Using the World Health Organization Health and Work Performance Questionnaire (HPQ) to Evaluate the Indirect Workplace Costs of Illness

Ronald C. Kessler, PhD
Minnie Ames, PhD
Pamela A. Hymel, MD, MPH
Ronald Loeppke, MD, MPH
David K. McKenas, MD, MPH
Dennis E. Richling, MD
Paul E. Stang, PhD
T. Bedirhan Ustun, MD

This report presents an overview of methodological issues in estimating the indirect workplace costs of illness from data obtained in employee surveys using the World Health Organization Health and Work Performance Questionnaire (HPQ). The HPQ is a brief self-report questionnaire that obtains three types of information: screening information about the prevalence and treatment of commonly occurring health problems; information about three types of workplace consequences (sickness absence, presenteeism, and critical incidents); and basic demographic information. The report considers two sets of methodological issues. The first set deals with measurement. The rationale for the HPQ approach to measurement is described in this section. In addition, data are presented regarding the accuracy of HPQ measures, documenting that the HPQ has excellent reliability, validity, and sensitivity to change. The second set of methodological issues deals with data analysis. A number of analysis problems are reviewed that arise in using self-report nonexperimental survey data to estimate the workplace costs of illness and the cost-effectiveness of treatment. Innovative data analysis strategies are described to address these problems. (J Occup Environ Med. *2004;46:S23–S37*)

From Harvard Medical School, Boston, Massachusetts (Drs Kessler and Ames); Hughes Electrics (Dr Hymel); CorSolutions, Inc., Rosemont, Illinois (Dr Loeppke); Logos Medical Consulting Services (Dr McKenas); Midwest Business Group on Health (Dr Richling); Galt Associates (Dr Stang); and World Health Organization, Geneva, Switzerland (Dr Ustun).

Address correspondence to: Ronald C. Kessler, PhD, 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.0000126683.75201.c5

Double-digit inflation in health care costs has led many employers to consider such health care cost control strategies as defined contributions, medical savings accounts, increased employee contributions for health insurance, and reductions in benefits.[1,2] It is clear that such approaches could create short-term savings in the direct costs of health care, but their effects on indirect workplace costs are less clear. Case studies document that the direct cost savings of some workplace health care interventions can be swamped by increases in indirect costs associated with offset, sickness absence, and disability,[3] whereas other interventions realize genuine savings.[4] A clear understanding of the indirect workplace costs of illness as well as the costs of changes in health care benefits is consequently needed to make rational decisions about changes in benefit structure. Yet, few employers have access to the data required to obtain such an understanding, making it impossible for them to optimize their health care-purchasing decisions.

Three types of data gaps can be distinguished. First, few employers have access to good data on untreated health problems of their employees unless they conduct annual physical examinations with all workers in the workplace. Second, few employers have access to good data that can be used to assess either the magnitude of the impact of illness, especially untreated illness, on workplace functioning or the effects of changes in health care interventions on changes in workplace functioning. Third, even when such data are available, employers typically lack accurate evidence-based transformation rules that can be used to estimate the effects of changes in workplace functioning on the corporate bottom line. Researchers are working on all three of these gaps. A number of self-report questionnaires have been developed to measure the indirect costs of illness and treatment on work performance in an effort to make up for the absence of archival data. Lynch and Reidel[5] and Loeppke et al[6] published recent reviews that discuss the pros and cons of available measures. In addition, empirical studies are currently underway to develop industry-specific transformation rules that can be used to convert information about employee-level effects of illness and treatment on workplace functioning into aggregate estimates of effects on the corporate bottom line.[7]

The current report focuses on the first two of these three data gaps by presenting new methodological data concerning the World Health Organization (WHO) Health and Work Performance Questionnaire (HPQ), the most widely used of the self-report instruments developed to assess the indirect workplace costs of illness.[8] The HPQ is a short instrument (10 minute average administration time) that screens for the presence of commonly occurring health problems and their treatment, assesses the three main domains of workplace performance that are traditionally assessed by organizational and industrial psychologists (absenteeism, presenteeism, and critical incidents),[9,10] and obtains basic demographic and occupation information. The HPQ can be self-administered using paper and pencil, interactive voice response, or internet modes of data collection either in a cross-section survey or in before-after test market studies or workplace experiments. A clinical trials version of the HPQ is also available.

WHO developed the HPQ as part of their Global Burden of Disease Initiative, a program of research aimed at documenting the human capital costs of illness and the cost-effectiveness of diverse health care interventions.[11] The HPQ is one component in the WHO Disability Assessment Schedule (WHO-DAS),[12] a multifaceted self-report scale of role functioning created by WHO to assess the global burden of disease in each of the core domains of the newly revised International Classification of Functioning, Disability and Health.[13] For example, the WHO-DAS is a primary outcome in a series of WHO-coordinated nationally representative general population health surveys in 28 countries around the world with a combined sample size of over 200,000 respondents.[14] A US sample of nearly 10,000 respondents is included in this larger data set. In addition to the nationally representative HPQ data embedded in these surveys, an electronic version of the HPQ is being administered to hundreds of thousands of employees of large US corporations in conjunction with the HPQ Data Consortium (www.hpq.org). The unprecedented size and diversity of the master HPQ benchmark data set available to this consortium makes the HPQ all the more attractive for use in future workplace health and productivity surveys.

Data on the validity of the HPQ absenteeism and presenteeism measures have been presented previously in this journal,[8] as have data on the effects of various health problems on HPQ outcome scores in a number of employee health surveys.[15,16] However, these previous reports omitted information about two important sets of methodological issues that are the focus of the current report. The first concerns two measurement issues: the sensitivity of the HPQ work performance measures to change and the accuracy of the HPQ assessments of chronic and acute conditions. Both of these issues are discussed in the first section of the article. We also present new data in this section on the validity of the HPQ presenteeism scale that adds to the validity data that we reported previously in this journal.[8]

The second set of issues concerns problems in data analysis that arise in using self-report nonexperimental survey data to estimate the workplace costs of illness and the cost-effectiveness of treatment. Four such problems and proposed solutions are discussed in the second section of the article. These problems include the confounding effects of common causes, the role of risk adjustment in evaluating differences between health plans, the evaluation of treatment effects on work performance using nonexperimental data, and the implications of comorbidity for evaluating the effects of individual conditions on work performance.

Measurement

Absenteeism

Most health surveys 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. Methodological studies led to four refinements of this basic assessment approach in the HPQ. First, the HPQ not only asks about days, but also about hours of work. This was done based on the fact that a “day of work” means something quite different to a person who works a regular 9-to-5 5-day-a-week schedule versus the increasingly large number of people who work 4-day weeks, half-days on Fridays, split shifts, rotating shifts, and the like. Second, in addition to asking about expected hours of work and hours missed on sickness absence days, the HPQ asks about hours missed on workdays (ie, coming in late or going home early) because of the fact that a substantial proportion of missed work time occurs on days when people come to work. Third, the HPQ asks 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, rather than focus on sickness absence, the HPQ considers total hours absent for any reason (eg, holidays and personal days in addition to sick days) because more and more employers are using integrated benefit schemes that combine vacation and personal days and sickness absence days, making the distinction among these categories artificial.

Methodological studies reported previously in this *Journal*[8] show that these four refinements resulted in the HPQ assessment of absenteeism having good validity. These studies compared HPQ self-reports with employer payroll records in multiple occupations. Good concordance was found, with Pearson correlations of 0.61 to 0.81 for 7-day recall and 0.66 to 0.71 for 4-week (28-day) recall of hours worked, days worked, hours missed, and days missed. Despite the good concordance between self-reports and payroll records, a consistent tendency was found in these methodological studies for HPQ self-reports to be biased in the direction of suggesting that workers spent somewhat more hours and days at work than recorded in payroll records. Fortunately, however, it was found that this bias can be corrected with a simple regression-based recalibration of self-reports. This correction is built into HPQ calculations of absenteeism.

