MAKING A CASE FOR QUALITY:
THE BUSINESS CASE FOR TRICARE HEALTH PROMOTIONS PROGRAMS

HEALTH AND WORK PERFORMANCE QUESTIONNAIRE (HPQ) DOD EDITION

BRIEFING BOOK

Submitted on: September 30, 2006

Submitted to: Health Program Analysis & Evaluation Division,
TRICARE Management Activity
5111 Leesburg Pike
Falls Church, VA 22041

Contract No.: HHSP23320045016XI

JOHN SNOW, INC.
44 FARNSWORTH STREET
BOSTON, MASSACHUSETTS 02210
VOICE: 617-482-9485
FAX: 617-482-0617
EMAIL: JSINFO@JSI.COM

This document was prepared by **John Snow, Inc**. in collaboration with **Harvard Medical School**, for TRICARE Management Activity (TMA) and funded by the U.S. Department of Defense under contract number HHSP23320045016XI. This document was written by James Maxwell, Ph.D., Eugenie Coakley, MPH, MA, Milly Krakow, Ph.D., and Emilie Pinkham, BA, for John Snow, Inc in collaboration with Ronald Kessler, Ph.D. of Harvard Medical School.

The web application for this project was developed and programmed by DataStat, an Ann Arbor, MI based survey research firm.

John Snow, Inc., and its nonprofit affiliate JSI Research & Training Institute, Inc., are public health research and consulting firms committed to improving public health and the functioning of the health care delivery systems in the United States and around the world. JSI conducts research on health policy as well as on a variety of clinical topics. For more than 25 years, Boston-based JSI and its other affiliates have provided high-quality research, technical, and managerial assistance to public and private sector clients worldwide. JSI has implemented projects in 97 countries, and currently operates from seven U.S. and 44 international offices.

Prepared September 2006 by John Snow, Inc.
44 Farnsworth Street, Boston, MA 02210-1211
JSI URL: http://www.jsi.com/
To obtain additional information, contact
Telephone: (617) 482-9485;
Fax: (617) 482-0617; Email: jsinfo@jsi.com

TABLE OF CONTENTS

ADMINISTRATIVE INFO AND APPROVALS

IRB MATIERALS
Claim of Exemption Application 4
NEIRB Exemption Approval Letter 8
DMDC APPLICATION 9
CITI TRAINING 20
SUMMARY OF POTENTIAL HUMAN SUBJECTS RESEARCH 22
ANONYMOUS DATA ASSURANCE 23
SAMPLE NOTIFICATION/REMINDER EMAILS 24

HEALTH AND WORK PERFORMANCE QUESTIONNAIRE

ORIGINAL HPQ SURVEY 28
ATLANTA HPQ SURVEY 40
THE WHO HEALTH AND WORK PERFORMANCE QUESTIONNAIRE 67
CHRONIC MEDICAL CONDITIONS AND WORK PERFORMANCE 86
USING THE HPQ TO EVALUATE COSTS OF ILLNESS 95
RATIONALIZING EMPLOYER HEALTH CARE SPENDING (ATLANTA HPQ) 110

INSTRUMENT VALIDATIONS AND RELATED LITERATURE

ADHD
ADHD Self-Report Scale 128
Prevalence and Effects of Adult ADHD (ASRS Validation) 131

CHRONIC FATIGUE SYNDROME
CIDI – Stem Questions, Neurasthenia Section 139

INSOMNIA
CIDI – Stem Questions, Chronic Conditions Section 140
Sleep Problems, Comorbid Mental Disorders, and Role Functioning 144

MENTAL DISTRESS
K6 Self-Administered Scale 172
K10 Self-Administered Scale 174
Short Screening Scales to…Psychological Distress 176

POST TRAUMATIC STRESS DISORDER (PTSD)
DD Form 2796 194

PERSONALITY/RISK TAKING BEHAVIORS
A Comparison of Three Structural Models for Personality 198

MOOD DISORDER QUESTIONNAIRE VALIDATION 210
EFFECTS OF MOOD DISORDERS ON WORK PERFORMANCE 213

Claim of Exemption

Please complete this form and send to New England IRB, 40 Washington Street, Suite 130, Wellesley, MA 02481 or fax to New England IRB at 781-237-0330. Contact New England IRB at 781-431-7577 with any questions.

Principal Investigator: Phone: 617 482-9485

James Maxwell , Ph.D.
Project or Protocol Title:

Making A Case For Quality: The Business Case for TriCare Health Promotions Programs

Contact person: James Maxwell

Address:

44 Farnsworth Street

Boston, MA 02210-1211

Phone: 617-482-9485

Fax: 617-482-0617

E-mail:Maxwell@jsi.com

SPONSOR / FUNDING INFORMATION

Will this project/protocol be supported by an external funding agency? ☐ No ☒ Yes

Department of Defense

List Sub-investigator/Co-investigators None ☐

Ronald Kessler Ph.D., Harvard University

Eugenie Coakley, M.S., Senior Statistician, JSI

Tom Wilkinson, Executive Vice President, DataStat

Faculty Sponsor (if applicable)

LOCATION OF RESEARCH:

Where will the study take place? *

Fort Bragg, North Carolina

Norfolk Naval Base, Virginia

MCB Camp Lejeune, North Carolina

Langley Air Force Base, Virginia

*Include all locations for study related activities here or on separate sheet.

Will the PI be conducting and/or supervising study related activity at any sites not under the jurisdiction of this IRB? If yes, please provide name and address for each location AND documentation of approval to conduct research at these sites. Note: Additional IRB approval may be required from these sites if an individual at this site is obligated to use another IRB. ☒ No ☐ Yes

EXEMPT CATEGORY CLAIMED

Identify all that apply to your research (check applicable boxes)

1. ☐ Research conducted in established or commonly accepted educational settings, involving normal educational practices. This category may include children.

2. ☐ Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement) for which subjects cannot be identified, or release of the information would not be harmful to the subject. This category may include children.

3. ☒ Research involving the use of survey procedures or interview procedures or observation of public behavior for which subjects cannot be identified, OR release of the information would not be harmful to the subject. This category may not include children. If subjects are 18 years of age or younger parental consent is required. Research may be reviewed by expedited procedures – do not use this form!

4. ☐ Survey or interview of public or elected officials. Testing of public officials.

5. ☐ Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available OR if the information is recorded by the Investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

6. ☐ Research and demonstration projects that are conducted by or subject to the approval of Federal Department or Agency heads, and which are designed to study or evaluate public benefits or services (e.g., evaluation of public benefits programs: Medicare, Public Assistance). This category may include children.

7. ☐ Taste and food quality evaluation and consumer acceptance studies. This category may include children.

8. ☐ Unidentifiable human body parts, sections or samples obtained from a morgue.

If your research involves only those procedures listed in one or more of the categories above, it may be exempt. Please provide a rationale for each exempt category claimed for this research.

RATIONALE FOR EXEMPT CATEGORY CLAIMED

The information **must** include a brief specific description of the procedure(s) involving the human subjects in sufficient detail to demonstrate to the IRB reviewer that the research protocol meets the requirements for each category of exemption claimed in this human subjects research protocol. The text should be approximately 300 words or less on separate sheets in sufficient detail to allow the reviewer to judge exemption criteria.

RATIONALE FOR EXEMPT CATEGORY # (s)

SYNOPSIS OF THE PROJECT OR PROTOCOL, INCLUDE:

1. The objective of the research project and background of study.
2. The rationale for the use of the selected subject population & plans for recruitment & consent.
3. The procedures that will be performed to generate research data & risks, if any, to subjects.
4 Steps to be taken to protect the privacy and/or confidentiality of subjects.
5 Include copy of questionnaires, surveys or brief outline of questions to be asked.

INVESTIGATOR'S ASSURANCE

I certify that the information provided in this claim of exemption is complete and correct.

I understand that as Principal Investigator, I have ultimate responsibility for the protection of the rights and welfare of human subjects and the ethical conduct of this research protocol. I agree to comply with all New England IRB policies and procedures, as well as with all applicable federal, state, and local laws regarding the protection of human subjects in research, including, but not limited to, the following:

- The project will be performed by qualified personnel according to the research protocol,
- Maintaining a copy of all questionnaires, survey instruments, interview questions, data collection instruments, and information sheets for human subjects for at least three years following termination of the project,
- Necessary review by New England IRB will be sought if changes made in the research protocol may result in the research no longer meeting the criteria for exemption.

I have read and understand the above policy concerning IRB exempt protocols.

Principal Investigator _____ Date 2/2/2016

SYNOPSIS OF THE PROJECT OR PROTOCOL

1. The Department of Defense contracted with John Snow, Inc to implement the Health Performance Questionnaire (HPQ) in a sample of active-duty military personnel. Developed by Dr. Ronald Kessler at Harvard University, the HPQ has been used extensively by corporations in the United States. It helps pinpoint health problems that have high indirect costs, such as poorer work performance, for the purposes of rationalizing health care investments. The HPQ will be used to study the prevalence, causes, and consequences of injury in military bases with high injury rates. The study has three major objectives: 1) to estimate the prevalence of various injuries; 2) to identify the worker- and environment- related risk factors most correlated with injury (for example, do sleep disorders or asthma, or shift work put soldiers at higher risk of injury?); 3) to estimate the numbers of lost work days and the impact of injury on work performance.

2. The Department of Defense has chosen four military bases with high injury rates for implementing this study, one in each of the service branches. The bases were selected because the Department of Defense has a major new initiative to reduce injuries among the active duty military population. The data will be used to help develop more effective health promotion and injury reduction programs at these and other military bases. The HPQ will be administered as an anonymous survey, with no personal information collected on the questionnaire. A senior officer at each of the bases will send an email to all active military personnel asking them to participate in this voluntary survey. No more than three reminders will be sent to potential participants in the study. Participants who have completed the questionnaire will be given a secure and non-identifiable method for removing themselves from the reminder list.

3. The survey will be administered in web based format. The survey and introductory letter or email will make it clear to service members that participation is completely voluntary and responses will be anonymous. The questionnaire has been used extensively in civilian populations and involves no risks to the participants. The questionnaire is not burdensome and takes less than a half hour on average to complete.

4. No information is collected on the survey that would allow individuals to be identified in any way. The survey data will be stored in a file that does not contain the e-mail address, nor the IP address, name, or any other identifying information. The data will be stored on a secure, stand-alone computer system in a locked office in which access is strictly limited to the project director, the senior statistician, and a data analyst.

