MEDICAL CARE
Volume 39, Number 1, pp 72–85
©2001 Lippincott Williams & Wilkins, Inc.

The Work Limitations Questionnaire

DEBRA LERNER, MS, PHD,*† BENJAMIN C. AMICK III, PHD,‡ WILLIAM H. ROGERS, PHD,*†
SUSAN MALSPEIS, SM,§ KATHLEEN BUNGAY, PHARMD,*† AND DIANE CYNN, BA*

OBJECTIVE. The objective of this work was to develop a psychometrically sound questionnaire for measuring the on-the-job impact of chronic health problems and/or treatment ("work limitations").

RESEARCH DESIGN. Three pilot studies (focus groups, cognitive interviews, and an alternate forms test) generated candidate items, dimensions, and response scales. Two field trials tested the psychometric performance of the questionnaire (studies 1 and 2). To test recall error, study 1 subjects were randomly assigned to 2 different questionnaire groups, a questionnaire with a 4-week reporting period completed once or a 2-week version completed twice. Responses were compared with data from concurrent work limitation diaries (the gold standard). To test construct validity, we compared questionnaire scores of patients with those of healthy job-matched control subjects. Study 2 was a cross-sectional mail survey testing scale reliability and construct validity.

SUBJECTS. The study subjects were employed individuals (18–64 years of age) from several chronic condition groups (study 1, n = 48; study 2, n = 121) and, in study 1, 17 healthy matched control subjects.

MEASURES. Study 1 included the assigned questionnaires and weekly diaries. Study 2 included the new questionnaire, SF-36, and work productivity loss items.

RESULTS. In study 1, questionnaire responses were consistent with diary data but were most highly correlated with the most recent week. Patients had significantly higher (worse) limitation scores than control subjects. In study 2, 4 scales from a 25-item questionnaire achieved Cronbach alphas of ≥0.90 and correlated with health status and self-reported work productivity in the hypothesized manner (*P* ≤0.05).

CONCLUSIONS. With 25 items, 4 dimensions (limitations handling time, physical, mental-interpersonal, and output demands), and a 2-week reporting period, the Work Limitations Questionnaire demonstrated high reliability and validity.

Key words: Work productivity; chronic disease and employment; disability. (Med Care 2001;39:72–85)

Approximately 55 million working-age individuals (18 to 65 years of age) have chronic illnesses and/or impairments and thus are vulnerable to disability.[1] Disabilities are a potential consequence of health problems and signify a partial or total inability to perform social roles in a manner con-

*From The Health Institute, Division of Clinical Care Research, New England Medical Center, Boston, Massachusetts.

†From Tufts University School of Medicine, Boston, Massachusetts.

‡From the University of Texas Houston Health Sciences Center, School of Public Health, Houston, Texas, and the Institute for Work and Health, Toronto, Canada.

§From Harvard School of Public Health, Boston, Massachusetts.

Sponsored by Glaxo Wellcome, Inc.

Address reprint requests to: Debra Lerner, MS, PhD, The Health Institute, New England Medical Center, NEMC Box 345, 750 Washington St, Boston, MA 02111. E-mail: dlerner@lifespan.org

Address requests for the Work Limitations Questionnaire to: wlq@lifespan.org

Received December 6, 1999; initial review completed February 22, 2000; accepted September 8, 2000.

sistent with norms or expectations.[2] National survey data suggest that 32% of employed adults have ongoing health problems that interfere with their ability to perform their job demands.[3] The national cost of lost work productivity resulting from chronic conditions has been estimated to be at least $234 billion annually.[4]

Statistics such as these underlie a growing effort to document the social and economic outcomes of various chronic health problems and their treatment[5,6] and have spawned interest in including work disability and work productivity loss, defined collectively as "work loss," as study end points. Because comprehensive archival work loss data are relatively scarce and difficult to obtain, research has relied principally on self-report.[7] Self-reports have addressed labor market participation, work absences, on-the-job effectiveness, and role disability.[8]

Degree of labor market participation is a useful work loss indicator when a condition or treatment is expected to influence a person's employment status and/or occupation. However, when these are infrequent outcomes, on-the-job performance measures have greater validity. One widely used indicator is the amount of time missed from work because of illness or treatment.[9,10] However, despite its acceptance, susceptibility to recall error remains a persistent concern.[11] Some studies have addressed on-the-job performance by asking individuals to rate their effectiveness on days when they are symptomatic,[12,13] although psychometric evidence is limited.

Scales such as the Role Limitation scales of the SF-36 represent another measurement approach using global, role-level disability indicators to capture disability in paid work and/or other activities (eg, "Were [you] limited in the kind of work or other activities?").[14] However, disability scales can be relatively coarse, distinguishing a limited range of disability levels.

We developed the Work Limitations Questionnaire (WLQ) to fill this gap in measuring the on-the-job impact of chronic conditions and treatment. A long-term goal is to facilitate the economic assessment of work loss.

We report on 3 pilot studies and 2 psychometric field trials (studies 1 and 2). Appendix 1 describes each sample. Appendix 2 illustrates the genealogy of the WLQ items and scales.

Methods

Pilot Studies

The WLQ content and format originated from focus groups, cognitive interviews, and an alternate forms comparison. Each pilot included patients 18 to 64 years of age who were employed ≥20 h/wk within the following condition groups: respiratory diseases (asthma), gastrointestinal diseases (Crohn's Disease and liver disease), psychiatric disorders (depression and/or generalized anxiety), or epilepsy (Appendix 1). We excluded patients with a planned or pending work disability claim and/or substance abuse problem. Participants received a monetary incentive ($40).

Focus Groups. To identify questionnaire content, 4 condition-specific focus groups were convened. Four participating physicians were asked to nominate 5 to 10 patients. Twenty-one were nominated; 18 (86%) participated. Next, we created a list of discussion topics and a focus group guide.[15] Each topic addressed a job demand category contained within 2 well-known work classification taxonomies.[16,17] Each 2.5-hour discussion was audiotaped. Tapes were transcribed and analyzed.

Initially, each participant was asked to describe his/her job, health status, a "good" health day at work, and a "bad" day. Participants were also asked whether their jobs required them to perform each type of demand and how their health and medical care affected its performance. As a result, we generated 70 job demand–level limitation items and 7 dimensions (column 1, Appendix 2).