Before leaving the discussion of absenteeism, it should be noted that focusing on hours rather than days worked may be a point of concern because we would expect the latter to be remembered more accurately than the former. This is a legitimate concern, although, as noted above, another problem exists when we focus on days from the perspective of monetizing results because the term "day of work" is ambiguous for workers with complex work schedules. Fortunately, the HPQ asks a series of questions about days of work (days absent because of health problems, days absent for any other reason, days with reduced hours because of health problems, days with reduced hours for other reasons) as a memory priming aid before asking about hours of work. Therefore, it is possible to repeat analyses of HPQ data twice, once using days and the second time using hours as the unit of analysis, and compare results for consistency.

It is also noteworthy that concerns can be raised about the decision to focus on overall absenteeism rather than on sickness absence. Indeed, a reviewer of this article raised exactly this concern, suggesting that we might be biasing the analysis against finding an adverse effect of illness or an ameliorative effect of treatment by including vacations and other sorts of absenteeism in the outcome measure. This concern reflects a misunderstanding about the logic of comparison. To appreciate this logic, imagine the situation in which we have two kinds of absenteeism, sickness absence and vacation absence, each of which we measure separately, and that we are interested in the effects of a given health problem on these two outcomes. If the health problem is associated with, say, 6.5 sickness absence hours per month but has no effect on vacation absence, then a statistical analysis of the effect of the condition on a measure of overall absenteeism will yield an estimate of 6.5 hours. There will be no bias caused by adding vacation absence to sickness absence, as the mean for this type of absence will be the same for respondents with and without the health problem. The standard error (a measure of the precision of the estimate) of the estimate will be larger when we use overall absenteeism rather than sickness absenteeism as the outcome measure. With the large sample sizes we typically use in HPQ surveys, though, the increase in standard error is of no real importance.

It is also informative to consider an alternative scenario to the one in the last paragraph. Imagine that the health problem led to an increase in 2.5 vacation hours per month. This is not implausible because many salaried workers fail to take all their vacation time each year, and illness might influence the decision to take vacation time. If this is the case, then exclusive focus on sickness absence would lead to under-estimating the true effect of the health problem (ie, 6.5 + 2.5 = 9.0 hours) on overall absenteeism, whereas an analysis that treated overall absenteeism as the outcome would yield an accurate estimate. The same is true if workers with the health problem under investigation forego some vacation to make up for sickness absence, in which case an analysis that focused exclusively on sickness absence as the outcome would overestimate the impact of the health problem on work absenteeism. Our decision to examine the effects of health problems on total absenteeism rather than only on sickness absence is based on these considerations.

A related issue is our decision to ask about overall absenteeism rather than about absenteeism because of a particular health problem. The second of these questions commonly is used in studies that focus on particular health problems. This can lead to bias, however, because respondents are often inaccurate reporters about reasons for their work absence. This is especially true in the commonly occurring situation where the worker suffers from comorbid disorders (eg, allergies and arthritis). If a bad night's sleep is associated with both joint pain and allergy symptoms, leading the worker to stay home from work the next day, it is likely that the worker would answer "yes" either to the question "Did your allergies keep you from work today?" or to the question "Did your arthritis keep you from work today?"

Because of this bias, questions of the sort “Did your allergies keep you from work today?” generally overestimate the effects of individual conditions. More accurate estimates can be obtained by using statistical analysis to tease out the relative effects of comorbid conditions. In the example given in the last paragraph, for example, this could be done in the aggregate by comparing workers with allergies-only, arthritis-only, and both conditions. Our decision to use statistical analysis to estimate aggregate effects of conditions rather than to ask respondents, in effect, to do this statistical analysis in their heads in answering questions about the effects of individual conditions is based on this reasoning.

Presenteeism

Inadequate work performance, often referred to as “presenteeism,” obviously is more difficult to assess than absenteeism. Indeed, the decision to develop the HPQ was based largely on a failure to find an existing self-report measure of work performance that met the needs of WHO. Objective performance-based assessments or self-report measures of work performance that include questions tailored to the unique demands of a single occupation[17,18] are ideal in this regard for a single occupation. However, such measures cannot be used for broad-based studies across diverse occupations.

Another possibility is to develop self-report measures that include questions about difficulties in many different concrete aspects of performance in an effort to cover the job demands of all existing occupations. Such an approach could try to be comprehensive, as in the Department of Labor’s Occupational Information Network (O*NET) system of job classification,[19] which contains over 50 dimensions of job performance, or it could either sample or aggregate these dimensions.[20,21] Among the self-report work performance measures that are based on this approach are the Endicott Work Productivity Scale,[22] the Stanford Presenteeism Scale,[23] and the Work Limitations Questionnaire.[24]

Although measures like these are useful for documenting the ways specific health problems affect work performance (eg, by decreasing the abilities to lift, read, concentrate etc.), none of them either covers all the dimensions of job performance included in the O*NET system or samples these dimensions in a representative way that guarantees unbiased coverage across occupations. Furthermore, even if the dimensions were representative, it would be extremely difficult to calculate the overall indirect costs of illness to the employer from scale results because no rules exist to combine dimensional 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 dimensions to 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. Many differences such as these would have to be taken into account in combining domain performance scores into an overall work performance score that applies equally well to workers in all the thousands of occupations in the labor force.

Given the current intractability of the problem described in the last paragraph, researchers who are more interested in arriving at an overall evaluation of the effects of health problems on work performance than in documenting effects on separate dimensions have gone to the other extreme of asking workers to provide a single global rating of their overall work performance rather than to report difficulties in a number of separate domains of work functioning. This is often done using a 0-to-10 global rating scale of overall work performance, as in the widely used Work Productivity and Activity Impairment Questionnaire.[25] The underlying assumption in using this approach is that workers can do a better job than researchers of implicitly reviewing the various dimensions of work functioning to arrive at a summary rating of their overall job performance.

The HPQ uses the global rating approach described in the last paragraph to assess work performance. Respondents are asked to rate their overall work performance during the past four weeks on a 0-to-10 scale where 0 means the “worst possible work performance” a person could have on this job and 10 means “top work performance” on this job. Our reasoning in selecting this simple aggregate approach was the one mentioned in the last paragraph: 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.

In administering this global rating question in the HPQ, more concrete memory priming and decomposition questions are asked first. These questions are explicitly designed to be sufficiently general that they apply to all occupations, but sufficiently focused that they facilitate relevant memory search and review. The goal is to force respondents to review critical aspects of their work performance before assigning themselves a rating on the global scale. Methodological research has shown that forced reviews of this sort increase the accuracy of responses to global ratings questions.[26–28] In addition, internal anchoring questions are used in the HPQ to facilitate interpretation of responses to the global questions by asking each respondent to give separate global ratings for the average worker on their job and for their own usual performance before rating their recent performance. Responses to these questions allow scores of recent performance to be calculated in comparison to (ie, divided by) the performance of other workers as a way of adjusting for possible between-worker differences in calibration on the 0-to-10 self-anchoring scale.

It is noteworthy that the HPQ component questions can be modified to blend the best features of self-report measures with measures that, like the HPQ, use global ratings to assess overall work performance. For example, the short version of the Work Limitations Questionnaire (WLQ), which assesses difficulties in a small number of work performance domains that are highly related to ratings in the larger set of WLQ domains, could be included among the HPQ component questions. Data of this sort could be analyzed using a statistical method known as path analysis[29] to study the extent to which proximate effects of health problems on these concrete aspects of work performance mediate the more distal effects of these same health problems on global ratings of overall work performance.

