristic curves for HPQ Global Ratings predicting archival and ESM measures of low work performance.

of the archival measures, this measure has the narrowest range of ratings as well (79 to 100 on the 0 to 100 scale). It is conceivable that this restricted range introduced imprecision into the definition of low performance, resulting in a dampening of the OR associated with low HPQ ratings for this outcome. The ORs associated with HPQ ratings of 8 across the remaining outcomes are all lower than those associated with ratings of 0 to 7. The ORs associated with HPQ ratings of 9 for these equations are generally lower than those associated with ratings of 8. The ROC curves for the strength of the HPQ ratings in predicting the archival and ESM outcomes are shown in Fig. 1. Areas under the ROC curve, which can be interpreted as the proportion of times randomly selected workers with low work performance could be distinguished from other workers based on differential HPQ ratings, range between 0.63 and 0.69 across the samples.

The results of logistic regression analyses linking HPQ ratings with the four archival or ESM measures of high work performance are presented in Table 6. There is a statistically significant relationship between HPQ ratings and the odds of high archival or ESM work performance in the equations for reservation agents and customer service representatives, but not in the one equation for executives. The equations in which the outcomes are the ESM scores show consistent monotonicity of odds. The equation in which the reservation agent supervisor ratings are the outcome, in comparison, shows a significant distinction between 0–7 and 8 to 10 ($\chi^2_1 = 6.4$, $P = 0.015$), but no meaningful variation in odds among respondents with HPQ ratings of 8, 9, or 10 ($\chi^2_2 = 2.5$, $P = 0.701$). This restricted range of ORs in predicting reservation agent supervisor ratings among respondents with HPQ ratings in the range 8 to 10 is similar to the pattern seen in Table 5. The ROC curves for the strength of the HPQ ratings scale in the three statistically significant equations are shown in Fig. 2. Areas under the ROC curve range between 0.59 and 0.72 across the samples.

We also evaluated the effects of the component questions about work performance that were administered in the survey before the global 0-to-10 work performance rating. As noted earlier in the report, these included questions about quantity of work, quality of work, interpersonal aspects of work, work successes and failures, and work-related accidents-injuries. Factor analysis showed that these measures, like other recently developed multi-item inventories of self-reported work performance,[34] form meaningful factors with good internal consistency reliabilities.

We found that these factors are significantly related to the archival and ESM work performance measures when considered one at a time. However, multivariate analyses

TABLE 6
Associations of HPQ Global Ratings with Highest 20 Percent of Archival and ESM Work Performance Outcome Measures

	Reservation Agent Supervisor Ratings[1]		Reservation Agent ESM		Customer Service Representative ESM		Executive Leadership Scores[1]	
Objective ratings Low work performance[2]	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
0–7	1.0	—	1.0	—	1.0	—	1.0	—
8	5.7*	(1.6–20.1)	3.7*	(1.3–10.7)	2.5*	(2.8–10.4)	1.0	(0.3–3.3)
9	3.8*	(1.1–13.1)	4.4*	(1.5–13.1)	5.5*	(2.8–10.8)	1.4	(0.4–4.6)
10	5.4*	(1.6–19.4)	6.4*	(1.8–22.2)	45.8*	(11.4–184.7)	1.0	(0.3–4.2)
χ^2_3	8.92*		38.8*		28.7*		1.09	
(n)	(441)		(816)		(1,926)		(269)	

* Significant at the .05 level, two-sided test
[1] Based on 30-day job performance rating
[2] Model includes sex, age, day of study, and beep number as controls