Cognitive Interviewing. Cognitive interviews potentially enhance the reliability and validity of a questionnaire.[18,19] Using a think-aloud methodology, we assessed how another sample of respondents interpreted and answered the candidate items. The performance of each item was rated on the basis of interview data.

With a research assistant (RA) present, each of 37 respondents completed an open-ended questionnaire. The question asked: "In the past 4 weeks, how much difficulty did you have performing each of the following because of your physical health or emotional problems. . . ?" A list of job demands followed (eg, concentrating on work). This open-ended format meant that each respondent could choose a response terminology.

Respondents were instructed to read each question silently or aloud, paraphrase it, and think aloud while answering. A probing segment followed in which respondents discussed work limitations reported during the interview, misinterpreted or difficult items, and suggestions for additional topics. Interviews were audiotaped and coded.

Using the data, we rated items for their comprehensibility, redundancy, relevancy to job demands and health problems, and ease of responding. Items with high problem frequencies and/or relatively low work limitation rates were eliminated. As a result, 32 of 70 items failed. Of 38 passing items, 23 were revised to reduce awkward or unnecessary words. Two items were added (total of 40).

Several candidate items had validity problems. For example, certain items did not apply to respondents' job demands. The Physical Demands section performed worst. This problem was cited in 23.6% of 407 administrations (37 subjects times 11 items). The corresponding rate for the best scale, Interpersonal Demands, was 3.2%. Other items lacked applicability to respondents' illnesses. This deficiency was cited most frequently in response to items in the Information Processing section (17.6% of administrations). The Time Management section performed best in this regard (4.3% of administrations). Within each section, item redundancy problems occurred in 3% to 10% of administrations.

The Information Processing section had 1 passing item. It was included with Mental Demands. The Physical Environment items were deleted entirely because of interpretation problems. Thus, 40 items and 5 dimensions remained (Appendix 2).

No single response pattern emerged. Among the terms respondents used to answer questions were "difficult"/"not difficult," "can do"/"can't do,""able to do"/"unable to do,"and "a problem"/ "not a problem."

Alternate Forms. In a third sample, we assessed the reliability of 3 different forms. Each contained the same 9 job demands embedded within the following stem/response options.

1. "In the past 4 weeks, how much difficulty have you had doing the following because of your physical health or emotional problems . . . " (5 responses ranging from "no difficulty" to "so much difficulty I couldn't do it")?
2. "How much time during the past 4 weeks were you able to do the following . . . " (5 responses ranging from "all of the time" to "none of the time")?
3. "On how many days during the past 4 weeks were you able to do the following . . . " (5 responses ranging from "more than 20 days" to "0 days")?

Each scale included the option "does not apply to my job."The last 2 contained a follow-up yes/no item ("If able to do less than all of the time, was it due to your health?").

We compared scales worded negatively (difficulty) and positively (able) and those measuring intensity (amount of difficulty) and frequency (amount of time). Questionnaires such as the SF-36 include intensity and frequency scales; however, the economic assessment of work loss usually involves a time factor (eg, lost work time).[10]

The forms were completed in the presence of an RA. Order bias was reduced by shuffling forms before each administration.

During an audiotaped portion, participants were asked to describe their reasons for choosing certain responses, identify events that would have led to selecting another response, and rate the accuracy of responses.

The analysis compared responses on form 1 versus 2 versus 3. Matching responses were considered reliable. If responses did not match, we attempted to determine which was correct by comparing the mismatched responses with the transcripts.

Of the 324 responses compared (9 items times 36 subjects), 79% were 3-way matches, 20% were 2-way mismatches, and <1% were 3-way mismatches. Of the 2-way and 3-way mismatches, 68% involved a disagreement with the "days" form, and it was rejected (mismatch rates for the "difficulty" and "time able" forms rates were 32% and 38%, respectively).

We compared mismatched responses on the 2 remaining forms with transcript data and found that the "difficulty"form captured events more accurately than the "time able" form. Consequently, we adopted a difficulty question stem for 4 sections (Time, Mental, Interpersonal, and Output Demands). For Physical Demands, we adopted, "How much of the time were you able to do the following without difficulty due to physical health or emotional problems?"A single response scale was chosen—eg, all of the time (100%), a great deal of the time, some of the time (~50%), a slight bit of the time, and none of the time (0%)—which could facilitate future economic analyses.

Field Trial Methods (Studies 1 and 2)

Designs. Using the 40 WLQ items and response scales developed in the pilot tests (Appendix 2), study 1 evaluated recall error. Two mail versions of the WLQ were tested: 1 with a 2-week reporting period and 1 with a 4-week reporting period. One randomly assigned group took the 2-week version, asking about work limitations in the past 2 weeks. It was administered at the end of study weeks 2 and 4. A different randomly assigned group took the 4-week version, asking about work limitations in the past 4 weeks. It was administered once at the end of study week 4. During the same weeks, both questionnaire groups also recorded work limitations on 4 weekly diaries (completed the last day of each week). These supplied a "gold standard" for judging the accuracy of the questionnaire data.

A case-control study was nested within study 1 comparing WLQ scores of patients and healthy coworkers matched on job and employer. Significantly higher (more limited) WLQ scores among patients provided initial evidence of construct validity.

Study 2 utilized a cross-sectional design to test 2 hypotheses: in H1, the WLQ contains internally consistent scales (a facet of reliability); in H2, scale scores correlate with measures of role disability and with self-reported work productivity (construct validity).

Study Populations

Study 1 included specialty clinic patients who met the pilot study criteria (Appendix 1). Site clinicians identified potentially eligible, interested patients. An RA called patients, explained the protocol, and assessed eligibility. Eligible study 1 patients were asked to nominate a job-matched coworker. Both were blinded to the fact that health status determined eligibility. To protect coworker confidentiality, each patient was asked to tell a coworker about the study and supply our phone number. During the call, the protocol was explained and eligibility was assessed. Eligible coworkers had the same job and employer as the patient, reported no major chronic conditions, and met the remaining study 1 criteria. Some patients, for privacy reasons, did not nominate a coworker or did not have a match. We included these patients in an "unmatched patient group" to participate in the questionnaire/diary protocol. All subjects received a monetary incentive to participate.