Validity

Methodological studies that were described previously in this journal[8] show that the HPQ is a valid assessment of presenteeism. These studies compared HPQ self-reported presenteeism with independent employer records of job performance and found statistically significant monotonic associations across a range of occupations (airline reservation agents, customer service representatives, automobile company executives, railroad engineers) and a variety of outcomes (work audits, supervisor ratings, peer ratings). One additional study of the HPQ presenteeism scale was performed subsequently and is reported here for the first time. This involved 551 call-center workers who completed the HPQ in an internet survey and were independently rated by their supervisors on a 1–3 scale of being inadequate, adequate, or superior on six different dimensions of work performance. These ratings were then averaged across the six dimensions to arrive at a summary rating of overall work performance. As with all internet HPQ surveys, respondents were informed that their participation was completely voluntary, that they were free to skip any question they did not want to answer, and that their responses would be de-identified after matching with their archival supervisor ratings data and prior to data analysis. These consent procedures were approved by the Human Subjects Committee of Harvard Medical School.

As is often the case with supervisor ratings, the summary call-center worker ratings were strongly skewed to the upper end of the distribution, with only 7% of workers classified as inadequate and the vast majority classified as superior. To deal with this skewed distribution, two dichotomous versions of the summary supervisor ratings were created. The first distinguished workers in the top 20% of the distribution (high performers) from all other workers. The second distinguished workers in the bottom 20% of the distribution (low performers) from all other workers. The HPQ presenteeism scale was used to predict these two supervisor rating dichotomies in separate logistic regression equations. Results are presented in Table 1.

Part I of Table 1 shows that a trichotomized version of the HPQ presenteeism scale significantly predicts supervisor ratings of high performance, which we have defined inversely in the table as the absence of a high rating for ease of comparison with the results in Part II of the table. Part II shows a similar pattern in predicting supervisor ratings of low performance. The odds ratios

TABLE 1

HPQ Presenteeism Scores as Predictors of Supervisor Rating Among Call Center Workers ($n = 551$)†

	% of Sample		Supervisor Ratings		Odds Ratio	
	%	(se)	%	(se)	OR	(95% CI)
I. Supervisor rated not high performance						
Low Performer	4.2	(0.9)	91.3	(6.0)	5.0*	(1.1–22.1)
Medium Performer	72.1	(1.9)	82.6	(1.9)	2.2*	(1.4–3.5)
High Performer	23.8	(1.8)	67.9	(4.1)	1.0	—
$\chi^2 = 14.4$, p = 0.001						
II. Supervisor rated low performance						
Low Performer	9.6	(1.3)	30.2	(6.4)	3.9*	(1.6–9.6)
Medium Performer	74.0	(1.9)	18.9	(1.9)	2.1*	(1.0–4.4)
High Performer	16.3	(1.6)	10.0	(3.2)	1.0	—
$\chi^2 = 8.7$, p = 0.013						

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

† In Part I, HPQ low performance was defined as an absolute score on the 0-to-10 scale <7. High performance was defined as a ratio score ≥1.4. Medium performance was defined residually as all other respondents. In Part II, HPQ low performance was defined as an absolute score <6 or a relative score <.75. High performance was defined as an absolute score of 10 or a relative score ≥1.8. Medium performance was defined residually as all other respondents.

(ORs) across the HPQ categories of low, medium, and high self-reported performance are monotonic in each equation. Compared with workers with HPQ scores in the high-performance category, who were defined as the contract category with an OR of 1.0, workers with HPQ medium and low scores had ORs of 2.2 and 5.0, respectively, in predicting not having supervisor ratings of high performance. The comparable ORs in predicting supervisor ratings of low performance were 1.0, 2.1, and 3.9, respectively, for HPQ high, medium, and low performers. These associations are similar in magnitude to those found in the previously reported calibrations of the HPQ presenteeism scale against independent archival measures of job performance.[8]

Sensitivity to Change

Although the results described in the last paragraph document that the HPQ presenteeism scale is valid, a separate issue is whether it is sensitive to change. As noted above, the HPQ asks respondents to rate their typical work performance and then separately to rate their performance over the past 30 days. It is important that the latter report is distinct from the former report to the extent that performance does, in fact, vary across time. If this is not the case, then longitudinal tracking studies that use the HPQ as an outcome will not be sensitive to true change in performance. We know from empirical analyses of many HPQ surveys that the two scales are distinct in the sense that their correlation is less than perfect. Pearson correlations between the two scales are in the range 0.5 to 0.7 across all HPQ surveys. However, with data collected only at one point in time, we have no way of knowing whether the lack of a perfect association between the two scales is due to unreliability or to the recent performance scale truly being sensitive to change.

This issue can be resolved by conducting a prospective study in which the HPQ scale of recent performance is assessed on two separate occasions. Even here, though, it is impossible to separate true change from lack of reliability by considering the HPQ presenteeism scale alone because change and unreliability are confounded when only a single measure is assessed at two points in time.[30] This problem can be resolved when two indicator variables are measured at two points in time and a linear structural equation model is specified in which true work performance at time t (WP_t) is assumed to cause the two indicators (I_{1t}, I_{2t}) and the observed correlations among the indicators are assumed to be induced by WP_t. An illustration of such a model, in the form of a path diagram,[31] is presented in Fig. 1.

The problem with a single-indicator model can easily be seen in Fig. 1, as the Pearson correlation between either of the single indicators measured at two points in time is the product of three path coefficients (eg, $I_{11}I_{12} = acd$). It is impossible to identify any of these three unknowns from a single observed correlation. This is true even if we assume, as is conventional, that the reliability of the indicators is constant over time (eg, $a = d$). With two indicators at two points in time, this underidentification can be resolved by virtue of there being six correlations among the four observed measures and only five unknown parameters to estimate. In particular, the stability parameter, c, can be estimated as follows:

$$c = [I_{11}I_{12} \times I_{21}I_{12}/(I_{11}I_{21} \times I_{12}I_{22})]^{1/2} \qquad (1)$$

Once the stability parameter is estimated, the reliability parameters can be identified simply by assuming temporal constancy and factoring out the effect of instability on the test-retest correlations. For example, the reliability of I_{1t} can be estimated as

$$a = d = \{I_{11}I_{12}/[(\mathrm{I}_{11}\mathrm{I}_{22} \times I_{21}I_{12})/(I_{11}I_{21} \times I_{12}I_{22})]\}^{1/2} \qquad (2)$$

$$= [(acd/c]^{1/2} = (ad)^{1/2}$$

$$= (a^2)^{1/2} = (d^2)^{1/2} \qquad (2a)$$

Note that reliability is identified even if the second indicator, I_2, is actually not a single indicator measured at two points in time but two different indicators measured at different points in time. Note, too, that in the case of the assumption of temporal constancy, the model is overidentified with three degrees of freedom, in which case a χ^2_3 test can be used to evaluate model fit. Substantive interpretation of the parameter estimates depends on good model fit.

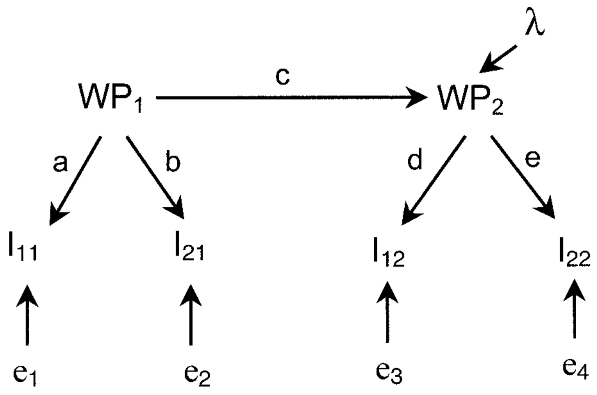

Fig. 1. A two-indicator two-time model of the stability of true presenteeism (WP_t) and the reliability of measured presenteeism indicators (I_{nt}). As in more conventional exploratory factor analysis models, measurement errors (e_{1-4}) are assumed to be independent of true presenteeism and of each other.

A data array that allows the test-retest reliability of the HPQ presenteeism scale to be estimated using this model was obtained as part of a calibration survey carried out in a sample of 105 airline reservation agents. A baseline HPQ was administered to this sample and these data were compared to independent supervisor ratings of work performance during the same month. The respondents then participated in a 1-week follow-up Experience Sample Method (ESM) evaluation[32] of moment-to-moment work experience two months after the baseline HPQ as well as in a repeat of the HPQ in a debriefing telephone interview the day after the end of the diary week. If we use the supervisor rating as a second indicator of presenteeism at Time 1 and the summary ESM presenteeism score as a second indicator of presenteeism at Time 2, we can estimate the parameters in the model specified in Fig. 1 and Equations 1–2a.