NEIRB
New England Institutional
Review Board

February 10, 2006

James Maxwell, PhD
44 Farnsworth Street
Boston, MA 02210-1211

RE: 06-025 The TriCare Health and Productivity Study

Dear Dr. Maxwell,

This is to inform you that New England Institutional Review Board (NEIRB) has reviewed your claim of exemption for the above-captioned research. NEIRB has determined that this research activity is exempt from NEIRB review.

Amendments and or changes to the research must be submitted to NEIRB for review, as changes may affect the exempt status.

Please call me if you have any questions about the terms of this determination.

M. Abel

Melinda Abelman, CIP
Assistant Lead Administrator

Copy: Alan M. Sugar, M.D., Chairman

40 WASHINGTON STREET, SUITE 130 • WELLESLEY, MASSACHUSETTS 02481 • VOICE (781) 431-7577 • FAX (781) 237-0330 • www.neirb.com

Implementation of DoD Instruction 1100.13, "Surveys of DoD Personnel"

In DoDI 1100.13, the Under Secretary of Defense (Personnel & Readiness) [USD(P&R)] tasked the Defense Manpower Data Center (DMDC) with the review of survey instruments and procedures that require participation of personnel in any DoD component other than the sponsoring Component.

I. Who Is Responsible for Licensing DoD Surveys?

A. Authority for and Scope of Responsibility. DoD Directive 8910.1 and DoD 8910.1-M assign Washington Headquarters Services/Directorate for Information Operations & Reports (WHS/DIOR) responsibility for the licensing of all information requirements that require responses from personnel in any DoD Component other than the requesting component. When such an information requirement is a survey, in accordance with DoDI 1100.13 WHS/DIOR cannot license it without DMDC's recommendation of approval. [WHS/DIOR also coordinates the clearance of all DoD data collection forms intended for members of the public and/or other Federal agencies.]

B. Materials Required by WHS/DIOR to License a Survey. WHS/DIOR requires a copy of DMDC's review to be included in documentation requesting approval and issuance of a Report Control Symbol (RCS). DIOR needs to review the survey instrument after the DMDC review to ensure that all DMDC requirements have been incorporated.

II. Materials Needed for DoDI 1100.13. The following list identifies major issues that are reviewed by DMDC in implementing DoDI 1100.13. While all of these issues need to be addressed by the submitting agency, the information does not need to comply with the ordering shown below.

A. Statement of Purpose for the Survey Effort (to determine if the survey instrument and procedures address the issues in an optimum manner)

The proposed project presents an enormous opportunity for the TRICARE program to use data obtained from the HPQ to establish baseline measures of health need and access to health care services, estimate the indirect workplace costs of high risk and chronic illnesses among the armed forces, and to estimate the value of TRICARE health promotion, and disease management programs. This project will be crucial to achieving the Department of Defense's objective of moving to a prevention based health system and maintaining the highest levels of military readiness (DoD, 2001). In addition, the improvement of TRICARE health promotion and disease management programs for active duty members can potentially yield long-term cost savings for the TRICARE for LIFE program.

In conjunction with the World Health Organization (WHO), the Harvard Medical School developed the Health and Work Performance Questionnaire (HPQ) in 2002 to estimate the workplace costs of health conditions in terms of reduced job performance, absenteeism, work-related accidents and injuries. The HPQ is a short, self-administered questionnaire to collect information about the prevalence of treated and untreated health

problems, and their effects on work performance. It can be administered either electronically or by paper, and either as a stand-alone survey or in conjunction with existing Health Risk Assessments.

The Department of Defense contracted with John Snow, Inc to implement the Health Performance Questionnaire (HPQ) in a sample of active-duty military personnel. Developed by Dr. Ronald Kessler at Harvard University, the HPQ has been used extensively by corporations in the United States. It helps pinpoint health problems that have high indirect costs, such as poorer work performance, for the purposes of rationalizing health care investments. The HPQ will be used to study the prevalence, causes, and consequences of injury in military bases with high injury rates. The study has three major objectives: 1) to estimate the prevalence of various injuries; 2) to identify the worker- and environment- related risk factors most correlated with injury (for example, do sleep disorders or asthma, or shift work put soldiers at higher risk of injury?); 3) to estimate the numbers of lost work days and the impact of injury on work performance.

B. Availability of Similar Information

1. Can the data from the proposed survey be obtained from record or other extant data?

   This data cannot be obtained from records or other extant data.

2. Is the survey redundant with other surveys?

   No.

3. Can the information be collected by piggybacking on another planned survey (e.g., a DMDC survey or another survey by the sponsoring agency)?

   No.

C. Sampling Plan

1. Is there a clear description of the population to which the findings are supposed to generalize?

Results of the study will be generalized to active-duty, military personnel – enlisted service men and women and officers – who have been members of either the Army, Navy, Air Force, or Marines for 1 year or longer. Neither Reservists, nor retired personnel, nor family members, nor members of the Coast Guard are part of the population of interest.

2. How will the sample be drawn?

Random samples will be selected within specific strata. There are 8 strata: military branch (4) and enlisted/officer status (2). *Warrant officers are classified as officers.*

3. What is the proposed sample size: overall and per cell?

Due to budget considerations, the overall sample frame size is **12,000** with an *assumed response rate of 25%.* Thus, the sample size is projected to be **3,000** completed surveys. The following table breaks the sample size down by cell:

	Army	*Navy*	*Air Force*	*Marines*
Total Sample Frame	3,000	3,000	3,000	3,000
Expected Sample	750	750	750	750
Enlisted:				
Total Sampled	1,800	1,800	1,800	1,800
Expected Returns	450	450	450	450
Officers:				
Total Sampled	1,200	1,200	1,200	1,200
Expected Returns	300	300	300	300

4. What anticipated response rate was used to derive the cell sizes for the sample?

The response rate is projected to be 25% and is a conservative figure, since this is a pilot study. However, based on a prior military study on survey response rates by Caplan and Quigley, we may actually attain a 30% response. Additional response will only improve our power to make inferences and generalizations.

5. What criteria were used to determine who will receive a survey?

As noted above, Active-duty personnel (in Army, Navy, Air Force, Marines) who have been in the military for 1 or more years are eligible for this survey. Eligible persons will then be randomly selected to receive an HPQ.

D. Analytic Considerations

1. Will data from the survey sample be weighted? If so, an explanation is needed.

The weights will be computed as the inverse of the probability of being selected and returning a survey instrument. In the population, officers comprise between 10% - 20% of active duty personnel. The weights will then be used to estimate overall (i.e., combined subgroups) statistics.

2. For which subgroups (e.g., officers vs. enlisted) will the findings be reported?

Results will be reported within branch, overall and broken out by enlisted and officer status. Results that combine branches will use the sample weights (i.e., weight as the inverse probability of being selected) to account for the fact that equal samples were drawn from the

4 branches even though they do not have equal enlistment. In the population, the Army, Navy, Air Force, and Marines comprise approximately 35%, 26%, 26%, and 13% of the armed forces, respectively.

3. What amount of confidence (e.g., 95%) and precision (e.g., ±3 percentage points) does the sponsor anticipate for the findings?

With 450 enlisted responses and 300 officer responses per branch, we will be able to estimate mean scores (within the officer or enlisted group) on the several HPQ 5-point Likert-scaled items to within **+/- 0.3** points with 95% confidence intervals. We will be able to estimate proportion of self-reported health conditions with 95% confidence with a margin of error of **+/-2-3%** when the underlying prevalence is between 5% and 10%; and with a margin of error of **+/-4-6%** when the underlying prevalence is between 10%-50%. Finally, we will be able to compare the mean productivity scores (e.g., a 5-pt Likert scale) within military branches (not stratified by officer status) of those with and without a specific condition with > 80% power to detect a 1-point difference when the common standard deviation is up to 1.75, alpha = 0.05, and the number with the condition ranges from 25 to 100 (prevalence in the 5%-25% range).

4. Does the Sampling Plan clearly support the Analytic Considerations and thus the stated objectives and requirements for the survey effort?

The sampling plan supports the analytic efforts.

E. Survey materials

1. Did the package submitted to DMDC include final or near final copies of survey instruments/ interview protocols, all cover letters, and instructions to potential respondents?

The package submitted to the DMDC includes a hard copy of the survey, which will be used, with slight modification (specifically: fitting page length to screen and automating the skip patterns rather than having text instructions) for use on the web. The package also includes all cover letters and instructions for potential respondents.

2. DMDC's review checks for basic survey concerns:
   a) Flow/eye appeal
   b) Interest
   c) Difficulty (e.g., reading and understanding instructions, questions, and response alternatives)
   d) Poorly constructed items (e.g., double-barreled, ambiguous, and misleading)
   e) Standard demographic questions

3. Have directions (e.g., "Blacken the circle" or "Mark all answers that apply") been provided to tell respondents how to complete each item or set of items.

Yes.

4. Does the wording for the demographic questions conform to the DoD-wide standards used in DMDC surveys?

   Yes

5. Does the survey contain items from copyrighted scales? If so, has the submitting agency obtained permission to use those scales?

   No

6. Has the survey developer consistently used highlighting? For example, are particular words or phases all emphasized by the same method (e.g., bolding, underlining, or all caps).

   Yes

F. Plans for Distributing and Collecting Survey Materials

1. Distribution of materials
   a) How will the survey questions be delivered to potential respondents (e.g., US mail, base mail, hand delivery by a installation POC or member of the survey team, telephone questions, or access through a computer)?

A total of 6- E-mails and up to 2 postal mailings will be sent for a 5-6 week data collection period. Solicitation letters and reminders will be signed by the Under Secretary (or other ranking official). The first e-mail will announce the survey to the total sample. This initial e-mail will include a link to the online survey and a unique ticket number for respondents to use. On each week after the launch of the survey, a postal mailing will go to the entire sample. This letter will also include the link to the online survey and a unique ticket number for the respondent to use. Once a respondent links to the survey and provides an e-mail address for future communication they will only be contacted via this e-mail address. A second (optional) postal mailing will be made one week later to non-respondents. All non-respondents will receive an e-mail reminder every 4-days for approximately two weeks for a maximum of 4-times. All communication will provide a deadline of the end of the 5th week for respondents to complete and submit the survey. The link will remain live for one week following the last reminder in order to maximize response.