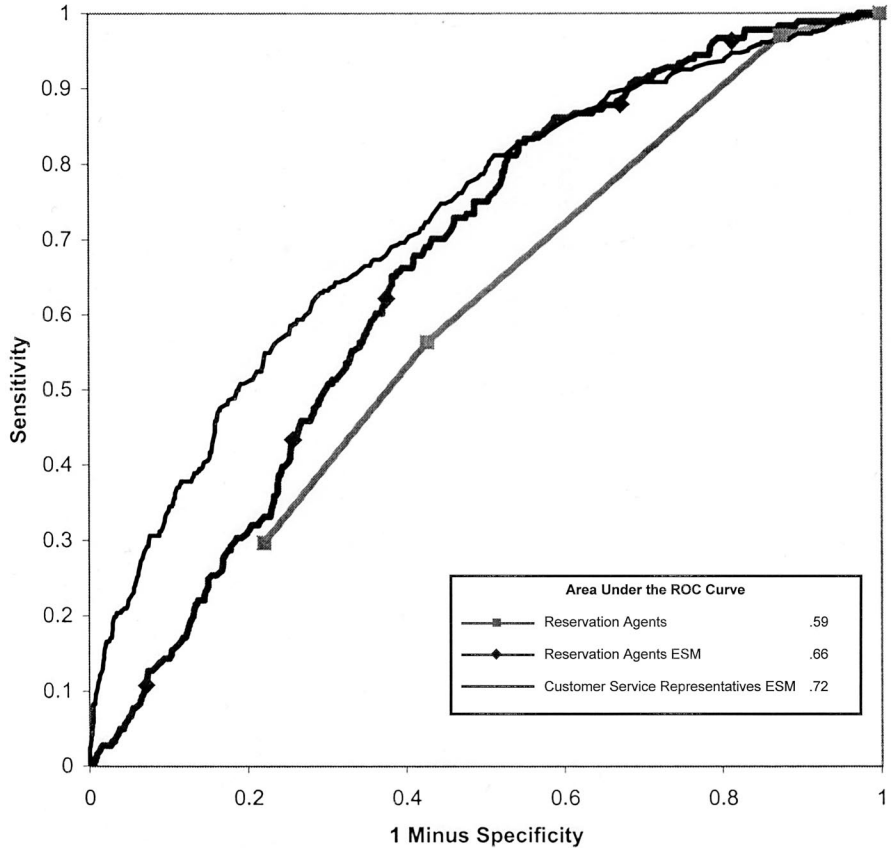

Fig. 2. Receiver operator characteristic curves for HPQ global ratings predicting archival and ESM measures of high work performance.

showed that they were generally not significant predictors of archival or ESM measures of work performance in a prediction equation that controlled the effects of the global 0-to-10 work performance rating. This means that the global rating out-performances the factor scales and, with one exception noted in the next paragraph, captures the effects of these scales. Why? The most likely reason is that the factor scales tap generic aspects of work that might vary in their importance for overall work performance across occupations and that doubtlessly fail to tap all relevant aspects of work for an assessment of performance on these jobs. The 0-to-10 self-rating, in comparison, asks the respondent, implicitly to use their knowledge of the relevant considerations in evaluating work performance to arrive at a global self-rating based on an assess-

TABLE 7
Associations of self-reported work absence with payroll work absence among reservation agents and railroad engineers

	Pearson Correlation	Means		Z-test
		Self-Report	Payroll	
I. Reservation Agents				
A. One-week recall (n = 414)				
Hours worked	.87	28.7	25.7	2.9*
Hours missed	.81	3.8	6.8	3.3*
B. Four-week recall (n = 414)				
Hours worked	.79	113.4	107.0	1.8
Hours missed	.71	27.3	33.7	2.1*
II. Railroad Engineers				
A. One-week recall (n = 847)				
Days missed	.61	1.1	1.5	3.6*
B. Four-week recall (n = 847)				
Days missed	.66	4.8	6.1	3.4*

* Self-report significantly different from payroll at the .05 level, two-sided test.

ment of all these considerations. Apparently workers are able to do a better job of this than we were able to do in developing our factor scales.

The only exception to this general statement was in the case of railroad engineers, where the component question about work failure was a significant predictor of the archival outcome measure. This was presumably the case because both the predictor and the archival outcome measure were highly skewed dichotomies and the archival measure was heavily influenced by the discovery of major performance errors that are presumably tapped more directly in the dichotomous question about failure than in the 0-to-10 rating.

The Associations of HPQ Absenteeism Reports with Payroll Records

Linear correlations and comparisons of means between HPQ self-reports about absenteeism and employer payroll records are reported in Table 7. In the case of reservation agents, the comparisons are for hours worked and hours missed over both one-week and four-week recall periods. These correlations are substantial in magnitude and higher for one-week (0.81–0.87) than 4-week (0.71–079) recall. Self-reports consistently overestimate hours worked and underestimate hours missed. Although these biases are fairly modest in absolute terms (1.5 to 3 hours per week), they represent rather substantial proportional underestimations of hours missed (19–44%). In the case of the railroad engineers, the comparisons are for days missed. These correlations are also substantial in magnitude (0.61–0.66), although somewhat smaller than the correlations for reservation agents. As with reservation agents, engineers underestimate absence by 0.3 to 0.4 days per week, which represent 21 to 26% underestimations of absence compared to payroll records. Unlike reservation agents, though, 4-week recall is as accurate as 1-week recall for railroad engineers.