This model was estimated using the LISREL 8.30 software package,[33] fitting the covariance matrix among the four observed variables and constraining the unstandardized slopes of the HPQ scale at time t (HPQ_t) to be constant across t = 1,2. The overall model fit was excellent ($\chi^2_2 = 1.1, P = 0.30$), indicating that the six observed covariances among the four measured variables can be accurately reproduced by the four model parameter estimates (the slope of HPQ_t on WP_t constrained to be constant for t = 1,2; the slope of WP_2 on WP_1, the slope of the supervisor rating on WP_1, and the slope of the ESM scale on WP_2). Standardized parameter estimates are presented in Fig. 2, where we see that the estimated stability of true presenteeism over two months is 0.59 and the estimated reliability of the HPQ presenteeism scale is 0.89 (0.96 × 0.93). It is noteworthy that the observed correlation between HPQ_1 and HPQ_2, which is 0.521 in this sample, is very close to the estimated product of stability multiplied by reliability (ie, 0.59 × 0.89), further confirming the excellent fit of the model to the observed data.

The most important result in Fig. 2 for purposes of evaluating the sensitivity to change of the HPQ is that the reliability of HPQ_t is considerably higher than the estimated stability of true presenteeism even over the short time interval considered here. This result implies that the HPQ presenteeism scale is sensitive to change. A formal analysis of this issue requires computing the reliability of the change score ($HPQ_2 - HPQ_1$). The latter is defined as $(R_H - S_T)/(1 - S_T)$, where R_H is the reliability of HPQ_t (0.89) and S_T is the stability of WP.[34] Note that this formula implies that the reliability of the change score increases as the time interval increases. This is because reliability is of true score variance to total variance (ie, true score variance plus variance due to unreliability in the HPQ). True score variance (ie, inter-temporal change) will increase as the time interval increases, but the variance due to unreliability in the HPQ will remain constant, leading to an increase in the ratio of true:total variance. With S_T equal to 0.59 over the time interval considered here, the reliability of the HPQ change score is 0.73. Over a longer time interval, such as 6 months or a year, which would be the typical range of the time intervals considered in the evaluation of workplace health care interventions, the stability of WP would decrease and the reliability of the HPQ change score would increase proportionally.

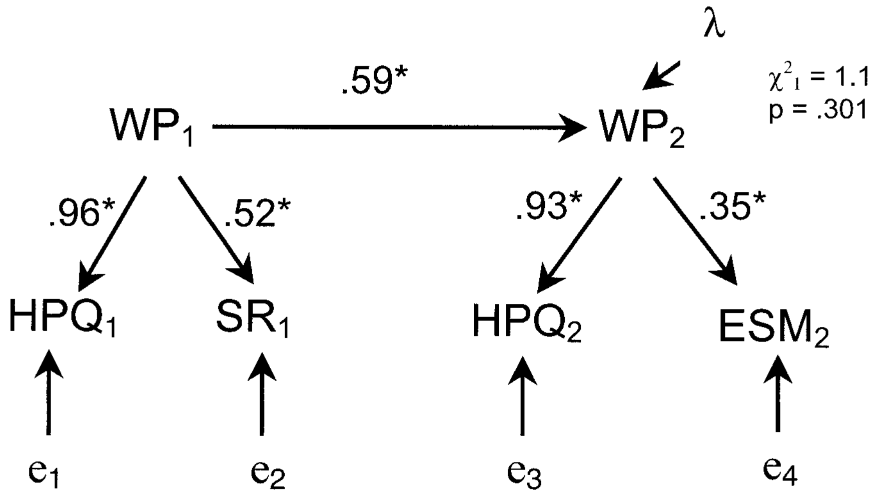

Fig. 2. Parameter estimates of the two-indicator two-time model of the stability of true presenteeism (WP_t) and the reliability of the HPQ presenteeism scale (HPQ_t). True presenteeism is included by the HPQ at both times by supervisor ratings at time 1 (SR_1), and by an aggregate performance measure based on moment-in-time sampling of work performance using the experience sampling method at time 2 (ESM_2). As in more conventional exploratory factor analysis models, measurement errors (e_{1-4}) are assumed to be independent of true work performance and of each other.

It is important to note that the data presented in Fig. 2 are not optimal because one might expect higher stability of a single scale (ie, the HPQ) than of different scales over the same interval of time (ie, SR_1 and ESM_2) due to correlated method variance (ie, a correlation between e_1 and e_2). A model that includes a term for correlated method variance cannot be identified, making it impossible to evaluate this possibility with only two times and two indicators. As a result, future work is needed that includes other indicators measured at multiple points in time. Despite this limitation, though, the results in Fig. 2 are consistent with the HPQ presenteeism scale being sensitive to change in true presenteeism over reasonable time intervals.

Critical Incidents

As noted above, job-related accidents are the third domain of work performance typically assessed by organizational and occupational psychologists. Even though they are uncommon, accidents are important because of their potential high cost. However, the same could be said of a number of other rare but important events in the workplace. Therefore, we decided to expand the assessment of accidents to include the broader domain of what we call "critical incidents," including big successes, big failures, and accidents. We explored a number of options for asking fully structured questions about these incidents. In the end, though, their rarity and great variety led us to include three separate open-ended question about successes, failures, and accidents-injuries-near-misses in the final HPQ. 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.

Conditions

Although the HPQ absenteeism, presenteeism, and critical incidents measures can be used on their own, the standard WHO HPQ instrument also includes separate series of questions about chronic conditions and the symptoms of acute conditions. For each reported condition, respondents are asked if they are currently in treatment and, if not, whether they were ever in treatment for this problem. The questions about conditions are included in the HPQ to obtain information about the prevalence of diverse health problems in the workplace as well as to study differences in the strength of association between individual conditions and the HPQ outcome measures. In considering either of these uses, though, the question arises as to the accuracy of the conditions measures.

The HPQ assesses chronic conditions using checklists modified from the US Health Interview Survey.[35] A number of methodological studies have found the self-reports obtained in these checklists to be valid for disorders brought to medical attention or that significantly limit activities when compared to independent medical records.[36–41] For example, moderate-to-high agreement (Cohen's κ)[42] has been found between self-reports and medical records regarding arthritis ($\kappa = 0.41$), asthma ($\kappa = 0.55$), diabetes ($\kappa = 0.82$), and high blood pressure ($\kappa = 0.73$).[36] These are lower bound estimates because the medical record is not a "gold standard," especially for chronic conditions that might not be brought to medical attention (eg, arthritis), for poorly defined conditions (eg, back pain), and for symptom-based conditions in which the medical record merely reproduces symptoms that are based on self-report (eg, chronic headaches).

In the case of symptom-based conditions, a number of more extensive scales exist that could have been used instead of the single yes-no question in the HPQ. For example, a brief and valid screener has been developed to assess migraine based on self-report.[43] A decision was made not to include symptom scales of this type in the basic version of the HPQ based on concerns about interview length and the realization that the HPQ checklist does a good job assessing most important chronic conditions. However, expanded versions of the HPQ are often used in workplace surveys that include a more in-depth assessment of a small number of conditions in conjunction with the standard HPQ checklist questions. This kind of tailored expansion can be guided by prior knowledge about conditions that are likely to have special importance in the workplace under study (such as musculoskeletal conditions in a blue-collar work setting in which heavy lifting is an important aspect of work performance).