We attempted to recruit 60 subjects: 20 patient/coworker pairs (n = 40) and 20 unmatched patients; 90 patients were screened, and we enrolled 17 matched pairs (n = 34) and 31 unmatched patients (total n = 65; Appendix 1). The main reason for exclusion was lack of availability for 4 consecutive weeks. Additionally, we reduced the number of matched pairs from 17 to 14 after 3 "healthy" controls were found to have SF-36 mental health scores indicative of clinical depression.[20]

Each subject was randomized to a questionnaire group (with matched pairs assigned to the same group). We assigned 29 subjects (45%) to the 2-week WLQ group and 36 (55%) to the 4-week group. Using χ^2 or *t* test statistics as appropriate, we found no significant differences between the questionnaire groups on mean age, percent male, mean education, occupation (percent manual versus nonmanual),[21] percent with a condition, and mean SF-36 scale scores.

Study 2 consisted of 3 groups: (1) rheumatoid arthritis patients from specialty clinics (A), (2) chronic daily headache syndrome patients from one clinic (H), and (3) an epilepsy group from the membership of 2 epilepsy foundations (E). Site investigators identified potentially eligible A and H subjects. E subjects received announcements in foundation newsletters. Interested individuals in all groups were asked to call a toll-free phone number.

Study 2 applied the study 1 condition criteria and a monetary incentive. Additionally, A subjects had moderate to severe functional limitations according to phone responses to the SF-36 Physical Functioning scale (ie, ≥2 "limited a little" responses, or 1 "limited a lot" response). E subjects reported ≥1 seizure in the past year. H subjects had clinic-documented impairments (eg, sleep disturbance). Of 188 screened, 133 enrolled. The final sample size was 121 (nonresponse=12; 9%).

Measurement. Study 1 and 2 subjects completed a background questionnaire assessing employment, health status,[14] comorbidities,[22] condition-specific and generic symptoms,[22,23] and demographics.

Additionally, study 1 subjects were required to complete their assigned WLQs (2-week or 4-week) and 4 weekly mail-out/mail-back diaries. Materials were mailed simultaneously for matched subjects.

To minimize the threat of repeated administration bias from completing diaries and questionnaires, we divided the 40-item pool among the 2 forms. Each form contained 5 WLQ dimensions with ≥2 items per dimension. We tried to equalize item content across forms, giving the diaries 18 items and the questionnaires 22 items (column 2, Appendix 2).

The study 2 sample completed a mail-out/mail-back WLQ (with a 2-week reporting period) containing the same 5 dimensions and 40 items, as well as 8 items suggested by the research team (column 3, Appendix 2). We also measured work absences and work hours, job effectiveness on symptom days ("0% not at all effective" through "100% completely effective"), and 2 work productivity items ("In the past 2 weeks, did you produce less than the required amount of products or services," and "did you produce less than the required quality of products or services?" "If yes, was this due to your health?"). Late responders received a call and/or second mailing.

Analyses

Study 1. Before performing the main analyses, we determined whether the 5 hypothesized WLQ scales met scaling assumptions established by classical test theory. MAP-R software was used.[24] Results suggested that 4 scales were present: Time, Physical, Mental-Interpersonal, and Output Demands. Scale Cronbach alphas[25] ranged from 0.90 ($\kappa = 7$) to 0.96 ($k = 11$).

Next, a scoring algorithm was created especially for these tests, incorporating both WLQ and diary data: (1) Scores for items administered weekly or biweekly were averaged across administration weeks; (2) the resultant average scores for items within a scale were summed, and the sum was divided by the total number of scale items (the summated average scale score ranged from 0–4); and (3) scores were multiplied by 25, generating a scale score of 0 (least limited) to 100 (most limited). "Does not apply to my job" responses were treated as missing. Thus, an Output Demands scale score of 30, for example, indicated that the respondent was limited in performing these demands during 30% of the reporting period.

Two-Week Versus Four-Week Recall. Both the 2-week and 4-week versions of the WLQ were assessed with regard to recall error. In 8 models (4 scales times 2 WLQ versions), the dependent variable was a scale score to which each subject contributed 2 data points: 1 score reflecting aggregated weekly diary data, and a corresponding score utilizing questionnaire data. The explanatory variables were indicators for "subject" and "method" (diary versus questionnaire).

F statistics and probability values generated by 2-way analysis of variance (ANOVA) indicated the significance of subject and/or method in explaining WLQ scores. An intraclass correlation coefficient (ICC) ≥0.70 indicated acceptable scale performance.[26]

Bias by Week. This second recall error test addressed the degree to which WLQ responses reflected limitations from all weeks within the specified reporting period. Ideally, responses should include information equally from all weeks.

With multiple linear regression, the dependent variable of each model was a WLQ scale score from a specific questionnaire administration (the first administration of the 2-week version, the second administration of the 2-week version, or the single administration of the 4-week version). The independent variables were work limitation scale scores reported on parallel diary weeks (eg, weeks 1–2 for the first administration of the 2-week WLQ, weeks 1–4 for the 4-week WLQ).

Regressions compared the relative influence of each week within the reporting period. Because results indicated that WLQ scores were explained mainly by events from the most recent week, subsequent regressions tested the importance of the most recent diary week versus the mean of all diary weeks in the reporting period (ie, whether scores reflected recent events and/or the average across weeks). Twenty-four models were tested (3 WLQs times 4 scales times 2 comparisons).

Case-Control Comparison. To test construct validity, the mean difference in each WLQ scale score between matched patient-coworker pairs was analyzed with paired *t* tests.

Study 2

Scale Reliability. Using MAP-R, the following characteristics of the 48-item WLQ were evaluated: (1) scale means, SDs, and floor (minimum) and ceiling (maximum) effects; (2) item-to-total scale correlations corrected for overlap; (3) Cronbach's alphas for internal consistency reliability; and (4) scaling success rates (percent of tests out of all possible tests in which the correlation of an item with its hypothesized scale is ≥2 standard errors higher than its correlation with other scales). Success rates ≥90% are considered excellent. Scale scores were the means of item responses within each scale multiplied by 25.