It is interesting to note that four-week estimates of absence are more than four times as large as 1-week estimates both for reservation agents and for railroad engineers. Similarly, 4-week estimates of hours worked are less than four times 1-week estimates. This is not caused by recall bias, as proven by the fact that the same patterns exist in employer payroll records. The reason is that respondents were allowed to postpone the start date of their ESM data collection if it was inconvenient for them to begin on the date selected by the research team. Debriefing showed that these postponements were often because of short-term illness. This means that the ESM weeks are downwardly biased in estimating the prevalence of absenteeism. The fact that this shows up not only in the means of the payroll record data but also in the means of the self-reports is an additional indirect indicator of the accuracy of the self-report data.

The results of logistic regression analyses of moment-in-time data on being at work versus not at work in the ESM person-time samples are reported in Table 8. The dependent variable is a dichotomy for whether the respondent was at work on not at the random moment-in-time, whereas the predictors are standardized HPQ self-reports obtained in the post-ESM debriefing interview of hours worked during the ESM week and days missed during the week. The results are consistent in the separate samples of reservation agents and customer service representatives in showing statistically significant associations between HPQ self-reports and the moment-in-time ESM data. A one standard deviation increase in the HPQ self-reported measure of hours worked during the ESM week is associated with a doubling of the relative-odds that a respondent will be at work during any randomly selected moment-in-time during that week. A one standard deviation increase in the HPQ self-report measure of hours worked

TABLE 8
Associations of self-reported hours/days worked with odds of being at work at randomly selected times in the ESM samples[1]

	Reservation Agents		Customer Service Representatives	
HPQ Self-Reports[2]	OR	(95% CI)	OR	(95% CI)
Hours worked	2.0*	(1.6–2.3)	2.0*	(1.6–2.5)
Days missed	0.5*	(0.4–0.7)	0.5*	(0.4–0.6)
(n_p)[3]	(105)		(181)	
(n_b)[3]	(3,675)		(6,335)	

* Significant at the .05 level, two sided test.

[1] Results are based on a two-level (person and 35 random moments of time within persons) mixed regression model that controlled both for within-person variables (number of days in the study at the time of the beep, ranging between 1–7; sequence of the beep within the day, ranging between 1–5) and for between-person variables (age, sex). Hours worked and days missed were included in separate models and were treated as between-person variables.

[2] The work measures were standardized to a mean of zero and variance of one.

[3] n_p = number of people; n_b = number of beeps.

during the ESM week, in comparison, is associated with a halving of the relative-odds of being at work during any randomly selected moment-in-time during that week.

Discussion

The results reported here document that the HPQ generates meaningful measures of work performance and absenteeism. The only negative result is the failure of the HPQ to predict high work performance among automobile executives. As these executives are the only white-collar workers in the sample, this failure might reflect a general weakness of the HPQ in predicting high performance among white-collar workers. Replication of the calibration study in other white-collar occupations is needed to evaluate this interpretation. Before that time, though, the HPQ should be seen as useful for white-collar workers largely in assessing low performance rather than high performance.

This usefulness of the HPQ in evaluating work performance is as a global measure, as no component measures are included in the scale. The HPQ can be used to assess the overall effects of allergies, migraine, and other illnesses on overall work performance in an entire workforce and, comparatively, across different types of occupations. It cannot tell us, though, what aspects of performance are affected by these illnesses (eg, motor skills, concentration etc.). As noted in the introduction, the decision to focus on global performance rather than on components is based on our interest to monetize the workplace costs of illness and the cost savings of health care interventions. The estimation of these monetary effects is much more easily achieved by assessing global work performance rather than selected components of performance. Monetizing component effects requires the researcher to determine the importance for overall work performance of, say, a decrement in ability to concentrate to a ditch digger or of a decrement in ability to lift heavy objects to a lawyer. We decided that these evaluations were better left to the worker-respondents themselves in arriving at global assessments of their overall work performance. Monetizing component effects also requires the researcher to assume that the components measured fully capture all relevant aspects of work that go into the creation of work performance. Given the enormous variety of work functions known to exist in the labor force, we were unwilling to make this assumption, preferring to allow respondents themselves to consider all functions that they consider relevant to the specific requirements of their jobs in making global evaluations of their performance.