In Summary:

E-mail 1: Announcement e-mail informing sample the survey was active and ready for completion.
Postal Mailing 1: 1- week after launch Reminder/tickler sent
Postal Mailing 2: 2- weeks after launch Reminder/tickler sent
E-mail 2 – 6: Every 4-day Reminder/ticklers sent

Mailing	**Week 1**	**Week 2**	**Week 3**							**Week 4**							**Week 5**						
E-mail 1:	x																						
Postal Mail 1:		x																					
Postal Mail 2:						x																	
E-mail 2:						x																	
E-mail 3:										x													
E-mail 4:														x									
E-mail 5:																		x					
E-mail 6:																						x	

b) Where will the survey be administered (e.g., in a dentist office, an auditorium as part of a group administration, or the respondent's home or office)?

The survey can be completed at any location with Web access. The introductory and follow-up letters will authorize the ability to complete the HPQ at duty station, and note that completing it at home or elsewhere is also permissible.

2. Collecting completed surveys

a) How many times will a potential respondent be contacted (e.g., a single time or a notification letter with a copy of the survey sent later)?

8 times (including combination of email and mail reminders) – see above.

b) Can the survey returns be monitored (e.g., by lithocodes on the survey) to determine who has or has not completed a survey?

The contract will use standard survey research methods in administering the questionnaire, and will assign each potential respondent with a unique identifier. Each SM will be assigned a "ticket number" that will allow tracking who has and has not completed a survey.

c) If a person does not return a survey, will the individual be contacted again? If so, who will be re-contacted (e.g., everyone in the sample or a subgroup tracked by numbered survey forms)?

Two weeks after receipt of the introductory letter, a follow-up e-mail will be sent to subgroups who did not already answer the survey tracked by numbered survey forms. See H1a above

G. Access to Surveys and Results

1. Who will have access to the completed surveys and the resulting electronic database(s)?

   Only authorized staff, as outlined in our SAQ and proposal, will have access to the completed surveys and the electronic database. The Contractor will create and clean the de-identified electronic database of survey responses.

2. How will the hard copy of the survey (e.g., paper surveys, diskettes used in computer-administered surveys, and notes from telephone surveys) be controlled, and eventually destroyed, to protect confidentiality?

   Documentation will be maintained in same secure facility as subject computer system as noted in SAQ.

3. What steps will be taken to prevent identification of data related to a specific respondent or non-respondent?

   The DoD will create the random sample of active-duty soldiers, according to the study criteria, and create a "ticket number", or study ID, which is not personally identifying information (e.g., SSN). DoD will provide a file with name, address, e-mail address, ticket number, and demographic (age, race/ethnicity, gender) and stratification (branch and officer/non-officer status) to DataStat in order to field the survey. Once the survey response period closes, DataStat will create a personnel file stripped of identifiers (no name, address, e-mail address, nor ticket number) but adds a variable indicating whether a survey was returned. This will allow for analysis of the demographics of responders and non-responders. A separate file containing the survey data, stratification variables, and ticket number but no personal identifiers - will also be created. These 2 files will be sent to JSI for analysis of non-respondents and of the survey results. They are not linkable.

   DataStat will also create a file of ticket numbers (no survey data) of respondents for DoD so that DoD can then extract claims data for respondents. DoD will then send to JSI the de-identified claims data, per the DUA, that is only linkable to the survey data by ticket number. DataStat will destroy all personally identifying files, according to an agreed upon timeframe.

4. What is the smallest size (*n*) that will be used for the presentation of results? Will the answers of a given person ever be associated with that person? For example, will John Doe's answers to a dental questionnaire

be shared with the dentist in such a manner that the dentist can associate the answers with John Doe with some degree of confidence?

The survey data file will never be linked to the personnel file so it will not be possible to directly associate responses with the person who completed the HPQ.

Care will be taken during data analysis to not over-stratify the data, particularly by demographic and geographic variables, to ensure the respondent cannot be indirectly associated with survey results. Any cell in a table with less than 5 respondents it will be represented with a dash. Further, when describing results of cross-tabulations of data, statistical inferences will not be made if one of the cells contains less than 5 respondents. This is the statistical "rule of thumb" that chi-square tests of association are not valid when one or more cells has less than 5 elements.

H. Type of Data in Electronic Dataset(s) and Access. Has the sponsoring agency defined the survey by type of data to be maintained in electronic dataset(s) and accepted all the conditions that apply? Definitions in this discussion are for the purposes of the discussion and are not the official definitions of the terms which are defined by law or regulation.

1. Are the survey data totally anonymous from the receipt of data?

   Yes. A non-confidential, de-identified data set will be maintained, in which the contents of the de-identified data set will be reviewed to determine whether beneficiaries could be identified from demographic and geographic information in the data set.

   a) To be considered totally anonymous no personal identifiers can remain attached to the survey once it has been received from the respondent. Further, demographic data that becomes part of the dataset cannot be so specific as to provide a possibility a respondent being identified. That is, a dataset user should not reasonably be able to simultaneously cross answers to multiple demographic questions (e.g., rank, gender, race/ethnicity, and geographic location) to identify unique individuals.
   b) If the survey data is totally anonymous from the receipt of data, then it follows that:
      (1) There are no issues with a requirement for a Systems of Record Notice. [DMDC Survey and Program Evaluation Division, however, does *not* make this determination.]
      (2) The only dataset that is created is a nonconfidential dataset. For purposes of this discussion, we define such a dataset as one where there are no personal identifiers and where a dataset user should not reasonably be able to

simultaneously cross answers to multiple demographic questions (e.g., rank, gender, race/ethnicity, and geographic location) to identify unique individuals. The Privacy Act is the basic source for what cannot be contained in a nonconfidential dataset.

2. Are the survey data *not* totally anonymous from the receipt of data?

   Survey data are totally anonymous.

   a) Personal identifiers and/or detailed demographics may be used by the data collector in the preparation and/or analysis of the data to prepare a report on the results *provided* that the dataset is part of a Systems of Record that has been announced in a Notice in the Federal Register.
   b) Contractors can collect and possess personal identifiers and/or detailed demographics from a Federal survey only while performing and for the purpose of fulfilling a contractual obligation to the Federal Government. Contractors' possession of the information is by extension of the Federal agencies Systems of Records. After Government acceptance of the report(s) requiring the use of the data, the contractor should only keep a copy of the nonconfidential survey dataset. With the expiration of a contract, the contractor is no longer covered by the Systems notice and has no legal authority to have a copy of confidential survey datasets.

3. If the survey data are not totally anonymous from the receipt of data, is it clear whether the agency will possess/maintain a copy of the confidential dataset?

   N/A

   a) An agency *may* possess and maintain a copy of the confidential dataset only as long as this is covered by a Systems of Record Notice and is consistent with the Privacy Act disclosure on the survey. For example, if the Privacy Act Notice said that personal identifiers would only be used for data collection, then the agency (and any contractor involved) would have to remove personal identifiers and very detailed demographics upon completion of the data collection phase which can include nonresponse adjustments and weighting, but not analysis.
   b) An agency need not possess at any time nor maintain a confidential dataset created for data collection and/or initial analysis. Keeping a copy of the confidential dataset is warranted as long as there is a reasonable expectation that the data

has continuing utility and that the utility warrants the risk of disclosure.

c)

3. Is it clear whether the agency will possess/maintain a copy of the nonconfidential dataset?

Yes, we will maintain both.

a) An agency should maintain a nonconfidential dataset even if they are maintaining a confidential dataset.

b) Only the nonconfidential dataset should be released to the public—even under almost all FOIA requests. [Note that an individual can obtain their own complete survey record.]

c) Typically, only the nonconfidential dataset should be released to contractors and other government offices, even other DoD offices.

4. If a contractor is gathering the data, is the contractor aware that all data except that in the nonconfidential database must be surrendered to the sponsoring agency at the completion of the contract or delivery order? That is, the contractor cannot keep a copy of the full database.

Yes.

5. Does the survey package submitted to DMDC acknowledge that DoDI 1100.13 can require that DMDC receive a copy of the full database at the completion of the project for surveys of continuing interest to the Office of the Under Secretary of Defense (Personnel and Readiness)?

Yes.

I. Privacy Act Statement

1. Does the survey's Privacy Act statement (PAS) adequately reflect the disclosure levels and risk described in the supporting statements for the survey? Is it in general conformity with typical statement used? [DMDC is not the official reviewer for PAS, but DMDC's reviews have alerted the Privacy Office with potential problems. See DoD Directive 5400.11 and 5400.11-R on the "DoD Privacy Program."]

Yes.

Title 10 of the United States Code, Sections 136 and 2358 are the statutes governing this survey. Your participation is voluntary and your answers will be kept private and your name confidential. The answers you give will not affect your benefits in any way.

If the submitter believes a PAS is not required, a letter from the sponsoring agency's Privacy Act officer should be attached. DoD Privacy Act Office will still examine the survey package to determine if they concur with ruling from the sponsoring agency.

Not applicable.

DMDC does not check whether the appropriate authority is cited. The sponsoring agency's Information Management Control Officer should provide the authority.

The authority will be checked by WHS/DIOR.

u:\sped\guides\dodi1100\chk_lst5.doc

Thursday, April 20, 2006

CITI Course Completion Record
for Jim Maxwell

To whom it may concern:

On *4/20/2006*, *Jim Maxwell* (username=JimMaxwell; Employee Number=) completed all *CITI Program* requirements for the *Basic CITI* Course in The Protection of Human Research Subjects.

Learner Institution: *Office of the Under Secretary of Defense (Personnel and Readiness)*

Learner Group: *Data Users*

Learner Group Description: *Data Users are researchers or research assistants who do not interact with human subjects but who have access to private or sensitive information about human subjects. Check with your institution to ensure this training program fulfills the institutional requirements for your research role.*

Contact Information:

Gender: Male
Department:
Which course do you plan to take?: Data Users
Role in human subjects research: Principal Investigator
Mailing Address:
Email: maxwell@jsi.com
Office Phone: 617-482-9485
Home Phone:

The Required Modules for *Data Users* are:	**Date completed**
Introduction	04/19/06
History and Ethical Principles - SBR	04/19/06
Defining Research with Human Subjects - SBR	04/19/06
Assessing Risk in Social and Behavioral Sciences - SBR	04/19/06
Privacy and Confidentiality - SBR	04/20/06
Internet Research - SBR	04/20/06

Conflicts of Interest in Research Involving Human Subjects	04/20/06
Office of the Under Secretary of Defense (Personnel and Readiness)	04/20/06

Additional optional modules completed:	**Date completed**

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI participating institution. Falsified information and unauthorized use of the CITI course site is unethical, and may be considered scientific misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Course Coordinator

SUMMARY OF POTENTIAL HUMAN SUBJECTS RESEARCH

P&R Organization: TMA TRICARE Management Activities/Health Affairs

Research Title: "Making a Case for Quality: The Business Case for TRICARE Health Promotion Programs"

Other ID/Serial: Contract Number: HHSP23320045016XI

Month/Year of Start: September 1, 2004

Lead Research Organization: John Snow, Inc.