Next, we attempted to create a shorter WLQ without sacrificing content, validity, and reliability. From the 48-item pool, 25 were chosen and tested (column 3, Appendix 2). They were selected for 3 reasons: excellent MAP-R results, significant correlation with productivity variables, and unduplicated content.

Construct Validity. In separate multiple linear regression models adjusted for age and gender, we tested the relationship of each WLQ scale score to the SF-36 Role/Physical scale (limitations resulting from physical health) and Role/Emotional scale (limitations resulting from emotional problems). We also assessed whether WLQ scores varied by condition (A, H, and E) using age- and gender-adjusted ANOVA.

Relative Validity. The association between each WLQ scale to self-reported work productivity (the sum of responses to the 2 productivity items) was compared with those of the following measures: percent of time absent because of health, effectiveness on symptom days (both for the past 2 weeks), and the SF-36 Role Limitation scales. Relative validity was quantified as a ratio of F statistics obtained from multiple linear regression. The numerator was the F statistic obtained from regressing work productivity on a specific scale. The denominator was the F value for the best scale in the comparison (maximum ratio = 1).

Results

Study 1

Two-Week Versus Four-Week Recall. Performance on this recall error test varied by scale and version (Table 1). The Time and Mental-Interpersonal Demands scales (2-week and 4-week versions) both exceeded the ICC criterion. The Physical and Output Demands scales, 4-week version, met the criterion, but method contributed in several models. Method had a small impact compared with subject. Initially, the ICC standard was not met by the Physical or Output Demands scales 2-week version (Physical = 0.64; Output = 0.58). However, 2 subjects with logically inconsistent data were excluded, and the criterion was met (Physical = 0.69; Output = 0.74).

Bias by Week. In 12 models assessing the degree to which data from individual weeks predicted WLQ scores, the most recent week tended to have the most influence (Table 2). When the most recent week was compared with the mean of the weeks, both variables were important. In 3 models, only the mean was significant ($P \leq 0.05$); in 2 models, only the most recent week was significant; and in 2 models, both were significant. In 5 of the 2-week version models, neither variable was significant. Thus, subjects tended to respond by reporting the average amount of the time they were limited during the reporting period and/or those limitations that occurred most recently. While results suggest that it is better to use a shorter reporting period such as a 2-week interval, the 4-week version also performed satisfactorily.

Case-Control Comparison. On each WLQ scale, patients had significantly higher (worse) work limitation scores than control subjects (Figure 1). The unmatched patient group had the highest WLQ scores, indicating the most limitation of the groups.

Study 2

Scale Reliability. On the 48-item WLQ, the percentages for "limited none of the time," "a slight bit of the time," "some of the time," "most of the time," and "all of the time" were 47.8%, 30.8%, 10.6%, 6.8%, and 3.8%, respectively. The frequency of "does not apply to my job" responses was small (range, 0–5 subjects per item).

The analysis confirmed 5 scales (Table 3). With a small number of exceptions, the correlation of each item to its hypothesized scale was ≥2 standard errors higher than its correlation with other scales, item-to-total scale correlation coefficients surpassed 0.40, and alphas were ≥0.90.

When the 25-item subset was assessed, the percentage of Interpersonal scale responses at the floor (zero) increased unacceptably. We tested whether its items could be combined with the Mental Demands scale. MAP-R results supported a 4-scale solution: Time, Physical, Mental-Interpersonal, and Output Demands (Table 3).

Construct Validity. In separate regression models, each WLQ scale explained a significant portion of the variance in the SF-36 Role/Physical scale, and 3 WLQ scales explained a significant amount of the variation in the SF-36 Role/Emotional scale (Table 4). The WLQ Physical Demands scale was appropriately unrelated to emotional disability.

TABLE 1. Study 1, Recall Error Test: WLQ Responses Compared With Concurrent Weekly Diary Data

	WLQ Scales							
	Time Demands		Physical Demands		Mental-Interpersonal Demands		Output Demands	
	2 wk	4 wk	2 wk	4 wk	2 wk	4 wk	2 wk	4 wk
Variables								
Subject (*df* = 28 or 35)	9.2‡	11.5‡	4.6‡	15.9‡	9.0‡	23.0‡	3.8†	20.1‡
Method (*df* = 1)	1.2	0.8	5.1*	9.0†	0.2	1.9	0.6	6.5*
Model								
r^2	0.91	0.92	0.83	0.94	0.90	0.96	0.80	0.95
F (*df* = 29 or 36)	9.0‡	11.3‡	4.6‡	15.7‡	8.7‡	22.4‡	3.7‡	19.7‡
ICC	0.80	0.84	0.64	0.88	0.80	0.92	0.58	0.90

Numbers in the body of the table are F statistics denoting associations between subject or method and WLQ scores. With 2 subjects removed from models, ICC = 0.69 for Physical Demands and 0.74 for Output Demands. Two-week group: n = 58 observations, 29 subjects; 4-week group: n = 72 observations, 36 subjects; total: 130 observations, 65 subjects.

*P <0.05, $^{\dagger}P$< 0.01, ‡ <0.001.

WLQ scores varied significantly by condition (Figure 2). Additionally, within each scale, the pattern of limitation was logically consistent with the characteristics of the different conditions. For example, headache syndrome involves sleep disturbance, fatigue, and extreme pain, which disrupt activities. H was the more limited than A (P = 0.02) or E (P <0.001) on the Time Demands scale. Headaches also involve visual and neurologic disturbances, depressed affect, and irritability. Compared with either A or E, H was most limited on the Mental-Interpersonal Demands scale (both P <0.01). On the Physical Demands scale, A was more limited than H (P <0.001) or E (P = 0.03).

Relative Validity. The WLQ Output Demands scale was the best predictor of productivity loss (Figure 3). The WLQ Mental-Interpersonal Demands and the SF-36 Role Limitation scales each exhibited half the predictive power of the Output Demands scale. The remaining measures had poorer predictive power.