Sensitivity of the HPQ Measures

In light of the fact that the HPQ work performance measure is relatively coarse, a question can be raised whether it is sensitive enough to detect effects of illnesses on work performance and of health interventions with moderate effects on the reduction in impaired work performance. It goes beyond the scope of the present report to present substantive results. However, in light of this important concern it is worth noting that substantive analyses of the data presented here, which will be reported in separate publications, show that the HPQ measure of work performance is sensitive to a variety of illnesses as well as to standard disorder-specific measures of illness severity within subsamples of respondents who suffer from specific disorders. There are also statistically and substantively significant associations of HPQ work absence and accident-injury measures with information collected in the surveys about the prevalences and severities of disorders.

Calibrating HPQ Global Ratings

Calibration rules were developed to convert ratings on the 0-to-10 HPQ global rating scale into predicted probabilities of high and low archival and ESM work performance using the results reported above. This was done bearing in mind that the predicted probabilities of high performance should not be evaluated for white-collar workers. These calibration rules were based on methods developed to promote the use of diagnostic screening scales in clinical decision-making.[51] These methods allow scores on screening scales like the HPQ to be interpreted in new samples by using the results of prediction equations developed in calibration samples to assign probability-of-illness scores (or, in the present case, probabilities of high and low work performance) to individual cases in the new samples.

The difficulty in developing these rules is that we cannot assume that the probabilities of high and low performance associated with a given HPQ score in the calibration sample (positive and negative predictive values) will be the same in new samples. This is true, importantly, even if the conditional distributions of HPQ ratings among people with true high and low performance (sensitivity and specificity) are constant across samples, because any deviation in the proportions of workers in the new samples with actual high and low performance from the 20% arbitrarily assumed in the HPQ calibration samples will lead to changes in the positive and negative predictive values at a given level of the HPQ rating.[52] As a result, it is not appropriate to specify a single threshold for the outcomes of interest (in this case, high and low work performance) for all populations based on given HPQ ratings.

This problem can be addressed in three ways, all of which are implemented in software developed for use with HPQ survey data. All three approaches rely on the method of stratum-specific likelihood-ratios (SSLRs). An SSLR is an odds-ratio that compares respondents with a specific score on a screening scale (in the case of the HPQ global rating, a rating of either 0–7,8,9, or 10) with those having all other ratings on the scale in terms of their odds of having a dichotomous outcome (in the case of the HPQ calibration, either low performance versus others or high performance versus others).[51] The assumption that the sensitivities and specificities of the relationship between HPQ ratings and true performance categorization are the same in new samples as in the calibration samples is equivalent to the assumption that SSLR's are constant across samples. Based on this assumption, the SSLRs estimated in the calibration samples can be used in conjunction with information about the true prevalence (converted to an odds) of the dichotomous outcome in the new sample to calculate individual-level predicted probabilities of the dichotomous outcome. This can be done by showing, based on Bayes' theorem, that

$$POO \times SSLR = ROO \qquad (1)$$

where POO = the population odds of the dichotomous outcome and ROO = the individual respondent's odds of the outcome. The individual's probability of the dichotomous outcome, p, can easily be derived from ROO by using the transformation

$$ROO = p/(1-p) \qquad (2)$$

There are three ways to use the results in Eqs. (1) and (2) to assign individual-level predicted probabilities of high and low work performance based on individual HPQ ratings. As noted above, all three are implemented in the software developed for use in HPQ surveys.

The first way to assign individual-level probabilities is to fix the aggregate prevalences of high and low work performance to 20% in new samples a priori in exactly the same way as in the calibration samples. High and low performance, in this approach, are arbitrarily defined fixed percentiles. The second way is to allow the estimated prevalences of high and low work performance to be fixed at different values based on external information and institutional knowledge. For example, managers in a particular corporation might conclude that the prevalences of high and low work performance, respectively, are 30 and 10% among their white-collar workers, 20 and 20% among their pink-collar workers, and 10 and 30% among their blue-collar workers. These assumptions can be used to convert HPQ global ratings into individual-level predicted probabilities of high and low performance separately in each occupational sub-sample by using the transformation in Eqs. (1) and (2). The third way to assign individual-level probabilities is to estimate the prevalences of high and low work performance empirically from the distribution of HPQ ratings in the new sample. This can be done by using maximum-likelihood to compare empirical distributions on this scale with the theoretical distributions generated by the sensitivities and specificities in the calibration sample applied to all logically possible combinations of high and low work performance. The maximum-likelihood estimates of the prevalences of high and low performance based on this approach are those associated with the theoretical distribution of HPQ global ratings most similar to the empirical distribution in the sample. Once these prevalence estimates are identified, they can be converted to odds and used in Eq. (2) to generate individual-level probabilities from individual-level HPQ ratings.