Lead Investigator: Jim Maxwell, Ph.D.

Funded by: DOD

Sponsoring Organization: Health Program Analysis and Evaluation

Performing Organization(s): John Snow, Inc.

Contracting Organization: DataStat, Inc.

Other Participating Organization(s) and Role(s): Ron Kessler, Ph.D., Harvard Medical School, Department of Health Care Policy, Senior Consultant

Number of Human Subjects (enrolled and/or anticipated): 10,000 (estimated sample size)

Location(s) of Human Subject Interactions: Web or mail

Location(s) of Identifiable Personal Data: DataStat, Inc. will manage and maintain any identifiable personal information in a secure database that will not be linked with the survey responses. This database will be used for survey administration and response tracking purposes only to determine recipients of reminder notices and surveys.

Research Objectives: To conduct a survey of active duty services members to establish baseline measures of health, health needs, and access to health care services, with additional data on productivity and absenteeism by using the Health and Productivity Questionnaire (HPQ) instrument.

Key Research Procedures: We will use eligibility files to identify potential survey respondents. 1. DataStat will have Identifiable personal data. 2. DataStat will contact potential survey respondents via the identifiable personal data by email to invite them to participate on the web. 3. DataStat will collect information from the web, and provide JSI with data files in which there is no Identifiable personal data. An expedited review will be pursued if deemed necessary.

IRB Reviews (Y/N): N __ **Data Monitoring Board Reviews (Y/N):** N__

Informed Consent Required (Y/N): N __ **More Than Minimal Risk (Y/N):** N__

Involve Military Personnel (Y/N): Y__ **Involve Other "Vulnerable Subjects" (Y/N):** N__

Remarks:

Anonymous Data Assurance

JSI/DataStat Project – Making a Case for Quality: The Business Case for TriCare Health Promotions Programs

I. Survey Solicitation

Solicitation emails and reminders will be signed by the Base Commander (or other ranking official). The first e-mail will announce the survey to the total sample. This initial e-mail will include a link to the online survey for respondents. The survey can be completed at any location with Web access (including duty stations). JSI will never have access to or see the email database.

II. Access

Only authorized John Snow, Inc. (JSI) staff will have access to the electronic survey database. DataStat will create and clean the anonymous electronic database of survey responses and provide them to JSI.

III. Storage

The survey data will be stored in a file that does not contain the e-mail address, nor the IP address, or any other identifying information. The data will be stored on a secure, stand-alone computer system in a locked office in which access is strictly limited to the project director, the senior statistician, and a data analyst. The survey data file will never be linked to the personnel file so it will not be possible to directly associate responses with the person who completed the HPQ.

IV. Participant Identification

No information is collected on the survey that would allow individuals to be identified in any way. Care will be taken during data analysis to not over-stratify the data, particularly by demographic and geographic variables, to ensure the respondent cannot be indirectly associated with survey results. Any cell in a table with less than 5 respondents per base will be represented with a dash. When describing results of cross-tabulations of data, statistical inferences will not be made if one of the cells contains less than 5 respondents. This is the statistical "rule of thumb" that chi-square tests of association are not valid when one or more cells have less than 5 elements. In addition, the anticipated sample sizes are large enough that there will be numerous, varied individuals within certain specificities (e.g., rank, sex, etc.)

NOTIFICATION

You have been selected to participate in the *Health and Work Performance Questionnaire*. The survey will soon be available at the Web site listed below. The survey will ask about your health behaviors and injuries as well as your treated and untreated health problems, and their effects on your work performance. This information will assist us in the development of policies and programs that will improve the health and reduce the accidents and injuries of all Military Service members.

I urge you to share your perspective on these important issues. This is your opportunity to directly impact the formulation of military health and safety policies. These surveys are *Official Business* and can be completed at your duty station using government equipment. If you choose, you can also complete the survey at home or elsewhere. I assure you that your responses will be kept anonymous. No information identifying you as an individual will be collected.

Please take the time today to access the survey by going to our Web site:

https://www.datastat.com/hs/ss.dss?pid=CFS

If you cannot access the Web site or you have questions pertaining to the survey, there are three ways to contact us: by calling 1-800, e-mailing, or sending a facsimile to

Thank you for your time and assistance in this very important effort.

Sincerely,

Base Commander/Other Ranking Official

REMINDER 1

The Department of Defense is seeking your help through your participation in the *Health and Work Performance Questionnaire*. If you have already completed the survey, I want to thank you for taking the time to do so. If you have not yet completed it, you can take the survey at

https://www.datastat.com/healthsurvey/ss.dss?pid=cfs

The survey should take 30 minutes or less to complete. Please keep in mind, these surveys are *Official Business* and that they can be completed at your duty station using government equipment. You can also complete the survey at home or elsewhere.

Your participation in this survey effort is important. You were scientifically selected, as part of a small group of people, to participate in this survey. Therefore, your answers will represent the views of many others like yourself. This is your opportunity to assist us in the development of policies and programs that will improve the health and reduce the accidents and injuries of all Military Service members. Be assured your responses are anonymous. Only group statistics will be compiled and reported. No information identifying you individually will be collected.

If you cannot access the Web or experience any other problem with the survey, please e-mail or leave a message anytime, toll-free, at 1-800.

Your time and cooperation in this very important effort are greatly appreciated.

Sincerely,

Base Commander

REMINDER 2

Recently, you were asked to participate in the *Health and Work Performance Questionnaire*. If you have already completed the survey, I want to thank you for taking the time to do so.

If you have not had a chance to complete the anonymous survey or were thinking about not participating, I would like to ask you to reconsider. Your participation is crucial and this really is a chance to assist in the development of policies that will improve the health of all Military Service members.

If you have not done so already, please take the survey by logging onto the following Web site:

https://www.datastat.com/healthsurvey/ss.dss?pid=cfs

If you cannot take the survey now, please ensure you take the survey soon. The survey should take 30 minutes or less to complete. The survey will only be available on the Web site for a few more weeks, so please make every effort to complete the survey soon. Since these surveys are *Official Business*, you can use the computer equipment at your duty station to complete the survey. If you cannot access the Web or experience any other problem with the survey, please e-mail or leave a message anytime, toll-free, at 1-800.

Your time and cooperation in this very important effort are greatly appreciated.

Sincerely,

Base Commander

REMINDER 3

If you have already completed the *Health and Work Performance Questionnaire,* I want to thank you for taking the time to do so. If you have not completed the survey, I want to ask you one last time to do so.

I cannot emphasize enough how important surveys are to the Department. Surveys are the only systematic way we have for finding out how you and your fellow Service members feel about important topics such as health and work performance. I want to assure you that the time you take to complete surveys is not wasted—this really is your chance to be heard. My staff and I are personally briefed on the results of these anonymous surveys. We use this information to shape policies and programs that directly affect you.

If you have not done so already, please take the survey by logging onto the following Web site:

https://www.datastat.com/healthsurvey/ss.dss?pid=cfs

Please ensure you take the survey very soon. The survey will only be available on the Web site until. Since these surveys are *Official Business*, you can use the computer equipment at your duty station to complete the survey.

If you cannot access the Web or experience any other problem with the survey, please e-mail or leave a message anytime, toll-free, at 1-800.

Your time and cooperation in this important effort are greatly appreciated.

Sincerely,

Base Commander

HARVARD MEDICAL SCHOOL
DEPARTMENT OF HEALTH CARE POLICY

[COMPANY]
HEALTH AND WORK SURVEY

Welcome to the [Company] Health and Work Survey. The survey takes approximately 10 minutes to complete. We appreciate your participation.

Background: [Company] is collaborating with researchers from the HPQ Data Consortium™ in a study on the health and quality of life of working people. This survey includes questions about your physical and mental health and about the work you do.

Confidentiality: Your responses to this survey will be kept confidential and only researchers at the HPQ Data Consortium will see them. Your identity will remain anonymous and your answers will only be reported in the aggregate. Information on individuals will not be shared with [Company]. To further protect your privacy, we have obtained a Certificate of Confidentiality from the National Institutes of Health. This helps avoid disclosure of your information without your express consent. However, if there were a risk of serious harm to yourself or others, we would attempt to get appropriate help. This on-line survey is a **secure website** run by the HPQ Data Consortium, a research organization located in Ann Arbor, Michigan.

The barcode at the bottom right of this page is an identification number used by DataStat to monitor survey responses so as not to send reminders to people who have already responded.

If you want to know more about this survey, you may contact the HPQ Data Consortium, toll free at 1-800-837-6322.

The Health and Work Survey was developed by the World Health Organization (WHO) as part of the WHO Composite International Diagnostic Interview (copyright © 2001 by WHO) and is used here with the permission of the World Health Organization

A. Your Health

Survey Instructions

Please be sure to fill the response circle COMPLETELY.
Use only BLACK or BLUE INK or DARK PENCIL
to complete the survey.

Correct

● Yes
○ No

Incorrect

☒ Yes ○ No
✓ Yes ○ No
◐ Yes ○ No

	Excellent	Very Good	Good	Fair	Poor
A1. In general, how would you rate your overall health now?	○	○	○	○	○
A2. In general, how would you rate your overall mental health now?	○	○	○	○	○

A3. Do you have any of the following conditions? If your answer is YES, mark whether you never, previously, or currently receive professional treatment. (Professional treatment is any treatment supervised by a health professional.) If you are unsure if you have a condition, please mark the NO response option.