Discussion

The WLQ is a reliable and valid self-report instrument for measuring the degree to which chronic health problems interfere with ability to

TABLE 2. Study 1, Relationship of WLQ Scores to Diary Data From Concurrent Weeks: Multiple Linear Regression Results

	WLQ Version					
	2-Week Recall, Week 2		2-Week Recall, Week 4		4-Week Recall, Week 4	
Predictor Variables	Week 1 vs Week 2	Mean 1 + 2 vs Week 2	Week 3 vs Week 4	Mean 3 + 4 vs Week 4	Week 1 vs 2 vs 3 vs 4	Mean 1–4 vs Week 4
WLQ scales						
Time demands	Week 1† Week 2*	Mean	NS	NS	Week 3	Mean
Physical demands	Week 2	NS	Week 4	NS	NS	Mean
Mental-interpersonal demands	Week 2	Week 2	Week 4	NS	Week 4	Mean* Week 4†
Output demands	Week 2	NS	NS	Week 4	Week 4	Mean* Week 4†
n	29		29		36	

Variables in cells are statistically significant predictors of questionnaire scores ($P \leq 0.05$). If both variables in models were statistically significant, values are as follows: *$P \leq 0.05$; $^{\dagger}P \leq 0.01$.

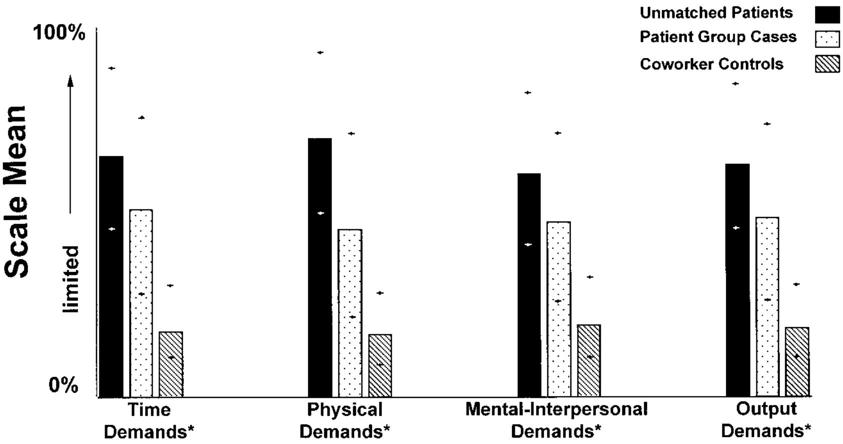

FIG. 1. Study 1. WLQ scores. Shown are means±95% CIs for job-matched cases and control subjects and unmatched patient sample. Probability values are from case-control matched pair *t* tests (n = 14 pairs). *Time Management: *t* = 3.1, *P* = 0.008; Physical Demands: *t* = 2.4, *P* = 0.032; Mental-Interpersonal Demands: *t* = 2.3, *P* = 0.040; Output Demands: *t* = 2.4, *P* = 0.031. Unmatched patients = 32.

perform job roles. Unlike available questionnaires, it addresses the content of the job through a demand-level methodology.

The WLQ performed well in studies 1 and 2. The study 1 diary/questionnaire comparison, while small and involving multiple comparisons, demonstrated that compared with diary data, both the 2-week and 4-week WLQs were relatively unbiased. However, the questionnaire responses were related more strongly to the most recent week of

TABLE 3. Study 2, WLQ Scaling Test Results

Scale	Items, n	Mean*	SD	% Floor	% Ceiling	Scale α	Range of Item-to-Total Correlations	% Scaling Success†
48-Item WLQ								
Time demands	9	32.5	27.4	10.1	0	0.89	0.51–0.77	97.2
Physical demands	11	32.4	34.5	19.3	1.8	0.96	0.72–0.88	100
Mental demands	14	32.9	25.0	6.4	0	0.94	0.34–0.82	96.4
Interpersonal demands	7	21.4	26.2	27.5	0	0.91	0.53–0.81	96.4
Output demands	7	25.7	26.2	15.6	0	0.91	0.60–0.80	100
25-Item WLQ								
Time demands	5	36.6	35.3	13.9	0.9	0.89	0.61–0.82	100
Physical demands	6	32.2	33.3	20.9	1.7	0.89	0.63–0.79	100
Mental-interpersonal demands	9	28.8	25.5	14.8	0	0.91	0.57–0.83	100
Output demands	5	26.0	26.0	16.5	0	0.88	0.53–0.82	100

n = 121.

*Minimum scale score (least limited) = 0; maximum scale score (most limited) = 100.

†Scaling success is the percent of tests out of all possible tests in which the correlation of an item with its hypothesized scale is ≥2 SEs higher than its correlation with other scales.

Table 4. Study 2, WLQ Construct Validity With SF-36 Role Limitation Scales: Multiple Regression Adjusted for Age and Gender

	Past 4-wk Role Limitations/Physical				Past 4-wk Role Limitations/Emotional			
WLQ Scale (Past 2 wk)	Estimate	SE	Scale P	r^2	Estimate	SE	Scale P	r^2
Time demands	−0.402	0.13	0.002	0.15*	−0.334	0.131	0.013	0.07*
Physical demands	−0.445	0.14	0.002	0.16*	−0.115	0.144	NS	0.02
Mental-interpersonal demands	−0.516	0.17	0.004	0.14*	−0.722	0.170	0.0001	0.15*
Output demands	−0.769	0.17	0.0001	0.22*	−0.771	0.168	0.0001	0.17*

n = 120, missing value = 1.

Low scores on WLQ indicate less limitation. Low scores on SF-36 indicate more limitation.

*Models are significant at $P \leq 0.05$.

the reporting period than to earlier weeks. The ease of remembering recent events may reflect the difficulty of the response task. Respondents must remember and integrate information about their health and work simultaneously. We recommend the 2-week WLQ to maximize accuracy. However, if it is important to match time periods across instruments within a study, the 4-week version is acceptable. In such situations, a single administration of the 4-week WLQ would achieve better precision than a single administration of the 2-week WLQ, and cost less than multiple administrations of the shorter version.

Study 2 indicated that the 25-item WLQ was reliable and valid for use among several different job and chronic condition groups. However, our sample included only adults working ≥20 h/wk, possibly excluding employed individuals with severe work limitations, and only certain diagnostic groups. The 25-item WLQ has been evaluated in additional patient and employee samples, and it has demonstrated excellent performance (data available from authors).