The HPQ assessment of acute conditions, in comparison, uses two standard symptom checklists, one for mental disorders and the other for physical disorders. Mental disorders are assessed with the K6 symptom checklist of nonspecific psychological distress.[44] The K6 is a six-question Likert scale that assesses symptom frequency over the past 30 days for common symptoms of anxiety and mood disorders. The K6 has excellent concordance with blind clinical evaluations of mental disorders.[45] Acute physical conditions are assessed with items selected from the Patient Health Questionnaire 15 (PHQ-15), a 15-question scale of acute somatic symptom severity.[46] The PHQ-15 captures over 90% of the presenting complaints for acute physical health problems seen in primary care settings and has strong monotonic relationships with independent measures of global perceived health and functioning.[47]

Analysis

Once the measures of conditions and outcomes are available, conventional regression analysis can be used to estimate associations between conditions and outcomes. However, a number of important methodological issues arise here that warrant discussion. Four of these will be considered in this section of the article: the confounding effects of common causes; the role of risk adjustment in evaluating differences between health plans; the use of innovative strategies to evaluate the effects of treatment on work performance using non-experimental data; and the implications of comorbidity for evaluating the effects of individual conditions on work performance.

Common Causes

In estimating the effects of conditions on work performance, it is important to recognize that morbidity is not randomly assigned. This means that the regression coefficients linking conditions to performance cannot unequivocally be interpreted in causal terms. If unmeasured variables cause both increased risk of illness and work performance, failure to control for the effects of these variables will lead to bias in the estimated effects of conditions. Perhaps the most obvious example of this problem involves the effects of age. Age is strongly related to increased risk of many kinds of chronic conditions (eg, cardiovascular disorders, musculoskeletal disorders). To the extent that age is also significantly related to changes in work performance independent of illness, failure to control statistically for age in multiple regression analysis will lead to biased estimates of the effects of illness on work performance. However, age is not the only potentially important confounding variable. Others include gender, marital status, number and ages of children, and education. It is important to include controls for all of these variables in multiple regression estimates that estimate the effects of conditions on work performance.

The major advantage of experimental manipulation over analysis of naturalistic variation is that the effects of unmeasured causes can be assumed independent of the effects of the focal predictor variables, making it unnecessary to recognize, measure, and control for the effects of all possible confounding variables in order to obtain unbiased estimates of the effects of the focal predictor variables. It is sometimes possible to gain part of this advantage of experiments in naturalistic analysis by working with prospective individual-level data and assuming consistent effects of the unmeasured unchanging causes.[48] To see how this can occur, consider an unstandardized regression equation for the effects at time t of a given condition (C_t), time-varying causes such as age (V_t), and time-invariant (unchanging) causes such as sex (U_t) on work performance (HPQt):

$$HPQ\mathrm{t} = b_{0t} + b_1\mathrm{C}_t + b_2\mathrm{V}_t + b_3\mathrm{U}_t \qquad (3)$$

If this survey was repeated in a given workplace over two years and individual-level responses were linked, first differences could be taken between the equations at times 1 and 2. Assuming consistent slopes over time, the resulting difference score equation would be:

$$HPQ_2 - HPQ_1 = (b_{02} - b_{01}) + b_1(C_2 - C_1) + b_2(V_2 - V_1) + b_3(U_2 - U_1 = 0) \qquad (4)$$

Note that the effects of U_t cancel out, which means that any bias introduced by failing to control for U_t in Equation 3 disappears in Equation 4. In addition, a comparison of the estimates of b_1 in Equations 3 and 4 can be used to evaluate the effects of unmeasured common causes on bias in estimating the effects of conditions. This kind of comparison can be made in any workplace that repeats HPQ surveys on an annual basis.

The data analysis strategy described in the last paragraph applies only to unmeasured causes that are time-invariant. It is not possible to correct for the bias introduced by unmeasured time-varying causes. These time-varying causes have to be measured and introduced as explicit controls to adjust for their effects. We noted in the last section that age is the most obvious example, but it is not the only important time-varying common cause of conditions and work performance. A second that is also very important for some conditions is seasonality. A number of conditions vary in prevalence by season of the year. Seasonal allergies and flu are the two most obvious examples, but less extreme seasonal variation is also found for other acute conditions (eg, strains-sprains), for exacerbations of some persistent chronic conditions (eg, arthritis), and for episodes of some chronic-recurrent conditions (eg, depression).

To the extent that seasonal variation also exists in work performance, failure to control for the effects of seasonality introduces bias into estimates of condition effects. To control for seasonality, it is necessary to carry out HPQ surveys across all seasons of the year, ideally using randomization to assign respondents to a data of survey administration. A convenient approach of this sort is to key administration to the worker's birth date. Most workplace health and productivity surveys do not randomize season. As a result, great care is needed in interpreting the estimated effects of conditions that vary seasonally. For example, the estimated effects of self-reported seasonal allergies on work performance in HPQ surveys conducted either in the winter or summer are dramatically lower than the estimated effects in surveys conducted in the spring or fall.

Risk Adjustment

In addition to estimating the effects of specific health conditions on work performance, another common research question is whether health plans differ in the extent to which they reduce the work impairments associated with particular conditions. This question can be addressed by estimating statistical interactions in multiple regression analyses between conditions and plan membership in predicting work performance. The estimated effects of conditions on decrements in performance would be expected to be lower in plans that do a better job of treating these conditions. However, this kind of analysis is subject to a special case of the problem discussed in the previous section on unmeasured common causes: that unmeasured individual-level determinants of selecting different health plans among workers whose employers offer choice among plans might introduce bias into estimates of health plan effects.

For example, workers who take least care of their health might be expected to select the health plan with the lowest employee contribution among those offered by their employer, leading to an induced association between membership in this plan and subsequent health-related decrements in work performance even if this plan is as effective as other plans in ameliorating the effects of health problems on work performance. Or workers with the most severe cases of a particular health problem might select the health plan with the best disease management program for that problem, leading to upward bias in estimating the impact of that health problem on decrements in work performance among enrollees of that particular health plan.

These kinds of bias in health plan selection have been the subject of considerable interest among health services researchers.[49–51] The same general types of strategies as those described in the previous section on common causes are used to address these biases: to measure and control for the biases in the estimation of regression equations; and to modify research designs to remove the biases. In general, if potential sources of bias are measured (eg, individual level health consciousness and health locus of control), risk adjustment for between-plan differences in these variables can be achieved by introducing these measures as additive control variables in multiple logistic regression equations that also include main effects for conditions and plans as well as interactions between conditions and plans. The notion here is that these biasing variables are expected to have the same effects across plans, but to differ in their distributions across plans, allowing these effects to be controlled with additive compositional adjustments.

A design-based approach to estimating the magnitude of selection bias can be taken by pooling survey results across samples collected across a number of different workplaces in a single health care market. Market-wide HPQ surveys of this sort are carried out by a number of local business coalitions. In cases of this sort, the participating businesses usually differ in the number of health plan options they offer their employees, from one extreme of businesses that have an exclusive contract with a single health plan to the other extreme of businesses that offer employees a choice among all health plans in the market. Individual-level selection bias will necessarily increase in businesses that offer employee choice versus no choice and might also increase as the number of choices increases from only two to many. This variation can be used in the analysis of market-wide HPQ surveys by comparing the estimated risk-adjusted within-plan effects of specific health conditions on work performance in sub-samples that differ in amount of employee choice among plans.

The Nonexperimental Analysis of Treatment Effects

A related kind of selection bias involves estimating the effects of treatment. The HPQ collects data on whether respondents who report specific health problems are or are not in treatment for these problems. This makes it possible to estimate multiple regression equations that include separate predictions for untreated conditions and treated conditions. A fairly consistent pattern found across a number of HPQ surveys is that the significant effects of certain conditions on decrements in work performance are confined to workers who are not in treatment for these problems (eg, seasonal allergies), while for other conditions these effects are confined to workers who are in treatment (eg, arthritis) and for still others the effects are unrelated to treatment (eg, depression). The question obviously arises whether such results tell us anything about the effects of treatment.