	NO, I don't have this condition	YES, but never received professional treatment	YES, previously received (but don't currently receive) professional treatment	YES, and I currently receive professional treatment
A3a. Arthritis or rheumatism?	○	○	○	○
A3b. Chronic back/neck pain?	○	○	○	○
A3c. Migraine headaches?	○	○	○	○
A3d. Other frequent or severe headaches?	○	○	○	○
A3e. Any other chronic pain?	○	○	○	○
A3f. High blood pressure or hypertension?	○	○	○	○
A3g. Congestive heart failure?	○	○	○	○
A3h. Coronary heart disease?	○	○	○	○
A3i. High blood cholesterol?	○	○	○	○

A4. **Do you have any of the following conditions? If your answer is YES, mark whether you never, previously, or currently receive professional treatment. (Professional treatment is any treatment supervised by a health professional.) If you are unsure if you have a condition, please mark the NO response option.**

	NO, I don't have this condition	YES, but never received professional treatment	YES, previously received (but don't currently receive) professional treatment	YES, and I currently receive professional treatment
A4a. An ulcer in your stomach or intestine?	○	○	○	○
A4b. Irritable bowel disorder?	○	○	○	○
A4c. Chronic heartburn or GERD?	○	○	○	○
A4d. Seasonal allergies or hay fever?	○	○	○	○
A4e. Asthma?	○	○	○	○
A4f. Chronic bronchitis or emphysema?	○	○	○	○
A4g. Chronic Obstructive Pulmonary Disease?	○	○	○	○
A4h. Urinary or bladder problems?	○	○	○	○
A4i. Diabetes?	○	○	○	○
A4j. Obesity?	○	○	○	○
A4k. Chronic sleeping problems?	○	○	○	○
A4l. Chronic fatigue or low energy?	○	○	○	○
A4m. Osteoporosis?	○	○	○	○
A4n. Skin cancer?	○	○	○	○
A4o. Any other kind of cancer?	○	○	○	○
A4p. Anxiety disorder?	○	○	○	○
A4q. Depression?	○	○	○	○
A4r. Any other emotional problem?	○	○	○	○
A4s. Substance problems (drugs or alcohol)?	○	○	○	○

A5. During the past 4 weeks (28 days), how much were you bothered by each of the following conditions?

	Not at all	A little	Some	A lot
A5a. Feeling tired or having low energy?	○	○	○	○
A5b. Trouble sleeping?	○	○	○	○
A5c. Headaches?	○	○	○	○
A5d. Back or neck pain?	○	○	○	○
A5e. Pain in your arms, legs, or joints (knees, hips, etc.)?	○	○	○	○
A5f. Muscle soreness?	○	○	○	○
A5g. Water eyes, runny nose, or stuffy head?	○	○	○	○
A5h. Cough or sore throat?	○	○	○	○
A5i. Fever, chills, or other cold/flu symptoms?	○	○	○	○
A5j. Constipation, loose bowels, or diarrhea?	○	○	○	○
A5k. Nausea, gas, or indigestion?	○	○	○	○

A6. During the past 4 weeks (28 days), how much of the time did you feel...

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
A6a. ...so sad nothing could cheer you up?	○	○	○	○	○
A6b. ...nervous?	○	○	○	○	○
A6c. ...restless or fidgety?	○	○	○	○	○
A6d. ...hopeless?	○	○	○	○	○
A6e. ...that everything was an effort?	○	○	○	○	○
A6f. ...worthless?	○	○	○	○	○

Survey Instructions

To record a response of "0 times," follow this example	→	Number of times (000-365)	0 0 0
To record a response of "8 times," follow this example	→	Number of times (000-365)	0 0 8
To record a response of "26 times," follow this example	→	Number of times (000-365)	0 2 6

A7.(Women Only) During the past 4 weeks (28 days), were you pregnant?

- ○ Yes
- ○ No
- ○ Not Sure
- ○ I am male

A8. In the past 12 months, did you have a work related accident, injury, or poisoning that required medical attention?

- ○ Yes
- ○ No ➔ ***Go to Question A9***

A8a. How many days of work did you miss in the past 12 months because of a work related accident, injury, or poisoning? (If less than 1 day, enter 000.)

☐ ☐ ☐ Number of days (000-365)

A9. How many times did you see each of the following types of professionals in the past 12 months? Include only visits regarding your own health, not visits when you took someone else to be examined.

Example: If you visited a dentist 2 times in the past year and an optician once, your answer to **A9c** would be 3.

	Number of times (000-365)
A9a. A doctor, hospital, or clinic for a routine physical check-up or gynecological exam (not counting pregnancy related care)	☐☐☐
A9b. (Women Only) A doctor, hospital, or clinic for pregnancy related care (If male, enter 000.)	☐☐☐
A9c. A dentist or optician for a routine check-up or exam	☐☐☐
A9d. A doctor, emergency room, or clinic for urgent care treatment (for example, because of new symptoms, an accident, or something else unexpected)	☐☐☐
A9e. A doctor, hospital, clinic, orthodontist, or ophthalmologist for scheduled treatment or surgery	☐☐☐
A9f. A psychiatrist, psychologist, or other mental health professional	☐☐☐

A10. How many nights did you stay in a hospital during the past 12 months (not including nights associated with childbirth)?

☐☐☐ Number of nights (000-365)

A10a. (Women Only) How many nights did you stay in a hospital during the past 12 months for nights associated with childbirth? (If male, enter 000.)

☐☐☐ Number of nights (000-365)

B. Your Work

B1. Please choose the category that best describes your main job. If none of the categories fits you exactly, please respond with the closest category to your experience. (Select only one.)

- ○ Executive, administrator, or senior manager
  (e.g., CEO, sales VP, plant manager)
- ○ Professional
  (e.g., engineer, accountant, systems analyst)
- ○ Technical support
  (e.g., lab technician, legal assistant, computer programmer)
- ○ Sales
  (e.g., sales representative, stockbroker, retail sales)
- ○ Clerical and administrative support
  (e.g., secretary, billing clerk, office supervisor)
- ○ Service occupation
  (e.g., security officer, food service worker, janitor)
- ○ Precision production and crafts worker
  (e.g., mechanic, carpenter, machinist)
- ○ Operator or laborer
  (e.g., assembly line worker, truck driver, construction worker)

B2. How many people do you personally supervise on your job?
(If more than 97, enter 97.)

☐☐ Number of people (00-97)

B3. About how many hours altogether did you work in the past 7 days?
(If more than 97, enter 97.)

☐☐ Number of hours (00-97)

B4. How many hours does your employer expect you to work in a typical 7-day week?
(If it varies, estimate the average. If more than 97, enter 97.)

☐☐ Number of hours (00-97)

B5. Now please think of your work experiences over the past 4 weeks (28 days). In the spaces provided below, write the number of days you spent in each of the following work situations.

In the past 4 weeks (28 days), how many days did you...

		Number of days (00-28)
B5a.	...miss an **entire** work day because of problems with your physical or mental health? (Please include only days missed for your **own** health, not someone else's health.)	☐☐
B5b.	...miss an **entire** work day for any other reason (including vacation)?	☐☐
B5c.	...miss **part** of a work day because of problems with your physical or mental health? (Please include only days missed for your **own** health, not someone else's health.)	☐☐
B5d.	...miss **part** of a work day for any other reason (including vacation)?	☐☐

B6. About how many hours altogether did you work in the past 4 weeks (28 days)? (See examples below.)

☐☐☐ Number of hours in the past 4 weeks (28 days)

Examples for Calculating Hours Worked in the Past 4 Weeks

40 hours per week for 4 weeks = 160 hours
35 hours per week for 4 weeks = 140 hours
40 hours per week for 4 weeks with 2 8-hour days missed = 144 hours
40 hours per week for 4 weeks with 3 4-hour partial days missed = 148 hours
35 hours per week for 4 weeks with 2 8-hour days missed and 3 4-hour partial days missed = 112 hours

B7. Did you have any of the following experiences at work in the past 4 weeks (28 days)?

	Yes	No
B7a. Any special work success or achievement?	○	○
B7b. Any special work failure?	○	○
B7c. An accident that caused either damage, work delay, a near miss, or a safety risk?	○	○

B7d. If you answered "Yes" to any of the questions B7a, B7b, or B7c, please describe what happened.

_____

_____

_____

_____

B8. The next questions are about the time you spent during your hours at work in the past 4 weeks (28 days). Select the one response for each question that comes closest to your experience.

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
B8a. How often was your performance higher than most workers on your job?	○	○	○	○	○
B8b. How often was your performance lower than most workers on your job?	○	○	○	○	○
B8c. How often did you do no work at times when you were supposed to be working?	○	○	○	○	○
B8d. How often did you find yourself not working as carefully as you should?	○	○	○	○	○
B8e. How often was the quality of your work lower than it should have been?	○	○	○	○	○
B8f. How often did you not concentrate enough on your work?	○	○	○	○	○
B8g. How often did health problems limit the kind or amount of work you could do?	○	○	○	○	○

B9. On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B10. Using the same 0-to-10 scale, how would you rate your usual job performance over the past year or two?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B11. Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked during the past 4 weeks (28 days)?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B12. How would you compare your overall job performance on the days you worked during the past 4 weeks (28 days) with the performance of most other workers who have a similar type of job? (Select only one.)

- ○ You were **a lot better** than other workers
- ○ You were **somewhat better** than other workers
- ○ You were **a little better** than other workers

- ○ You were about **average**

- ○ You were **a little worse** than other workers
- ○ You were **somewhat worse** than other workers
- ○ You were **a lot worse** than other workers

C. DEMOGRAPHICS

C1. How old are you?

☐ ☐ YEARS OLD (00-99)

C2. Are you male or female?

- ○ Male
- ○ Female

C3. What is your current marital status?

- ○ Married or Cohabiting
- ○ Separated
- ○ Divorced
- ○ Widowed
- ○ Never Married

C4. How many children do you have?

- ○ None
- ○ One
- ○ Two
- ○ Three
- ○ Four or more

C5. What is the highest grade or level of school that you have <u>completed</u>?

- ○ 8th grade or less
- ○ Some high school, but did not graduate
- ○ High school graduate or GED
- ○ Some college or 2-year degree
- ○ 4-year college graduate
- ○ More than 4-year college degree

C6. What is your height?

☐ Feet (0-9) ☐ ☐ Inches (00-11) (Please round to the nearest inch)

C7. How much do you weigh?

☐ ☐ ☐ Pounds (Please round to the nearest pound 000-999)

C8. What is your annual income from your job, before taxes?