The analyses also confirmed 4 distinct dimensions of on-the-job disability (limitations handling Time, Physical, Mental-Interpersonal, and Output Demands). The multidimensionality of the WLQ is likely to appeal to clinicians, other disability management professionals, and employers. Because the WLQ is context specific and focused on

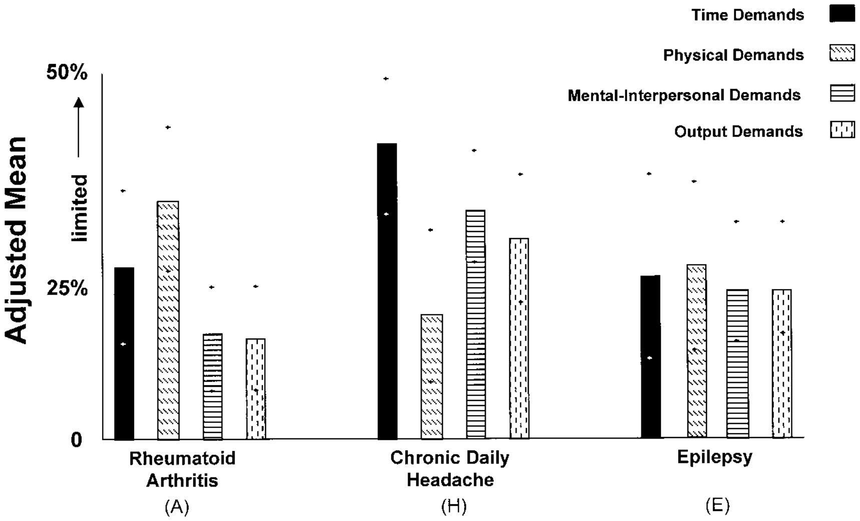

Fig. 2. Study 2. Adjusted WLQ scores. Shown are means±95% CIs by condition group. Values are adjusted for age and gender (n = 121). ANOVA results are as follows. Time Demands scale: F = 5.19; $P = 0.007$; Physical Demands scale: F = 7.58, $P = 0.0008$; Mental-Interpersonal Demands scale: F = 8.59; $P = 0.0003$; Output Demands scale: F = 4.15, $P = 0.0181$.

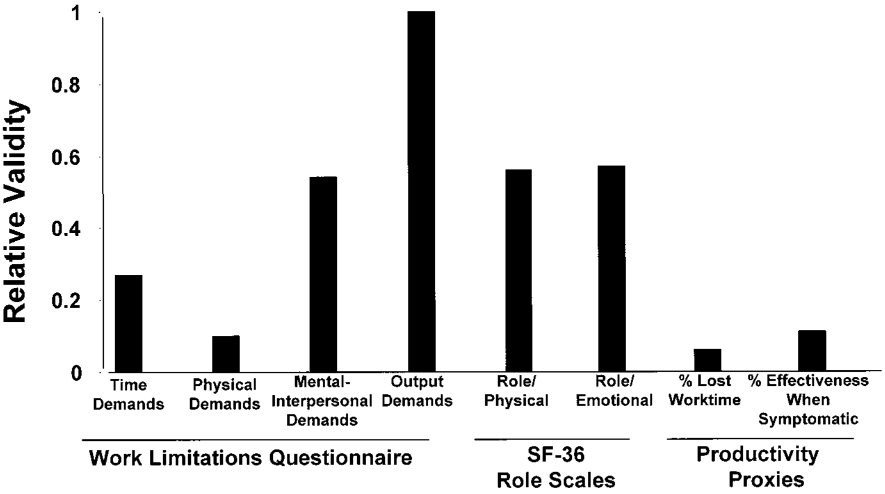

FIG. 3. Study 2. Relative validity for predicting self-reported work productivity loss. Relative validity is a ratio of F values (maximum = 1). The numerator is the F value from regressing self-reported work productivity on a scale included in the comparison. The denominator is the F value for the best scale in the comparison. Models are adjusted for age and gender (n = 121). SF-36 is based on 4-week recall. WLQ and productivity proxies are 2-week recalls.

job demand performance, it can be used to identify both the magnitude and type of impact that health problems are having in the workplace. In contrast, role disability scales are pitched at too high a level of generality to be of practical value. Moreover, construct validity test results indicated that the WLQ Output Demands scale had superior performance for predicting productivity. The Mental-Interpersonal Demands and the SF-36 Role Limitation scales had moderate validity. Thus, the WLQ provides more specific information than available instruments while increasing the depth and breadth of information generated. However, there is a trend in health status assessment toward using summary scores, and future WLQ users may prefer a similar approach.

While this project involved multiple psychometric assessments, our tests stopped short of addressing certain issues. We did not attempt to measure abilities that exceed demands, the positive end of the ability spectrum. We did not assess test-retest reliability and responsiveness to change within condition groups. The value of the WLQ as a productivity indicator was addressed briefly; criterion validity tests linking scores to objective work output were not performed. We did not explore how job demand variations may impact WLQ data. Finally, we did not fully assess our scoring method, which combines within-scale limitations by averaging them. Ideally, a scale would capture the intensity of each limitation measured and its frequency; however, this may result in a cumbersome instrument.

Study results provide important evidence of the reliability and validity of the WLQ. It is a promising new tool for assessing chronic health problems and their social and economic impact.

Acknowledgments

We wish to acknowledge Glaxo Wellcome, Inc, Research Triangle Park, North Carolina, for its sponsorship of the research project and the Henry J. Kaiser Family Foundation of Palo Alto, California. We also wish to express our gratitude to the following site investigators: Leonard Sicilian, MD, Bruce Ehrenberg, MD, David Adler, MD, Peter Bonis, MD, Anne Marie Brown, BS, RN, Laurie Olans, MD, and Arthur A. Wills III, MD, all from the New England Medical Center; Saralynn Allaire, ScD, of the Boston University School Of Medicine Multi-Purpose Arthritis Center; and Lawrence C. Newman, MD, and Margie Russell, RN, of the Montefiore Hospital Headache Clinic. Additionally, we wish to extend our appreciation to Anita Wagner, PharmD, and Constance Kelley for their invaluable participation and advice throughout the study. Patients in this study were recruited from the New England Medical Center (Respiratory, Gastroenterology, Psychiatry, and Neurology departments), Downtown Medical Associates (a New England Medical Center affiliate), Boston University, The Epilepsy

Foundation of Massachusetts and Rhode Island, the Epilepsy Foundation of Connecticut, and the former Massachusetts Respiratory Hospital in Weymouth, Massachusetts. We gratefully acknowledge the participation of each.