The main difficulty with making inferences about treatment effects from such nonexperimental data (ie, data in which the researcher does not use some type of probability mechanism to manipulate exposure, intensity, or quality of treatment) is selection bias: that the severity and impairment caused by an illness strongly predict whether or not a person with that illness will seek treatment. This selection bias leads to a conservative bias in estimating treatment effects from non-experimental data. Indeed, this bias often swamps treatment effects, leading to a pattern in which workers with a particular illness have lower work performance if they are in treatment than if they are not in treatment. This does not mean that treatment hurts work performance (although there are certainly instances in which that may be the case).

Two important conclusions can sometimes be drawn from nonexperimental HPQ survey data on treatment despite the existence of selection bias. First, in instances where the aggregate work impairment of untreated cases is no greater than that of people without the condition, the most plausible interpretation is that selection processes keep mild cases out of treatment. This means that additional outreach efforts to increase the proportion of workers with the condition who obtain treatment would not be cost-effective from the employer perspective. Downstream cost savings (eg, early intervention to prevent future costs) are another matter and cannot be evaluated in cross-section HPQ surveys. The same is true for the extent to which increases in barriers to care might reduce the number of other mild cases who are currently in treatment without affecting work performance. Both of these issues can be examined by using longitudinal HPQ surveys to evaluate the effects of workplace health care interventions, but not in non-experimental cross-sectional analyses.

Second, in instances where the significant effects of the condition on decrements in work performance are greater among workers who are not in treatment than those in treatment, one can reasonably conclude that treatment is effective. This interpretation is based on the assumption that selection bias works in the opposite direction from the observed data pattern (ie, more serious cases seek treatment). The magnitude of the treatment effect cannot be estimated by comparing the slopes of treated and untreated cases because of the selection bias. However, in the absence of other information, it is plausible to assume that the treatment effect is at least as large as the difference between the slopes of treated and untreated cases. It is also reasonable in such cases to consider the possibility that outreach to increase the treatment rate among workers with this condition might be cost-effective from the employer perspective.

It is difficult to draw firm conclusions about treatment effects in the more typical instance where the significant effects of the condition on decrements in work performance are greater among workers who are in treatment than those not in treatment. In such instances, it might be that the performance of workers in treatment, albeit lower than the performance of untreated workers with the same condition, might have been even lower in the absence of treatment. Similarly, the performance of workers not in treatment, even though it is better than that of treated workers, might nonetheless improve significantly with treatment. Yet we have no way to know if either of these possibilities is the case in the absence of other data.

The obvious way to resolve these uncertainties is to implement a controlled treatment intervention in which probability or quasi-probability mechanisms are used to assign some, but not all, individuals or units of individuals (eg, some, but not other, branches of a bank; some, but not other, departments in a large corporation; some, but not other, businesses in a local business coalition) to greater access, intensity, or quality of treatment than others. Comparisons between respondents in the different treatment arms can be used to make inferences about treatment effects that are much less subject to selection bias than the comparisons based on non-experimental data. Weaker, but nonetheless useful, inferences about treatment effects can be made from before-after comparisons in a single business setting.[3]

Is there any way short of such an intervention to make useful, although necessarily incomplete, inferences about likely treatment effects from non-experimental data? This is possible in some instances, but access to additional information is required that can be used to introduce the equivalent of quasi-randomization into the analysis. A good example is the work of Weiss et al,[52] who estimated the cost-effectiveness of implantable cardioverter defibrillators in patients with ventricular arrhythmias by making use of the two observations that (1) the proportion of patients who receive this procedure varies enormously across hospitals and (2) that the vast majority of patients who seek treatment for this condition do so by going to the hospital that is closest to their home. Weiss and colleagues used these observations to create a score for the predicted probability of receiving this procedure for each patient eligible for the procedure across a large sample of hospitals based on the track record of the hospitals for performing the procedure in the past. The amount of variance in actual use of the procedure that was predicted by this score, which was assumed to be independent of any individual selection bias, was used to implement an econometric estimation procedure that allows treatment effect to be estimated with only minimum bias. Although we are aware of no comparable within-market studies, one could easily imagine similar analyses being carried out to evaluate the effects of disease management programs that are unique to individual health plans in a single market by pooling data across members of a business coalition in the market that differ in the access they give their workers to participation in that health plan.

Comorbidity

The discussion of data analysis issues has so far focused on the effects of individual conditions and their treatment. However, the majority of working people with chronic disorders suffer from more than one chronic condition. This is illustrated in Table 2, which shows the distribution of the number of chronic conditions reported by respondents in a series of HPQ surveys carried out in the summer of 2003. Only 13.2% of respondents reported that they had none of the 27 chronic conditions in the HPQ checklist, while an additional 15.9% reported having only one of these conditions and the remaining 70.9% reported having two or more conditions. Among workers who reported having at least one chronic condition, the median number of conditions was four.

Previous research has shown consistently that comorbid disorders are, in general, more impairing than pure disorders in both clinical samples[53] and community samples.[15] These same studies have also shown that the incremental effects of additional disorders on diverse measures of functioning generally decrease as the number of comorbid conditions increases. This is illustrated in Table 3,

TABLE 2
The Distribution of Number of Chronic Conditions in Recent HPQ Surveys (n = 10,050)

Number of Conditions	%	Cumulative %
0	13.2	13.2
1	15.9	29.1
2	15.6	44.7
3	13.0	57.7
4	10.8	68.5
5	8.3	76.8
6	6.7	83.4
7	4.8	88.3
8	3.5	91.8
9	2.8	94.6
10	2.0	96.5
11+	3.5	100.0

TABLE 3
The Estimated Effects of Chronic Pain Disorder on Annualized Absenteeism Days in Recent HPQ Surveys Among Respondents Who Differ in Number of Comorbid Conditions (n = 10,050)†

	Prevalence (%)	Annualized Effect (days/year)
Pure	2.1	7.2
Low comorbidity (1–3)	23.9	5.5
Medium comorbidity (4–6)	32.6	2.5
High comorbidity (7–9)	25.3	0.2
Very high comorbidity (10+)	16.1	0.2
Total	14.5	3.0

† Effect size estimates are based on linear regression equations that control for age, sex and occupation in predicting 30-day absenteeism. Annualized estimates are projected by simple linear extrapolation from 30-day estimates.

which presents data on chronic pain from the same HPQ surveys as in Table 2. The first column shows that 14.5% of respondents reported chronic pain and that the vast majority of these workers also reported a number of other comorbid chronic conditions. The median number of conditions reported by workers with chronic pain was five. The second column of Table 3 presents the estimated annualized effects of chronic pain based on a series of multiple regression equations that included separate dummy variables for each of the 27 chronic conditions in the HPQ checklist plus controls for several sociodemographic variables (age, sex, occupation).

The equations used to estimate the results in the second column of Table 3 differed only in the sub-samples of respondents that they included. The equation used to estimate the effect in the first row (7.2 days/year) was based on the 85.5% of respondents who did not report chronic pain plus the 2.1% who reported that chronic pain was their only chronic condition. The equations used to estimate the effects in the next four rows (5.5, 2.5, 0.2, and 0.2 days/year) were based on the 85.5% of respondents who did not report chronic pain plus those who reported chronic pain in addition to either a low (1 to 3), medium (4 to 6) high (7 to 9), or very high (10 or more) number of other comorbid conditions. The equation used to estimate the effect in the last row (3.0 days/year) was based on the entire sample. The results are very clear in showing that the estimated effect of chronic pain decreases monotonically as the number of comorbid conditions increases.

Two conclusions can be drawn from this pattern of results. The first is that the estimated effect of chronic pain in the total sample, 3.0 days/year, is not a very adequate descriptor of the actual impact of this condition because of the interaction between chronic pain and number of comorbid conditions in predicting absenteeism. A disaggregated analysis of the sort shown in Table 3 provides a much more accurate description. The second conclusion is that the number of conditions is sometimes more important than the nature of conditions among workers with high comorbidity. This is borne out in more extensive analyses of the same data set (results not presented in the table), which show that very few individual conditions are significant predictors of absenteeism among workers with high comorbidity even though workers with high comorbidity have an extremely high rate of absenteeism.