- ○ $1 - $999
- ○ $1,000 - $1,999
- ○ $2,000 - $2,999
- ○ $3,000 - $3,999
- ○ $4,000 - $4,999
- ○ $5,000 - $5,999
- ○ $6,000 - $6,999
- ○ $7,000 - $7,999
- ○ $8,000 - $8,999
- ○ $9,000 - $9,999
- ○ $10,000 - $10,999
- ○ $11,000 - $11,999
- ○ $12,000 - $12,999
- ○ $13,000 - $13,999
- ○ $14,000 - $14,999
- ○ $15,000 - $15,999
- ○ $16,000 - $16,999
- ○ $17,000 - $17,999
- ○ $18,000 - $18,999
- ○ $19,000 - $19,999
- ○ $20,000 - $24,999
- ○ $25,000 - $29,999
- ○ $30,000 - $34,999
- ○ $35,000 - $39,999
- ○ $40,000 - $44,999
- ○ $45,000 - $49,999
- ○ $50,000 - $74,999
- ○ $75,000 - $99,999
- ○ $100,000 - $149,999
- ○ $150,000 - $199,999
- ○ $200,000 - $299,999
- ○ $300,000 - $499,999
- ○ $500,000 - $999,999
- ○ $1,000,000 or more

Thank You! You have just completed the [Company] Health and Work Survey. We appreciate your interest and participation in the study, and we thank you for your time and patience.

Please return the completed survey in the postage-paid envelope to:

DataStat Inc.
3975 Research Park Drive
Ann Arbor, MI 48108

[Georgia Healthcare Leadership Council/Midwest Business Group on Health]
HEALTH AT WORK SURVEY

Welcome to the [Georgia Healthcare Leadership Council/Midwest Business Group on Health] Health at Work Survey. The survey takes approximately 15 minutes to complete. It would be a great help to us if you took the time to participate.

Background: Your company is collaborating with the [Georgia Healthcare Leadership Council/Midwest Business Group on Health] and a consortium of researchers from Harvard Medical School, the Institute for Health and Productivity Management, and a number of affiliated health research organizations (CorSolutions, DataStat, Galt Associates) in a survey about the health and quality of life of working people. The study is funded by the federal government's Centers for Disease Control (CDC). The survey includes questions about your physical and mental health and about the work you do. We urge you to participate, as the survey results will help your employer design health benefit packages that better meet your needs.

Confidentiality: Your participation in the survey is strictly voluntary and your responses will be strictly confidential. Your employer will never know whether or not you participated. Only researchers in the consortium will have access to the anonymous survey responses. Survey results will be reported to employers only in the aggregate. If there were a risk of serious harm to yourself or others, we would attempt to get appropriate help. A **secure website** is used to administer the survey on-line.

Survey instructions: As you answer the questions, please do not use your browser's BACK and FORWARD buttons. Please only use the BACK and NEXT buttons below each question to move backward and forward through the survey. If you cannot finish all the questions in one sitting, remember to click on the FINISH LATER button and you may finish answering the rest of the questions later. In the unusual event that you receive an error message, the server may be temporarily busy. Please wait 60 seconds and press the REFRESH key to continue. If you encounter any technical difficulties with the site, you may contact DataStat's toll free number at 1-800-<NUMBER>. Please do not contact your employer's technical support department.

If you want to know more about the survey, you may call our survey firm, DataStat, toll free at 1-800-<NUMBER>

If you are ready to complete the questionnaire right now, please click the NEXT button below. If you want to complete the survey at a more convenient time or not to participate at all, click the EXIT button. Thanks very much for your help.

NEXT EXIT

The Health and Work Survey was developed by the World Health Organization (WHO) as part of the WHO Composite International Diagnostic Interview (copyright © 2001 by WHO) and is used here with the permission of the World Health Organization

A. Your Health

Survey Instructions

Please be sure to fill the response circle COMPLETELY.
Use only BLACK or BLUE INK or DARK PENCIL
to complete the survey.

Correct

● Yes
○ No

Incorrect

⊗ Yes ○ No ✓ Yes ○ No ◐ Yes ○ No

	Excellent	Very Good	Good	Fair	Poor
A1. In general, how would you rate your overall health now?	○	○	○	○	○
A2. In general, how would you rate your overall mental health now?	○	○	○	○	○

A3. Do you have any of the following conditions? If your answer is YES, mark whether you never, previously, or currently receive professional treatment. (Professional treatment is any treatment supervised by a health professional.) If you are unsure if you have a condition, please mark the NO response option.

	NO, I don't have this condition	YES, but I never received professional treatment	YES, I previously received (but don't currently receive) professional treatment	YES, and I currently receive professional treatment
A3a. Arthritis?	○	○	○	○
A3b. Chronic back/neck pain?	○	○	○	○
A3c. Migraine headaches?	○	○	○	○
A3d. Other frequent or severe headaches?	○	○	○	○
A3e. Any other chronic pain?	○	○	○	○
A3f. High blood pressure or hypertension?	○	○	○	○
A3g. Congestive heart failure?	○	○	○	○
A3h. Coronary heart disease?	○	○	○	○
A3i. High blood cholesterol?	○	○	○	○

CHECKPOINT: If R HAS ARTHRITIS (A3a = ANY OF THE THREE "YES" RESPONSES), GO TO A4. OTHERWISE, GO TO A7.

A4. You mentioned having arthritis. Most people with arthritis have osteoarthritis, which is caused by the cartilage in joints wearing down until bones rub against each other and cause pain. When a doctor tells you that you have "arthritis," he means osteoarthritis unless he explicitly says otherwise. The other kind of arthritis is rheumatoid arthritis. This is a relatively rare auto-immune disease that causes inflammation of the tissues that line joints. Rheumatoid arthritis usually begins in early adulthood. With these definitions in mind, which of the two do you have: osteoarthritis or rheumatoid arthritis?

- ○ Osteoarthritis
- ○ Rheumatoid Arthritis

A5. Think about the pain you felt due to your arthritis during the past 4 weeks (28 days). How much pain did you have…

	None	Mild	Moderate	Severe	Extreme
A5a. …when walking on a flat surface?	○	○	○	○	○
A5b. …when going up or down stairs?	○	○	○	○	○
A5c. …at night while in bed (i.e., pain that disturbs your sleep)?	○	○	○	○	○
A5d. …when sitting or lying?	○	○	○	○	○
A5e. …when standing upright?	○	○	○	○	○

A6. Think about the difficulty you had in doing the following daily physical activities due to arthritis during the past 4 weeks (28 days). (By "difficulty," we mean your ability to move around and to look after yourself.) How much difficulty did you have…

	None	Mild	Moderate	Severe	Extreme
A6a. …when walking up stairs?	○	○	○	○	○
A6b. …while standing?	○	○	○	○	○
A6c. …when walking on a flat surface?	○	○	○	○	○
A6d. …when putting on your socks or panty hose or stockings?	○	○	○	○	○
A6e. …while sitting?	○	○	○	○	○

A7. **Do you have any of the following conditions? If your answer is YES, mark whether you never, previously, or currently receive professional treatment. (Professional treatment is any treatment supervised by a health professional.) If you are unsure if you have a condition, please mark the NO response option.**

	NO, I don't have this condition	YES, but never received professional treatment	YES, previously received (but don't currently receive) professional treatment	YES, and I currently receive professional treatment
A7a. An ulcer in your stomach or intestine?	○	○	○	○
A7b. Either frequent diarrhea or frequent constipation?	○	○	○	○
A7c. Frequent nausea, gas, or indigestion?	○	○	○	○
A7d. Chronic heartburn or GERD?	○	○	○	○
A7e. Seasonal allergies or hay fever?	○	○	○	○
A7f. Asthma?	○	○	○	○
A7g. Chronic bronchitis or emphysema?	○	○	○	○
A7h. Chronic Obstructive Pulmonary Disease?	○	○	○	○
A7i. Urinary or bladder problems?	○	○	○	○
A7j. Diabetes?	○	○	○	○
A7k. Chronic sleeping problems?	○	○	○	○
A7l. Chronic fatigue or low energy?	○	○	○	○
A7m. Osteoporosis?	○	○	○	○
A7n. Multiple Sclerosis?	○	○	○	○
A7o. Skin cancer?	○	○	○	○
A7p. Any other kind of cancer?	○	○	○	○
A7q. Anxiety disorder?	○	○	○	○
A7r. Depression?	○	○	○	○
A7s. Any other emotional problem?	○	○	○	○
A7t. Substance problems (drugs or alcohol)?	○	○	○	○

CHECKPOINT: If R HAS DIABETES (A7J = ANY OF THE THREE "YES" RESPONSES), GO TO A8. OTHERWISE, GO TO A14.

A8. You mentioned having diabetes. How old were you when you were first diagnosed with diabetes?

☐☐ Age (00-99)

A9. Which of the following things do you do to control your diabetes:

	Yes	No
A9a. Do you exercise?	○	○
A9b. Do you have a special diet?	○	○
A9c. Do you take insulin?	○	○
A9d. Do you take prescription pills or medications	○	○

A10. How often do you check your urine?

- ○ More than once a day
- ○ Once a day
- ○ Most days
- ○ Several days a week
- ○ One or two days a week
- ○ Several days a month
- ○ One or two days a month
- ○ Less than once a month
- ○ Never

A11. How often do you check your blood?

- ○ More than once a day
- ○ Once a day
- ○ Most days
- ○ Several days a week
- ○ One or two days a week
- ○ Several days a month
- ○ One or two days a month
- ○ Less than once a month
- ○ Never

A12. Have you ever had any of the following problems because of your diabetes?

	Yes	No
A12a. Circulation problems in your legs?	○	○
A12b. Foot ulcers?	○	○
A12c. Vision problems?	○	○
A12d. An amputated limb?	○	○

A13. How many times in the past 12 months did you see each of the following kinds of doctors because of your diabetes?

	Number of Times (00-99)
A13a. A foot doctor?	☐☐
A13b. An eye doctor?	☐☐
A13c. An endocrinologist?	☐☐
A13d. Any other medical doctor?	☐☐

A14. (Women Only) Are you currently pregnant?

- ○ Yes
- ○ No
- ○ Not Sure
- ○ I am male

A15. Do you smoke cigarettes?

- ○ Currently
- ○ Ex-smoker → **GO TO A16**
- ○ Only smoked a few times → **GO TO A16**
- ○ Never → **GO TO A16**

A15a. How many cigarettes do you smoke a day?

- ○ 10 or less
- ○ 11 – 20
- ○ 21 – 30
- ○ 31 or more

A15b. How soon after you wake do you smoke your first cigarette?

- ○ Within 5 minutes
- ○ 6-30 minutes
- ○ 31-60 minutes
- ○ After 60 minutes

A15c. Which cigarette would you hate most to give up?

- ○ The first one in the morning
- ○ All others

A15d. Do you find it difficult to refrain from smoking in places where it is forbidden, such as the library, theater, or doctor's office?