References

1. **Kaye HS, LaPlante MP, Carlson D, et al.** Trends in disability rates in the United States, 1970–1994. Disabil Stats Abstract 1996;17:1–6.

2. **Pope A, Tarlov AR.** Disability in America: Toward a national agenda for prevention. Washington, DC: Division of Health Promotion and Disease Prevention, Institute of Medicine, National Academy Press; 1991.

3. **Lerner DJ, Amick BC III, Malspeis S, et al.** A national survey of health-related work limitations among employed persons in the United States. Disabil Rehabil 2000;22:225–232.

4. **Hoffman C, Rice D, Sung HY.** Persons with chronic conditions: Their prevalence and costs. JAMA 1996;276:1473–1479.

5. **Freudenheim E, ed.** Chronic care in America: A 21st century challenge. Princeton, NJ: Robert Wood Johnson Foundation; 1996.

6. National Institute for Occupational Safety and Health. National occupational research agenda. Cincinnati, Ohio: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health; 1996. DHHS publication No. 96-115.

7. **Lerner D, Lee J.** Measuring health-related work productivity with self-reports. In: Stang P, Kessler RC, eds. Health and work productivity: Emerging issues in research and policy. Chicago, Ill: University of Chicago Press. In press.

8. **Lerner DJ, Bungay KM.** Measuring work outcomes. In: Pharmacoeconomics and outcomes: Applications for patient care, module 3: Assessment of humanistic outcomes. Kansas City, Mo: American College of Clinical Pharmacy; 1997:171–185.

9. **Ormel J, Von Korff M, Oldehinkel AJ, et al.** Onset of disability in depressed and non-depressed primary care patients. Psychol Med 1999;29:847–853.

10. **Greenberg PE, Stiglin LE, Finkelstein SN, et al.** The economic burden of depression in 1990. J Clin Psychiatry 1993;54:405–418.

11. **Johns G.** Absenteeism estimates by employees and managers: Divergent perspectives and self-serving perceptions. J Appl Psychol 1994;79:229–239.

12. **Osterhaus JT, Gutterman DL, Plachetka JR.** Healthcare resource and lost labor costs of migraine headache in the US. Pharmacoeconomics 1992;2:67–76.

13. **Reilly MC, Zbrozek AS, Dukes EM.** The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353–365.

14. **Ware JE, Snow KK, Kosinski M, et al.** SF-36 Health Survey: Manual and interpretation guide. Boston, Mass: Health Institute, New England Medical Center; 1993.

15. **Krueger RA.** Group dynamics and focus groups. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia, Pa: Lippincott-Raven; 1996:397–402.

16. Employment and Training Administration, US Employment Service. Dictionary of occupational titles. 4th ed. Washington, DC: US Department of Labor; 1991.

17. **McCormick EJ, Jeanneret PR, Mecham RC.** The Position Analysis Questionnaire (PAQ). Palo Alto, Calif: Consulting Psychologists Press Inc; 1989.

18. **Sudman S, Bradburn NM, Schwarz N.** Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, Calif: Jossey-Bass Publishers; 1996.

19. **Streiner DL, Norman GR.** Health measurement scales: A practical guide to their development and use. 2nd ed. New York, NY: Oxford University Press; 1995.

20. **Berwick D, Murphy J, Goldman P, et al.** Performance of a five-item mental health screening test. Med Care 1991;29:169–176.

21. US Department of Commerce. Classified index of industries and occupations. Washington, DC: US Government Printing Office; 1992.

22. **Stewart A, Ware JE Jr, eds.** Measuring functioning and well-being. Durham, NC: Duke University Press; 1992.

23. **Baker GA, Smith DF, Dewey M, et al.** The development of a seizure severity scale as an outcome measure in epilepsy. Epilepsy Res 1991;8:245–251.

24. **Ware JE, Harris WJ, Gandek B, et al.** MAP-R for Windows: Multitrait/multi-item analysis program. Boston, Mass: Health Assessment Lab; 1997. Computer program.

25. **Cronbach LJ.** Essentials of psychological testing. 4th ed. New York, NY: Harper and Row; 1984.

26. **Shrout PE, Fleiss JL.** Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:40.

Appendix 1

Table 5 of Appendix 1 gives the sample characteristics.