An important methodological implication of these two conclusions is that the best way to estimate a single multiple regression equation for the joint effects of many different chronic conditions on absenteeism, if a single equation is desired, is to define separate dummy variables for each of the conditions under consideration among workers who do not have high comorbidity in addition to one or more separate dummy variables for workers with high comorbidity that ignore the nature of their conditions. An even more useful approach is to abandon the use of a single equation and to examine the joint effects of number of conditions and, among people with a given number of conditions, types of conditions, in predicting absenteeism. This kind of joint investigation of number and nature is a standard feature of the analysis of HPQ survey data.

A final consideration in the analysis of comorbidity is that certain types of comorbidity might have especially powerful synergistic effects on workplace functioning. One class of comorbidities that has become the subject of special interest in this regard involves comorbidities of mental disorders with chronic physical disorders. Strong patterns of mental-physical comorbidity have been found for a number of commonly occurring physical disorders both in general population samples[54] and in primary care samples.[55] In addition, clinical studies have found substantial impairment associated with co-occurring mental disorders among people with chronic physical conditions.[56] Furthermore, recent analysis of HPQ data has shown that the effects of several common chronic physical conditions on absenteeism increase dramatically when these conditions are comorbid with anxiety disorders or clinical depression.[15] These results raise the intriguing possibility that expanded outreach and treatment of workplace mental disorders might be cost-effective by virtue of the indirect effects on reductions in the work impairments associated with comorbid physical disorders.

Discussion

The results presented in the first section of the article demonstrated that the HPQ work performance measures are reliable, valid, and sensitive to change. Data also were reviewed that demonstrate the reliability and validity of the HPQ measures of chronic and acute conditions. Taken together, these results support the use of the HPQ as an efficient and accurate method to obtain information on workplace health and productivity. However, we also noted that the HPQ is designed to estimate overall indirect costs of individual health problems rather than to collect data on the ways in which individual health problems influence overall work performance (eg, by decreasing the abilities to lift, read, concentrate etc.). It is important for users to be clear about which type of data they need for their research purposes. In cases where both types of data are of interest, domain-specific questions (eg, the short version of the WLQ) can be added to the HPQ.

Endemic data analysis problems that confront health and productivity researchers irrespective of the measurement tools they use were discussed in the second section of the article. We also discussed ways these problems are addressed in current HPQ studies using innovative approaches to research design and data analysis. All of these approaches focused on making causal inferences about the effects of health care interventions. Such inferences are required for the employer to calculate a return on investment in health care. The ideal way of doing this, of course, is to carry out an experiment (eg, randomly assigning workers or business units to the intervention). We are involved in several such experiments using the HPQ as one of the primary outcomes. When this is not possible, though, quasi-experimental test market studies and naturalistic studies are required. The HPQ can be used in all these types of studies.

As cross-section surveys are the beginning step for more complex designs, electronic HPQ report-generating software has been developed to produce easy-to-interpret reports from cross-sectional HPQ surveys. These reports allow employers to answer a number of basic, but important, questions that must be addressed in order to have rational health care decision-making: (1) Which conditions are most common in my workforce? (2) Which conditions are associated with the greatest lost productivity in my workforce? (3) Are the presumed effects of the latter conditions confined to workers who are in treatment for these conditions, to workers who are not receiving treatment, or both? (4) What is the monetary value of the lost performance associated with these conditions? (5) Are the presumed effects of these conditions the same or different across health plans?

Answers to these questions can help employers choose health plans that adequately treat costly conditions. Of course, HPQ data will only be one of several important inputs to this strategic planning. Nonetheless, the part of the relevant data provided by the HPQ is currently absent from the decision-making information available to most employers. The answer to question (5) in the last paragraph is of special importance. As noted earlier in the paper, a number of sophisticated statistical methods have been developed for the valid non-experimental comparison of outcomes across health plans.[49–51] In the ideal case, such analyses would be carried out at a market level in collaboration with an employer coalition that is able to generate a database far larger and more varied than the one that could be generated by focusing on the employees of a single company. Based on this realization, we are currently working with the Midwest Business Group on Health in collaboration with the National Business Coalition on Health to develop a model health and productivity evaluation and quality assurance system that uses the HPQ to assess indirect costs of illness.

It is important to realize that sophisticated comparisons of health plans using thoughtful case-mix risk adjustment methods have been carried out in the past. However, these evaluations have focused almost entirely on health care outcomes rather than on workplace outcomes.[57,58] There has been a particular focus on process measures (ie, various clinical quality measures) rather than on outcome measures (ie, morbidity, mortality, speed of recovery), with some notable exceptions.[59,60] There are two reasons why workplace outcomes (ie, presenteeism, absenteeism, duration of work disability) have not been included in these studies. First, research in this area has been under the direction of health care professionals whose main interest is in health care outcomes rather than workplace outcomes. Second, health care outcomes measures are much more readily available than workplace outcomes measures.

To obtain more and better data on between-plan differences in workplace outcomes, employers need to make it clear to their health plans that workplace outcomes are of central importance to their contracting decisions. In addition, they have to facilitate access to workplace outcomes data. Both of these things can be done most effectively when employer coalitions work with their local health plans to develop a coordinated approach to collecting workplace outcomes data in parallel across a number of different workplaces. This is the approach we are taking in our work with the Midwest Business Group on Health.

In addition, it is important to recognize that one-time evaluation of health plan performance has to be followed with ongoing quality assurance monitoring. The HPQ can be extremely useful in the latter regard, as annual HPQ tracking surveys can be used to monitor trends in rates of treatment among workers with costly conditions as well as trends in both within-plan effects of specific conditions on work performance and between-plan differences in these effects. In cases where trends of this sort are monitored carefully across a number of collaborating corporations in a market-wide employer health care coalition, it is feasible to envision the development of health plan incentives based on these trends to increase productivity rather than merely to decrease direct costs. The field of health and productivity management is too new for good models of this sort to exist, but this is clearly a feasible goal for the future as implementation of health and productivity surveys become a routine ongoing part of employer data gathering in the service of maximizing returns on health care investments.