- ○ Yes
- ○ No

A15e. Do you smoke more frequently during the first hours after waking than the rest of the day?

- ○ Yes
- ○ No

A15f. Do you smoke when you are so ill that you are in bed most of the day?

- ○ Yes
- ○ No

A16. How often do you usually have at least one drink of alcohol?

- ○ Nearly everyday
- ○ Several days per week
- ○ 1-2 days a week
- ○ 1-3 days a month
- ○ Less than once a month
- ○ Never → **GO TO A17**

A16a. On the days you drink, about how many drinks do you usually have per day?

- ○ 1-2 drinks
- ○ 3-4 drinks
- ○ 5-10 drinks
- ○ 10+ drinks

A16b. How often do you drink <u>5 or more</u> drinks per day?

- ○ Nearly everyday
- ○ Several days a week
- ○ 1-2 days a week
- ○ 1-3 days a month
- ○ Less than once a month
- ○ Never

A17. Some people have times lasting several days or longer when they feel much more excited, full of energy, or hyper than usual. Their minds go fast. They sometimes talk a lot more than usual or they are very restless or they are unable to sit still. They sometimes do things that are unusual for them, such as driving too fast or spending too much money. Have you ever had times like this in your life?

- ○ Yes
- ○ No → **GO TO A21**

A18. During the times you felt much more excited, full or energy, or hyper than usual, how often did you…

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
A18a. … feel so good or so hyper that people thought you were not your normal self?	○	○	○	○	○
A18b. …feel so hyper that you got into trouble?	○	○	○	○	○
A18c. … sleep much less than usual without missing it?	○	○	○	○	○
A18d. …have thoughts race through your head so much that you couldn't slow your mind down?	○	○	○	○	○
A18e. … have much more energy than usual?	○	○	○	○	○
A18f. … have problems because of these experiences – like getting into arguments, being unable to work, having family problems, having money problems, or having legal problems?	○	○	○	○	○

CHECKPOINT: If R RESPONDED "NONE OF THE TIME" TO A18f, GO TO A21. OTHERWISE, GO TO A19.

A19. How serious were these problems at their worst – would you say minor, moderate, serious, or very serious?

○ Minor
○ Moderate
○ Serious
○ Very Serious

A20. About how many weeks out of 52 in the past year did you have an episode of feeling much more excited, full of energy, or hyper than usual with some of the other problems that we just reviewed? You can use any number between 0 and 52 to answer.

☐☐ Number of weeks (00-52)

A21. The next questions are about problems you may have with attention or concentration.

	Never	Rarely	Sometimes	Often	Very Often
A21a. How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?	○	○	○	○	○
A21b. How often do you have difficulty getting things in order when you have to do a task that requires organization?	○	○	○	○	○
A21c. How often do you have problems remembering appointments or obligations?	○	○	○	○	○
A21d. When you have a task that requires a lot of thought, how often do you avoid or delay getting started?	○	○	○	○	○
A21e. How often do you fidget or squirm with your hands or feet when you have to sit down for a long time?	○	○	○	○	○
A21f. How often do you feel overly active and compelled to do things, like you were driven by a motor?	○	○	○	○	○

A22. The next questions are about how often you got tired over the past twelve months. How often did you become very tired, weak, or exhausted while performing each of the following kinds of activities?

	Never	Rarely	Sometimes	Often	Very Often
A22a. ...minor everyday <u>physical tasks</u> like working, shopping, housekeeping, and walking?	○	○	○	○	○
A22b. …minor everyday <u>mental tasks</u> like reading, writing, and doing paperwork?	○	○	○	○	○

<u>CHECKPOINT:</u> IF R checked "SOMETIMES," "OFTEN," or "VERY OFTEN" TO ONE OR BOTH OF A22a or A22b, GO TO A23. OTHERWISE, GO TO A24.

A23. During the times you became very tired while performing minor everyday tasks, what would happen when you tried to rest or relax? Would you…

- ○ …fully regain your energy and strength? → **GO TO A24**
- ○ …still feel tired or weak?

A23a. When this problem was most severe over the past 12 months, how often did you get tired?

- ○ Nearly everyday
- ○ Several days a week
- ○ 1-2 days a week
- ○ 1-3 days a month
- ○ Less than once a month

A23b. How often were you too tired to carry out your daily activities?

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Often
- ○ Very often

A24. During the past twelve months, have you had at least one week each month when you had frequent pain or discomfort in your stomach or lower abdomen that was relieved when you had a bowel movement?

- ○ No
- ○ Yes, but I never received any professional treatment
- ○ Yes, I previously received (but don't currently receive) professional treatment
- ○ Yes, and I currently receive professional treatment

A25. **The next few questions are about problems with your sleep. During the past twelve months, how many weeks did you have problems …**

	0 Weeks	1-2 Weeks	3-4 Weeks	5-8 Weeks	9-12 Weeks	13-26 Weeks	27-51 Weeks	52 Weeks
A25a. …getting to sleep, when nearly every night it took you two hours or longer before you could fall asleep?	○	○	○	○	○	○	○	○
A25b. …staying asleep, when you woke up nearly every night and took an hour or more to get back to sleep?	○	○	○	○	○	○	○	○
A25c. …waking too early, when you woke up nearly every morning at least two hours earlier than you wanted to?	○	○	○	○	○	○	○	○
A25d. …feeling sleepy during the day?	○	○	○	○	○	○	○	○

A26. **Have you ever had surgery on the stomach or esophagus (food pipe)?**

○ Yes
○ No

A26a. **Have you had a burning pain or discomfort behind the breast bone in your chest (HEARTBURN) in the last year? (Please do not count pain in your stomach or pain from heart trouble.)**

○ Yes
○ No → **GO TO 27a**

A26b. **How many times have you had heartburn in the last year?**

○ Daily
○ Several times a week
○ About once a week
○ About once a month
○ Less than once a month

A26c. How bad is your heartburn usually?

- ○ Mild – can be ignored if I don't think about it
- ○ Moderate – cannot be ignored, but does not affect my life-style
- ○ Severe – affects my life-style
- ○ Very severe – markedly affects my life-style

A27a. Have you had a bitter or sour tasting fluid coming up into your throat or mouth (ACID REGURGITATION) in the last year?

- ○ Yes
- ○ No → **GO TO A28**

A27b. How many times have you had acid regurgitation in the last year?

- ○ Daily
- ○ Several times a week
- ○ About once a week
- ○ About once a month
- ○ Less than once a month

A27c. How bad is your acid regurgitation usually?

- ○ Mild – can be ignored if I don't think about it
- ○ Moderate – cannot be ignored, but does not affect my life-style
- ○ Severe – affects my life-style
- ○ Very severe – markedly affects my life-style

A28. In the past 6 months, how often have you taken each of the following types of medication for heartburn or acid regurgitation?

	Daily	Weekly	Monthly	Never
A28a. Over-the-counter Antacids (For example: Gaviscon, Rolaids, Mylanta, Maalox, Tums, Gelusil, Amphogel, Di-Gel, Titralac etc.)	○	○	○	○
A28b. Over-the-counter acid blockers (For example: Pepcid AC, Pepcid Complete, Prilosec OTC, Tagamet HB 200, Axid AR, Zantac 75, Zantac 150)	○	○	○	○
A28c. Any of the following prescription medications: Zantac/Zantac Efferdose (ranitidine), Tagamet (cimetidine), Axid (nizatidine), or Pepcid/Pepcid RPD (famotidine)	○	○	○	○
A28d. Any of the following prescription medications: Nexium (esomeprazole magnesium), Protonix (pantaprozole sodium), Aciphex (rabeprazole sodium), Prevacid (lansoprazole), Prilosec (omeprazole), Zegerid (omeprazole), or generic omeprazole	○	○	○	○
A28e. The following prescription medication: Reglan (metoclopramide), Urecholine (bethanechol)	○	○	○	○

A29. About how many times in the past twelve months did you have an attack of anger when all of a sudden you lost control and broke or smashed something worth more than a few dollars?

☐☐☐ Number of Times (000-999)

A30. About how many times in the past twelve months did you have an attack of anger when all of a sudden you lost control and threatened, hit, or hurt someone?

☐☐☐ Number of Times (000-999)

A31. In the past 12 months, how many accidents, injuries, or poisonings did you have that required medical attention?

☐☐☐ Number of accidents (000-999)

CHECKPOINT: IF NO ACCIDENTS IN A31, GO TO A32. OTHERWISE, GO TO A31a

A31a. About how many days of work did you miss in the past 12 months because of a work related accident, injury, or poisoning? (If less than 1 day, enter 000.)

☐☐☐ Number of days (000-365)

Survey Instructions

To record a response of "0 times," follow this example	→	Number of times (000-365)	0 0 0
To record a response of "8 times," follow this example	→	Number of times (000-365)	0 0 8
To record a response of "26 times," follow this example	→	Number of times (000-365)	0 2 6

A31b. Which of the conditions on this list resulted from your most recent accident, injury, or poisoning? Please check all that apply.

- ○ Broken or dislocated bones
- ○ Sprain, strain, or pulled muscle
- ○ Cuts, scrapes, or puncture wounds
- ○ Head injury, concussion
- ○ Bruise, contusion, or internal bleeding
- ○ Burn, scald
- ○ Poisoning from chemicals, medicines, or drugs
- ○ Other, please describe: _____

_____

A31c. What caused that most recent accident, injury, or poisoning? Briefly describe what you were doing and what happened. (For example, fell down while playing basketball and sprained ankle.)

_____

_____

_____

A31d. In what month did the most recent accident, injury, or poisoning occur?

_____ (month)

A32. In the past 12 months, how many work-related accidents did you have that either damaged company property, led to a work delay, or otherwise had a financial cost to your company?

☐☐☐ Number of accidents (000-999)

CHECKPOINT: IF NO ACCIDENTS IN A32, GO TO A33. OTHERWISE, GO TO A32a

A32a. What is your best estimate of the financial loss to your company caused by your accident(s) over the past 12 months?

$_____ (dollar amount)

A33. During the past 4 weeks (28 days), how much were you bothered by each of the following conditions?