TABLE 5. Sample Characteristics

	Pilot Studies			Study 1 Recall Error and Construct Validity	Study 2 Scale Reliability and Construct Validity
	Focus Groups	Cognitive Interviews	Alternate Forms Comparison		
Sample, n	18	37	36	65	121
Gastrointestinal, %	27.8	23.2	13.2	21.5	. . .
Psychiatric, %	16.7	25.6	26.3	15.4	. . .
Respiratory, %	22.2	25.6	26.3	21.5	. . .
Epilepsy, %	33.3	25.6	34.2	15.4	29.8
Rheumatoid arthritis, %	. . .	. . .	. . .	. . .	38.8
Chronic daily headache, %	. . .	. . .	. . .	. . .	31.4
Controls, %	. . .	. . .	. . .	26.2	. . .
Demographics					
Male, %	50.0	33.3	34.2	27.3	27.3
White, %	94.1	92.1	84.2	90.9	85.1
Married, %	33.3	48.7	36.8	47.0	48.8
Age, mean (SD)	39.9 (11.4)	45.4 (13.1)	41.3 (10.1)	41.3 (11.1)	43.0 (10.0)
Education, mean (SD)	14.9 (2.2)	14.0 (3.0)	13.8 (2.8)	14.9 (2.0)	15.1 (1.9)
Income, $1,000, mean (SD)	31.6 (27.6)	35.6 (24.5)	37.4 (25.3)	33.3 (22.0)	42.5 (23.0)
Health status, mean (SD)					
Comorbid conditions, mean (SD)*	0.4 (0.8)	0.6 (0.7)	0.2 (0.5)	0.3 (0.7)	0.3 (0.5)
SF-36 scales, mean (SD)					
Physical functioning	76.9 (18.6)	84.2 (19.8)	85.7 (15.7)	81.9 (22.4)	72.2 (25.0)
Role/physical	51.4 (40.6)	50.6 (45.7)	76.4 (34.8)	69.8 (40.2)	45.9 (38.6)
Pain	64.7 (25.4)	70.2 (23.4)	76.9 (18.3)	76.5 (20.8)	64.0 (18.8)
General health	46.5 (19.5)	59.6 (25.3)	60.6 (21.3)	56.9 (25.1)	53.6 (22.1)
Vitality	40.3 (22.5)	49.7 (24.6)	47.8 (23.4)	49.9 (23.6)	43.9 (20.6)
Social functioning	64.6 (22.4)	68.9 (25.3)	68.1 (26.6)	71.0 (27.1)	55.9 (24.6)
Role/emotional	53.7 (47.3)	64.9 (40.2)	73.0 (36.7)	79.8 (34.0)	69.3 (37.6)
Mental health	62.7 (18.1)	65.8 (20.0)	63.4 (22.3)	71.3 (19.6)	67.0 (17.9)
Work measures, %					
Full time (≥31 h)	82.4	89.7	76.3	89.4	86.7
Occupation					
Nonmanual	33.3	25.6	35.1	32.3	45.5
Service	55.6	61.5	59.5	60.0	49.6
Manual	11.1	12.8	5.4	7.7	5.0
Company size, n					
<100	22.2	38.9	29.7	31.8	30.8
100–499	33.3	25.0	18.9	19.7	18.8
≥500	38.9	33.3	43.2	28.8	40.2
Don't know	5.6	2.8	8.1	19.7	10.3
Time with company, y					
<1	29.4	7.7	23.7	19.7	12.4
1–5	29.4	23.1	31.6	27.3	35.5
6–10	5.9	25.6	18.4	19.7	15.7
>10	35.3	43.6	26.3	33.3	36.4

*Medically diagnosed conditions include hypertension, myocardial infarction, congestive heart failure, diabetes, angina, and cancer.

APPENDIX 2. TABLE 6. Items Tested and/or Retained for WLQ

Cognitive Interview Items	Study 1 Items	Study 2 Items
Time demands		
Get to work on time	D	T
Work required hours	D	T*
Work required days	. . .	. . .
Stay within sick, vacation, personal day limits	. . .	. . .
Get going beginning of work day	Q	T*
Start on work soon after arriving	Q	T*
Work without breaks or rests	D	T*
Stick to routine/schedule	Q	T*
Give tasks time needed	Q	T
Adjust to work pace changes	. . .	. . .
Put off tasks	Q	T
Let work pile up	. . .	. . .
Put in extra hours to keep up	. . .	. . .
Pace yourself	. . .	. . .
Work without watching clock	. . .	. . .
Item added during study 2		
Stop before work is finished		T
Physical demands		
Get to work from parking, bus, train	Q	T
Walk/move around work locations	D	T*
Lift, carry, move objects	See below	. . .
Walk >1 block, climb flight of stairs	Q	T
Sit, stand, stay in 1 position	D	T*
Work in awkward or unusual positions	. . .	. . .
Repeat motions	D	T*
Bend, twist, or reach	Q	T*
Use handheld tools, equipment	Q	T*
Use upper body to operate tools, equipment	Q	T
Use lower body to operate tools, equipment	Q	T
Items added during study 1		
Lift, carry, move objects ≤10 lb	D	T
Lift, carry, move objects, ≥10 lb	D	T*
Mental demands		
Keep mind on work	D	T
Keep track of >1 task	Q	T
Think clearly	D	T*
Remain alert	Q	T
Work carefully	Q	T*
Do precise work	. . .	. . .
Concentrate on work	Q	T*
Remember things important for work	D	T
Avoid confusion	. . .	. . .
Handle demanding/stressful work	D	T
Adjust to high-pressure periods	. . .	. . .
Maintain morale during demanding/stressful periods	. . .	. . .
Become tense/frustrated	Q	T
Remain calm	. . .	. . .
Stay interested in job	. . .	. . .

Continues

Appendix 2. Table 6. (Continued)

Cognitive Interview Items	Study 1 Items	Study 2 Items
Items added during study 2		
Learn new things on the job		T
Work without watching clock		T
Lose train of thought		T*
Personal problems affect work		T
Easily read/use eyes (see below)		T*
Information processing		
Easily read/use eyes	. . .	. . .
Understand written instructions, assignments	. . .	. . .
Understand spoken instructions, assignments	. . .	. . .
Work around noise/activity	. . .	. . .
Interpersonal demands		
Speak in person/on phone	D	T*
Control irritability/anger	D	. . .
Get along	. . .	. . .
Keep your cool	. . .	. . .
Work near others	Q	T
Communicate well	Q	T
Be supportive	. . .	. . .
Maintain contacts	. . .	. . .
Item added during study 1		
Help others to work	Q	T*
Items added during study 2		
Control temper		T*
Present your ideas		T
Limit contact with others		T
Output demands		
Handle workload	D	T*
Work fast enough	D	T*
Finish all work	Q	. . .
Finish work on time	Q	T*
Meet simultaneous demands	. . .	. . .
Put in extra hours to keep up	. . .	. . .
Work without mistakes	. . .	T*
Do work over	. . .	. . .
Work safely	. . .	. . .
Satisfy others	D	T
Feel sense of accomplishment	D	T
Item added during Study 1		
Do all you're capable of	Q	T*
Work environment		
Work in available physical conditions	. . .	. . .
Work without fresh air	. . .	. . .
Work in hot, cold, damp	. . .	. . .
Work with fumes, odors, smells	. . .	. . .
Work near bright/flashing lights	. . .	. . .
Work close to others	. . .	. . .

D indicates Study 1 diary item; and Q, questionnaire item. Ellipses indicate item excluded from test. Cognitive interviews = 70 items. Study 1 = 40 items. Study 2 = 48 items. Final WLQ = 25 items.

*T indicates tested and included in final WLQ (2-week recall version).