Monetizing Absenteeism and Work Performance Ratings

It is important to remember that absenteeism and low work performance have quite different costs across occupations and industries.

These differences are not necessarily proportional to salaries. The unscheduled absence of an airline reservation agent, for example, might lead to customers spending somewhat more time waiting before they speak to an agent and to agents on duty having a somewhat more hectic day than usual. But the costs of these inconveniences to the employer are probably minimal unless the delays are so long and persistent that customers go to a different airline to purchase their tickets. The unscheduled absence of an airline stewardess, in comparison, can cause a flight departure delay, due to FAA staffing requirements for number of personnel needed for a flight to depart, that costs the airline at much as $5000 per hour in additional gate fees. Similarly, the low performance of a salesman, if it leads to the loss of a major contract, can cost a corporation millions of dollars even though the salesman's salary is only a fraction of that amount. Because of situations such as these, even when we have good estimates of absenteeism and work performance, an additional step is required to estimate the monetary costs of poor performance and absenteeism to the employer. A number of approaches have been proposed to make these estimates,[6] several of which have been implemented in software developed for use in analyzing HPQ surveys. Although it is beyond the scope of this report to discuss these approaches here, it is important to note that this additional step is needed to monetize the HPQ results.

Future Directions

Nationally representative general population HPQ surveys are currently being performed in 28 countries around the world as part of a larger WHO initiative aimed at estimating the societal costs of mental and physical illness.[53] We anticipate that over 200,000 respondents will complete these HPQ surveys once they are finished. In addition, both paper-pencil and internet versions of the HPQ have been developed and are being used to carry out ongoing annual surveys of the employees of a number of large corporations in the U.S., either as part of existing Health Risk Appraisal surveys or as standalone surveys. A number of these surveys are being carried out in collaboration with the National Business Coalition on Health. We will soon be distributing an HPQ survey toolkit to all members of the National Business Coalition on Health throughout the United States. A number of other collaborations are also in development in the United States, Canada, and Europe.

The dissemination of HPQ surveys is the first step in a larger program of research aimed at pinpointing health problems that are associated with high indirect workplace costs, developing and evaluating interventions to reduce these costs, and establishing quality assurance procedures to monitor the success of efforts to disseminate and maintain these interventions. In order to implement this program of research, it is necessary to begin by linking HPQ absenteeism, work performance, and work-related accident-injury reports to information about specific health problems. This is done in the HPQ surveys by asking respondents if they suffer from a number of common chronic conditions and, if so, whether they are currently under professional treatment for these conditions. The chronic conditions checklist in the HPQ interview schedule is based on the checklist used in the U.S. National Health Interview Survey (NHIS). Data from the NHIS and a number of other nationally representative general population surveys were analyzed to select the conditions in the HPQ checklist. The criteria for selection were that the conditions are commonly occurring among working people and are associated either with excess work absence, low work performance, or elevated rates of work-related accidents-injuries.[54–56] We also included in the HPQ surveys an acute symptoms checklist developed by Khroenke et al[57] to capture the most common complaints of acute-care patients in primary care treatment.

It is noteworthy that the HPQ survey distinguishes between health problems that are being treated and those that are not being treated. This distinction is important because most common health problems that influence workplace functioning vary widely in severity. Some people with seasonal allergies, for example, have very mild symptoms while other seasonal allergy sufferers have very severe symptoms. People with severe symptoms are more likely to be in treatment than those with mild symptoms. This makes it is impossible to determine whether low rates of treatment should be considered a problem from the perspective of employers in the absence of separate analyses to determine whether untreated cases are associated with impairments in work performance. This is performed in standardized analyses of HPQ data by distinguishing the separate effects of treated conditions and untreated conditions on workplace outcomes. A single yes or no question about treatment of each health problem is included in the HPQ for this purpose. More extensive questions were considered for inclusion in the surveys, but subsequently rejected based on the realization that detailed information about the treatment of specific health problems could be obtained by employers from health claims records. The HPQ treatment question asks about "professional" treatment, defined as treatment by a doctor, nurse, or other health professional, to exclude self-treatment and complementary and alternative medical treatment not provided by a health professional. These exclusions are important in light of the growing importance of self-treatment and complementary and alternative medical treatment.[5,58]