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 initial HPQ 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 conducting the calibration survey. We appreciate the helpful comments of two anonymous reviewers and of John Kochevar on an earlier draft of this article. The complete text of the HPQ is posted at http://www.hpq.org. Additional information about the HPQ can be found at 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. Scanlon DP. Overcoming barriers to managing health and productivity in the workplace. In: Kessler RC, Stang PD, eds. *Health and Work Productivity: Emerging Issues in Research & Policy.* Chicago: University of Chicago Press; In Press.
2. Parmenter EM. Controlling health care costs: Components of a new paradigm. *J Financial Services Prof.* 2003;57:59–68.
3. Rosenheck RA, Druss B, Stolar M, Leslie D, Sledge W. Effect of declining mental health service use on employees of a large corporation. *Health Aff (Millwood).* 1999;18:193–203.
4. Aldana SG. Financial impact of health promotion programs: a comprehensive review of the literature. *Am J Health Promot.* 2001;15:296–320.
5. Lynch W, Riedel JE. *Measuring Employee Productivity: A Guide to Self-Assessment Tools.* Scottsdale, AZ: The Institute for Health and Productivity Management; 2001.
6. Loeppke R, Hymel PA, Lofland JH, et al. Health-related workplace productivity measurement: general and migraine-specific recommendations from the ACOEM Expert Panel. *J Occup Environ Med.* 2003;45:349–359.
7. 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.
8. Kessler RC, Barber C, Beck A, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ). *J Occup Environ Med.* 2003;45:156–174.
9. Harbour JL. *The Basics of Performance Measurement.* New York: Productivity, Inc.; 1997.
10. Grote D. *The Complete Guide to Performance Appraisal.* New York: AMACOM; 1996.
11. Murray CJL, Lopez AD. *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries and Risk Factors in 1990 and Projected to 2020.* Cambridge, MA: Harvard University Press; 1996.
12. 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 Psychiatric Res.* 1999;8:110–123.
13. World Health Organization. *International Classification of Functioning, Disability and Health.* Geneva: World Health Organization; 2001.
14. Kessler RC, Ustun TB. The World Health Organization World Mental Health 2000 Initiative. *Hosp Manag Int.* 2000:195–196.
15. Kessler RC, Ormel J, Demler O, Stang PE. Comorbid mental disorders account for the role impairment of commonly occurring chronic physical disorders: results from the National Comorbidity Survey. *J Occup Environ Med.* 2003;45:1257–1266.
16. Wang PS, Beck A, Berglund PA, et al. Chronic medical conditions and work performance in the HPQ Calibration Surveys. *J Occup Environ Med.* In Press.
17. Holloway J, Lewis J, Mallory G. *Performance Measurement and Evaluation.* London: Sage; 1995.
18. Warr PB, Conner MT. *The Measurement of Personal Effectiveness for Review and Guidance.* London: Department of Education and Employment; 1999.
19. U. S. Department of Labor Employment and Training Administration. *Testing and Assessment: An Employer's Guide to Good Practices.* Washington, DC: US Government Printing Office; 1999.
20. Whetzel DL, Wheaton GR, eds. *Applied Measurement Methods in Industrial Psychology.* Cleveland: Davis-Black; 1997.
21. Pritchard RD, Holling H, Lammers F, Clark BD, eds. *Improving Organizational Performance with the Productivity Measurement and Enhancement System: An International Collaboration.* Huntington, NY: Nova Science; 2002.
22. Endicott J, Nee J Endicott Work Productivity Scale (EWPS): a new measure to assess treatment effects. *Psychopharmacol Bull.* 1997;33:13–16.
23. Koopman C, Pelletier KR, Murray JF, et al. Stanford presenteeism scale: health status and employee productivity. *J Occup Environ Med.* 2002;44:14–20.
24. Lerner D, Amick BC 3rd, Rogers WH, Malspeis S, Bungay K, Cynn D. The Work Limitations Questionnaire. *Medical Care.* Jan 2001;39:72–85.
25. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics.* 1993;4:353–365.
26. Means B, Loftus EF. When personal history repeats itself: Decomposing memories for recurring events. *Appl Cognit Psychol.* 1991;5:297–318.
27. Menon A. Judgments 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.
28. 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, DC: Department

of Health, Education, and Welfare; 1979: 74–108.

29. Kline RB. *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press; 1998.
30. Crocker LM, Algina J. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston; 1986.
31. Bollen KA. *Structural Equations with Latent Variables.* 1st ed. New York: John Wiley & Sons, Inc.; 1989.
32. Csikszentmihalyi M, Larson R. Validity and reliability of the Experience Sampling Method. In: deVries M, ed. *The Experience of Psychopathology*. Cambridge: Cambridge University Press; 1992:43–57.
33. Joreskog KG, Sorbom D. *LISREL 8: A Guide to the Program and Applications.* Chicago: Scientific Software International; 1993.
34. Kessler RC, Greenberg DF. *Linear Panel Analysis: Models of Qualitative Change.* New York: Academic; 1981.
35. National Center for Health Statistics. *Summary Health Statistics for US Population: National Health Interview Survey, 1999.* Vol PHS 2003–1539. Hyattsville, MD: US Department of Health and Human Services; 2003.
36. National Center for Health Statistics. Evaluation of National Health Interview Survey Diagnostic Reporting. *Vital Health Stat 2.* 1994;120:1–116.
37. Halabi S, Zurayk H, Awaida R, Darwish M, Saab B. Reliability and validity of self and proxy reporting of morbidity data: a case study from Beirut, Lebanon. *Int J Epidemiol.* 1992;21:607–612.
38. Heliovaara M, Aromaa A, Klaukka T, Knekt P, Joukamaa M, Impivaara O. Reliability and validity of interview data on chronic diseases. The Mini-Finland Health Survey. *J Clin Epidemiol.* 1993; 46:181–191.
39. Kriegsman DM, Penninx BW, van Eijk JT, Boeke AJ, Deeg DJ. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy. *J Clin Epidemiol.* 1996;49:1407–1417.
40. Gross R, Bentur N, Elhayany A, Sherf M, Epstein L. The validity of self-reports on chronic disease: characteristics of underreporters and implications for the planning of services. *Public Health Rev*. 1996;24:167–182.
41. Mackenbach JP, Looman CW, van der Meer JB. Differences in the misreporting of chronic conditions, by level of education: the effect on inequalities in prevalence rates. *Am J Public Health.* May 1996;86:706–711.
42. Cohen J. A coefficient of agreement for nominal scales. *Ed Psychol Meas.* 1960; 20:37–46.
43. Lipton RB, Dodick D, Sadovsky R, Kolodner K, Hettiaracher J, Harrison W. A self-administered screener for migraine in primary care: the ID Magrain (TM) validation study. *Neurology*. 2003;61: 375–382.
44. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in nonspecific psychological distress. *Psychol Med.* 2002;32:959–976.
45. Kessler RC, Barker PR, Colpe LJ, et al. Screening for serious mental illness in the general population. *Arch Gen Psychiatry*. 2003;60:184–189.
46. 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.
47. Kroenke K. Patients presenting with somatic complaints: Epidemiology, psychiatric comorbidity and management. *Int J Methods Psychiatric Res*. 2003;12:34–43.
48. Duncan GJ, Liker J, Augustyniak S. Panel data and models of change: a comparison of first difference and conventional two-wave models. *Soc Sci Res*. 1985;14:80–101.
49. Benneyan JC, Borgman AD. Risk-adjusted sequential probability ratio tests and longitudinal surveillance methods. *Int J Qual Health Care*. 2003;15:5–6.
50. Austin PC. A comparison of Bayesian methods or profiling hospital performance. *Med Decis Making*. 2002;22: 163–172.
51. Zhao Y, Ash AS, Ellis RP, Slaughter JP. Disease burden profiles: an emerging tool for managing managed care. *Health Care Manag Sci*. 2002;5:211–219.
52. Weiss JP, Saynina O, McDonald KM, McClellan MB, Hlatky MA. Effectiveness and cost-effectiveness of implantable cardioverter defibrillators in the treatment of ventricular arrhythmias among medicare beneficiaries. *Am J Med.* 2002;112:519–527.
53. Ormel J. Functioning, well-being, and health perception in late middle-aged and older people: comparing the effects of depressive symptoms and chronic medical conditions. *J Am Geriatr Soc.* 1998; 46:39–48.
54. Neeleman J, Ormel J, Bijl RV. The distribution of psychiatric and somatic ill-health: associations with personality and socioeconomic status. *Psychosom Med*. 2001;63:239–247.
55. Berardi D, Berti Ceroni G, Leggieri G, Rucci P, Ustun TB, Ferrari G. Mental, physical and functional status in primary care attenders. *Int J Psychiatry Med*. 1999;29:133–148.
56. Sullivan MD, LaCroix AZ, Russo JE, Walker EA. Depression and self-reported physical health in patients with coronary disease: mediating and moderating factors. *Psychosom Med*. 2001;63:248–256.
57. Solomon LS, Zaslavsky AM, Landon BE, Cleary PD. Variation in patient-reported quality among health care organizations. *Health Care Financ Rev.* 2002; 23:85–100.
58. Zaslavsky AM, Shaul JA, Zaborski LB, Cioffi MJ, Cleary PD. Combining health plan performance indicators into simpler composite measures. *Health Care Financ Rev.* 2002;23:101–115.
59. Mukamel DB, Weimer DL, Zwanziger J, Mushlin AI. Quality of cardiac surgeons and managed care contracting practices. *Health Serv Res.* 2002;37:1129–1144.
60. Hannan EL, Wu C, Ryan TJ, et al. Do hospitals and surgeons with higher coronary artery bypass graft surgery volumes still have lower risk-adjusted mortality rates? *Circulation.* 2003;108:795–801.