The fact that treatment is strongly influenced by illness severity means that cross-sectional HPQ surveys cannot be used to help employers

estimate the likely return on their investment (ROI) because of expanding treatment for a particular condition. Experimental or quasi-experimental studies are required to make such estimates. Cross-sectional HPQ surveys are better suited to address the prior questions: 1) Which of the health problems assessed in the HPQ survey have the greatest indirect costs in my workforce? 2) Are these costs associated with low rates of treatment (ie, high workplace costs among untreated workers with the conditions), inadequate treatment (ie, high workplace costs among treated workers with the conditions), or both? 3) How do the indirect workplace costs of target illnesses in my workforce compare with those in other benchmark populations?

Answers to these questions can help employers pinpoint health problems that have particularly high indirect workplace costs. Systematic reviews of the treatment effectiveness literature can then be used to evaluate the likelihood that enhanced outreach and/or treatment efforts aimed at these conditions would yield a large enough reduction in workplace costs to have a positive return on investment. Ongoing HPQ monitoring surveys can then be used to calculate ROI of new interventions based on such considerations. This can be done in a single corporation with a universal intervention using a before-after interrupted time series design or a quasi-experimental case-control design that compares changes in the HPQ ratings among workers with the target conditions in corporation that do, versus do not, implement the intervention. In a corporation that has multiple sites and that can assign new treatment programs to a subset of these sites, a more powerful before-after case versus control test market design can be used to evaluate the ROI of the intervention.

A large experimental treatment effectiveness trial is currently underway in conjunction with ongoing HPQ surveys that illustrates some of these ideas about the evaluation of treatment interventions. This trial is evaluating the effects of detecting and treating working people with major depression.[11,21,59] The intervention features outreach and best-practices treatment provided by United Behavioral Health (UBH), one of the largest behavioral health carve-out companies in the country. Baseline HPQ surveys are being used to screen for major depression among workers with UBH coverage in a number of large corporations. UBH care managers are implementing an outreach and treatment program to a random sub-sample of these workers using an intent-to-treat experimental design. Expanded follow-up HPQ surveys are being used to evaluate the return on investment (ROI) of the intervention over a 2-year follow-up period. Our hope is that this experiment will serve as a model for future interventions and evaluations using the HPQ.

Acknowledgments

The authors gratefully acknowledge the help of questionnaire wording experts Stephanie Chardoul, Lou Magilavy, Beth Ellen Pennell, Kent Peterson, Tom Wilkinson, and Debbie Zivan in developing and revising the HPQ, the assistance of Bill Whitmer and his colleagues at the Health Enhancement Research Organization (HERO) in recruiting companies to participate in the calibration survey, the staff of the participating companies in facilitating the sample selection and abstraction of objective performance measures and absenteeism records, and Joe Myer, Tom Wilkinson, and their staff at DataStat, Inc. in carrying out the calibration survey. The complete text of the HPQ is posted at the http://www.hcp.med-.harvard.edu/hpq.

Supported by The World Health Organization, the John D. and Catherine T. MacArthur Foundation, and by unrestricted educational grants From Glaxo-Wellcome, Pfizer, Searle and Schering-Plough.

References

1. Sloan FA. *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*. New York: Cambridge University Press; 1996.
2. Gold MR, Siegel JE, Russell LB, et al. *Cost-Effectiveness in Health and Medicine*. Oxford: Oxford University Press; 1996.
3. Patrick DL, Erickson P. *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resourced Allocation.* New York: Oxford Univerity Press; 1993.
4. Tarlov AR, Ware JE Jr., Greenfield S, et al. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *JAMA.* 1989;262: 925–930.
5. Kessler RC, Davis RB, Foster DF, et al. Long-term trends in the use of complementary and alternative medical therapies in the United States. *Ann Intern Med.* 2001;135:262–268.
6. Pauly MV, Nicholson S, Xu J, et al. A general model of the impact of absenteeism on employers and employees. *Health Econ.* 2002;11:221–231.
7. Greenberg PE, Kessler RC, Nells TL, et al. Depression in the workplace: an economic perspective. In: Feighner JP, Boyer WF, eds. *Selective Serotonin Reuptake Inhibitors: Advances in Basic Research and Clinical Practice.* New York: John Wiley & Sons Ltd.; 1996.
8. Kessler RC, Zhao S, Katz SJ, et al. Past-year use of outpatient services for psychiatric problems in the National Comorbidity Survey. *Am J Psychiatry*. 1999;156:115–123.
9. Coulehan JL, Schulberg HC, Block MR, et al. Treating depressed primary care patients improves their physical, mental, and social functioning. *Arch Intern Med.* 1997;157:1113–1120.
10. Wells KB, Sherbourne C, Schoenbaum M, et al. Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *JAMA.* 2000;283:212–220.
11. Kessler RC, Barber CB, Birnbaum HG, et al. Depression in the workplace: effects on short-term work disability. *Health Aff.* 1999;18:163–171.
12. Boonen A, van der Heijde D, Landewe R, et al. Work status and productivity costs due to ankylosing spondylitis: comparison of three European countries. *Ann Rheum Dis.* 2002;61:429–437.
13. Reginster JY, Khaltaev NG. Introduction and WHO perspective on the global bur-

den of musculoskeletal conditions. *Rheumatology.* 2002;41:1–2.
14. Kessler RC, Almeida DM, Berglund P, et al. Pollen and mold exposure impairs the work performance of employees with allergic rhinitis. *Ann Allergy Asthma Immunol.* 2001;87:289–295.
15. Rosenheck RA, Druss B, Stolar M, et al. Effect of declining mental health service use on employees of a large corporation. *Health Aff.* 1999;18:193–203.
16. Lynch W, Riedel JE. *Measuring Employee Productivity: A Guide to Self-Asessment Tools.* Scottsdale, AZ: The Institute for Health and Productivity Management; 2001.
17. Rehm J, Ustun TB, Saxena S, et al. On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *Int J Methods Psychatric Res.* 1999; 8:110–123.
18. World Health Organization. *International Classification of Functioning, Disability and Health.* Geneva: World Health Organization; 2001.
19. Converse JM, Presser S. Survey Questions: Handcrafting the Standardized Questionnaire. In: Sage University Paper Series, 63:*Quantitative Applications in the Social Sciences.* Beverly Hills: Sage; 1986.
20. Sudman S, Bradburn NM, Schwartz N. *Thinking about Answers: The Applications of Cognitive Processes to Survey Methodology.* San Francisco, CA: Jossey-Bass; 1996.
21. Wang PS, Beck AL, McKenas DK, et al. Effects of efforts to increase response rates on a workplace chronic condition screening survey. *Med Care.* 2002;40: 752–760.
22. Blum TC, Roman PM. Alcohol consumption and work performance. *J Stud Alcohol.* 1993;54:61–70.
23. Harbour JL. *The Basics of Performance Measurement.* New York: Productivity, Inc.; 1997.
24. Grote D. *The Complete Guide to Performance Appraisal.* New York: AMACOM; 1996:400.
25. U. S. Department of Labor Employment and Training Administration. *Testing and Assessment: An Employer's Guide to Good Practices.* Washington, DC: U. S. Government Printing Office; 1999.
26. Matheson LN, Kaskutas V, McCowan S, et al. Development of a database of functional assessment measures related to work disability. *J Occup Rehabil.* 2001; 11:177–199.
27. Holloway J, Lewis J, Mallory G. *Performance Measurement and Evaluation.* Thousand Oaks, Ca: Sage; 1995.
28. Pritchard RD, Holling H, Lammers F, et al. *Improving Organizational Performance with the Productivity Measurement and Enhancement System: An International Collaboration.* Huntington, NY: Nova Science; 2002.
29. Whetzel DL, Wheaton GR. *Applied Measurement Methods in Industrial Psychology.* Cleveland: Davis-Black; 1997.
30. Neal A, Griffen MA, Paterson J, et al. Development of measures of situational awareness, task performance, and contextual performance in air traffic control. In: Lowe AR, Hayward BJ, eds. *Aviation Resource Management.* Aldershot, UK: Ashgate; 2000:241–267.
31. Sawyer JE, Latham WR, Pritchard RD, et al. Analysis of work group productivity in an applied setting: Application of a time series panel design. *Personnel Psychol.* 1999;52:927–967.
32. Warr PB, Conner MT. *The Measurement of Personal Effectiveness for Review and Guidance.* London: Department of Education and Employment; 1999.
33. Endicott J, Nee J. Endicott Work Productivity Scale (EWPS): a new measure to assess treatment effects. *Psychopharmacol Bull.* 1997;33:13–16.
34. Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: health status and employee productivity. *J Occup Environ Med.* 2002;44:14–20.
35. Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revisions of a health status measure. *Med Care.* 1981;19:787–805.
36. Lerner D, Amick RC, Rogers WH, et al. The Work Limitations Questionnaire. *Med Care.* 2001;39:72–85.
37. U. S. Department of Labor. *Dictionary of Occupational Titles.* Washington, DC: Labor Dept., Employment and Training Administration, United States Employment Service; 1991:1445.
38. Cain P. An assessment of the Dictionary of Occupational Titles as a source of occupational information. In: Miller AR, Treiman DJ, Cain PS, et al. (Eds.), *Work, Jobs, and Occupations: A Critical Review of the Dictionary of Occupational Titles.* Washington: National Academy Press; 1980:148–195.
39. Cullum CM, Saine K, Chan LD et al. Performance-based instrument to assess functional capacity in dementia: The Texas Functional Living Scale. *Neuropsychiatry Neuropsychol Behav Neurol.* 2001;14:103–108.
40. Whetstone LM, Fozard JL, Metter EJ, et al. The physical functioning inventory: a procedure for assessing physical function in adults. *J Aging Health.* 2001;13:467–493.
41. Means B, Loftus EF. When personal history repeats itself: Decomposing memories for recurring events. *Appl Cognit Psychol.* 1991;5:297–318.
42. Menon A. Judgements of behavioral frequencies: Memory search and retrieval strategies. In: Schwartz N, Sudman S, eds. *Autobiographical Memory and the Validity of Retrospective Reports.* New York: Springer-Verlag; 1994:161–172.
43. Oksenberg L, Vinokur A, Cannell CF. Effects of commitment to being a good respondent on interview performance. In: Cannell CF, Oksenberg L, Converse JM, eds. *Experiments in Interviewing Techniques. DHEW Publication No. (HRA) 78–3204.* Washington: Department of Health, Education, and Welfare; 1979: 74–108.
44. Schwarz N. Self-reports: how the questions shape the answers. *Am Psychol.* 1999;54:93–105.
45. Revecki DA, Irwin D, Reblando J, et al. The accuracy of self-reported disability days. *Med Care.* 1994;32:401–404.
46. Csikszentmihalyi M, Larson R. Validity and reliability of the Experience Sampling Method. In: deVries M, ed. *The Experience of Psychopathogy*: Cambridge University Press; 1992:43–57.
47. Stone AA, Kessler RC, Haythornthwaite JA. Measuring daily events and experiences: decisions for the researcher. *J Pers.* 1991;59:575–607.
48. Campbell D. *Campbell Leadership Index.* Greensboro, NC: Center for Creative Leadership; 1991.
49. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148: 829–843.
50. Peirce JC, Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. *Med Decis Making.* 1993;13:141–151.
51. Guyatt G, Rennie D. *User's Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* Chicago: AMA Press; 2001.
52. Rothman KJ. *Epidemiology: An Introduction.* New York: Oxford University Press; 2002:223.
53. Kessler RC, Ustun TB. The World Health Organization World Mental Health 2000 Initiative. *Hosp Manag Int.* 2000;195–196.
54. Kessler RC, Frank RG. The impact of psychiatric disorders on work loss days. *Psychol Med.* 1997;27:861–873.
55. Kessler RC, Greenberg PE, Mickelson KD, et al. The effects of chronic medical conditions on work loss and work cut-

back. *J Occup Environ Med.* 2001;43: 218–225.
56. Kessler RC, Mickelson KD, Barber CB, et al. The association between chronic medical conditions and work impairment. In: Rossi AS, ed. *Caring and Doing for Others: Social Responsibility in the Domains of Family, Work, and Community.* Chicago, IL: University of Chicago Press; 2001: 403–426.
57. Kroenke K, Spitzer RL, Williams JB. The PHQ-15: validity of a new measure for evaluating the severity of somatic symptoms. *Psychosom Med.* 2002;64: 258–266.
58. Eisenberg DM, Kessler RC, Van Rompay MI, et al. Perceptions about complementary therapies relative to conventional therapies among adults who use both: Results from a national survey. *Ann Intern Med.* 2001;135:344–351.
59. Simon GE, Barber C, Birnbaum HG, et al. Depression and work productivity: the comparative costs of treatment versus nontreatment. *J Occup Environ Med.* 2001;43:2–9.