	Not at all	A little	Some	A lot
A33a. Feeling dizzy	○	○	○	○
A33b. Feeling tired or having low energy	○	○	○	○
A33c. Trouble sleeping	○	○	○	○
A33d. Headaches	○	○	○	○
A33e. Back or neck pain	○	○	○	○
A33f. Pain in your arms, legs, or joints (knees, hips, etc.)	○	○	○	○
A33g. Muscle soreness	○	○	○	○
A33h. Watery eyes, runny nose, or stuffy head	○	○	○	○
A33i. Cough or sore throat	○	○	○	○
A33j. Fever, chills, or other cold/flu symptoms	○	○	○	○
A33k. Constipation, loose bowels, or diarrhea	○	○	○	○
A33l. Nausea, gas, or indigestion	○	○	○	○

A34. During the past 4 weeks (28 days), how much of the time did you feel...

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
A34a. ...so sad nothing could cheer you up?	○	○	○	○	○
A34b. ...nervous?	○	○	○	○	○
A34c. ...restless or fidgety?	○	○	○	○	○
A34d. ...hopeless?	○	○	○	○	○
A34e. ...that everything was an effort?	○	○	○	○	○
A34f. ...worthless?	○	○	○	○	○
A34g. …unable to relax?	○	○	○	○	○
A34h. …impatient or irritable?	○	○	○	○	○

A35. How many times did you see each of the following types of professionals in the past 12 months? Include only visits regarding your own health, not visits when you took someone else to be examined.

Example: If you visited a dentist 2 times in the past year and an optician once, your answer to **A29c** would be 003.

	Number of times (000-365)
A35a. A doctor, hospital, or clinic for a routine physical check-up or gynecological exam (not counting pregnancy related care)	☐☐☐
A35b. (Women Only) A doctor, hospital, or clinic for pregnancy related care (If male, enter 000.)	☐☐☐
A35c. A dentist or optician for a routine check-up or exam	☐☐☐
A35d. A doctor, emergency room, or clinic for urgent care treatment (for example, because of new symptoms, an accident, or something else unexpected)	☐☐☐
A35e. A doctor, hospital, clinic, orthodontist, or ophthalmologist for scheduled treatment or surgery	☐☐☐
A35f. A psychiatrist, psychologist, or other mental health professional	☐☐☐

A36. How many nights did you stay in a hospital during the past 12 months (not including nights associated with childbirth)?

☐ ☐ ☐ Number of nights (000-365)

A36a. (Women Only) How many nights did you stay in a hospital during the past 12 months for nights associated with childbirth? (If male, enter 000.)

☐ ☐ ☐ Number of nights (000-365)

NOTE: THE FOLLOWING QUESTION WILL BE DELETED FOR SOME COMPANIES AND THE OPTIONS MODIFIED FOR OTHERS.

A37. What is the name of your health plan(s)? (*Please refer to your Health Plan Card. Check all that apply.)*

- ○ Local List #1
- ○ Local List #2
- ○ Local List #3
- ○ Local List #4
- ○ Local List #5
- ○ Local List #6
- ○ Local List #7
- ○ Local List #8
- ○ Local List #9
- ○ Some other plan *(Please print the name of the plan)*

_____

_____

A38. It would greatly help our research if we could merge your questionnaire responses with your health plan's computerized records of your medical and pharmacy claims. This would allow us to examine the comparative effects of different treated health problems on employees. All records will be kept strictly confidential and will be used only by the project researchers from Harvard Medical School and CorSolutions to carry out aggregated analyses. Can we have your permission to use your claims records in this way?

- ○ Yes
- ○ No ➔ **GO TO B1**

A38a. We will need your health plan identification number to link your questionnaire responses to your medical and pharmacy claims records.

— — — — — — — — — — —

B. YOUR WORK

B1. Please choose the category that best describes your main job. If none of the categories fits you exactly, please respond with the closest category to your experience. (Select only one.)

- ○ Executive, administrator, or senior manager
  (e.g., CEO, sales VP, plant manager)
- ○ Professional
  (e.g., engineer, accountant, systems analyst)
- ○ Technical support
  (e.g., lab technician, legal assistant, computer programmer)
- ○ Sales
  (e.g., sales representative, stockbroker, retail sales)
- ○ Clerical and administrative support
  (e.g., secretary, billing clerk, office supervisor)
- ○ Service occupation
  (e.g., security officer, food service worker, janitor)
- ○ Precision production and crafts worker
  (e.g., mechanic, carpenter, machinist)
- ○ Chemical/Production Operator
  (e.g., shift supervisors and hourly employees)
- ○ Laborer
  (e.g., truck driver, construction worker)

B2. Is your work schedule best described as a regular schedule (roughly the same hours every day), a rotating schedule (e.g., working a day shift some days and a night shift other days), or an irregular schedule (e.g., unpredictable hours controlled by situations or workload)?

- ○ Regular schedule → **GO TO B4**
- ○ Rotating schedule
- ○ Irregular schedule

B3. **What percent of your total work hours in an average week are in each of the following times of day? (The sum should add up to 100%)**

	%
Morning (6:00AM-12:00PM)	___
Afternoon (12:00PM-6:00PM)	___
Evening (6:00PM-12:00AM)	___
Nights (12:00AM-6:00AM)	___
Total	**100**

CHECKPOINT: GO TO B6 IF RESPONDENT ANSWERED B3

B4. **What time do you usually begin work?**
*****create dropdown that begins at midnight**

B5. **What time do you usually end work?**
*****create drop down that begins at midnight**

B6. **How many people do you personally supervise on your job?**
(If more than 97, enter 97.)

☐☐ Number of people (00-97)

B7. **About how many hours altogether did you work in the past 7 days?**
(If more than 97, enter 97.)

☐☐ Number of hours (00-97)

B8. **How many hours does your employer expect you to work in a typical 7-day week?**
(If it varies, estimate the average. If more than 97, enter 97.)

☐☐ Number of hours (00-97)

B9. **Now please think of your work experiences over the past 4 weeks (28 days). In the spaces provided below, write the number of days you spent in each of the following work situations.**

In the past 4 weeks (28 days), how many days did you...

	Number of days (00-28)
B9a. **...miss an entire work day because of problems with your physical or mental health? (Please include only days missed for your own health, not someone else's health.)**	☐☐
B9b. **...miss an entire work day for any other reason (including vacation)?**	☐☐
B9c. **...miss part of a work day because of problems with your physical or mental health? (Please include only days missed for your own health, not someone else's health.)**	☐☐
B9d. **...miss part of a work day for any other reason (including vacation)?**	☐☐
B9e. **...come in early, go home late, or work on your day off?**	☐☐

CHECKPOINT: IF R HAS NOT MISSED AN ENTIRE DAY OR A PARTIAL DAY (R ANSWERED "00" FOR ALL QUESTIONS IN B9 SERIES) GO TO B11. OTHERWISE GO TO B10.

B10. **Think of all the days in the past four weeks (28 days) when you missed either a full day of work or a partial day of work. Count partial days as whole days.**

How many of these days did you …

	Number of days (00-28)
B10a. **…not receive pay?**	☐☐
B10b. **…get paid as part of regular salary**	☐☐
B10c. **…use earned sick leave (while receiving regular pay)**	☐☐
B10d. **…use earned vacation time (while receiving regular pay)**	☐☐
B10e. **…get paid as short-term or long-term disability**	☐☐
B10f. **…get paid as a result of an injury at work**	☐☐
TOTAL (The sum of the numbers in B10a-B10f should equal the total number of days you missed work in the past 28 days.)	☐☐

B11. About how many hours altogether did you work in the past 4 weeks (28 days)? (See examples below.)

☐ ☐ ☐ Number of hours in the past 4 weeks (28 days)

Examples for Calculating Hours Worked in the Past 4 Weeks

40 hours per week for 4 weeks = 160 hours
35 hours per week for 4 weeks = 140 hours
40 hours per week for 4 weeks with 2 8-hour days missed = 144 hours
40 hours per week for 4 weeks with 3 4-hour partial days missed = 148 hours
35 hours per week for 4 weeks with 2 8-hour days missed and 3 4-hour partial days missed = 112 hours

B11a. In the past 4 weeks (28 days), did you have any special work success or achievement?

○ Yes
○ No ➔ **GO TO B12a**

B11b. Please describe your special work success or achievement.

_____

_____

B12a. In the past 4 weeks (28 days), did you have any special work failure?

○ Yes
○ No ➔ **GO TO B13**

B12b. Please describe your special work failure.

_____

_____

B13. The next questions are about the time you spent during your hours at work in the past 4 weeks (28 days). Select the one response for each question that comes closest to your experience.

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
B13a. How often was your performance higher than most workers on your job?	○	○	○	○	○
B13b. How often was your performance lower than most workers on your job?	○	○	○	○	○
B13c. How often did you do no work at times when you were supposed to be working?	○	○	○	○	○
B13d. How often did you find yourself not working as carefully as you should?	○	○	○	○	○
B13e. How often was the quality of your work lower than it should have been?	○	○	○	○	○
B13f. How often did you not concentrate enough on your work?	○	○	○	○	○
B13g. How often did health problems limit the kind or amount of work you could do?	○	○	○	○	○

B14. On a scale from 0 to 10 where 0 is the worst job performance anyone could have at your job and 10 is the performance of a top worker, how would you rate the usual performance of most workers in a job similar to yours?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B15. Using the same 0-to-10 scale, how would you rate your usual job performance over the past year or two?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B16. Using the same 0-to-10 scale, how would you rate your overall job performance on the days you worked during the past 4 weeks (28 days)?

Worst Performance										Top Performance
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

B17. How would you compare your overall job performance on the days you worked during the past 4 weeks (28 days) with the performance of most other workers who have a similar type of job? (Select only one.)

- ○ You were **a lot better** than other workers
- ○ You were **somewhat better** than other workers
- ○ You were **a little better** than other workers

- ○ You were about **average**

- ○ You were **a little worse** than other workers
- ○ You were **somewhat worse** than other workers
- ○ You were **a lot worse** than other workers

C. Demographics

C1. How old are you?

☐☐ YEARS OLD (00-99)

C2. Are you male or female?

- ○ Male
- ○ Female

C3. What is your current marital status?

- ○ Married or Cohabiting
- ○ Separated
- ○ Divorced
- ○ Widowed
- ○ Never Married

C4. How many children do you have?

- ○ None
- ○ One
- ○ Two
- ○ Three
- ○ Four or more

C5. What is the highest grade or level of school that you have completed?

- ○ 8th grade or less
- ○ Some high school, but did not graduate
- ○ High school graduate or GED
- ○ Some college or 2-year degree
- ○ 4-year college graduate
- ○ More than 4-year college degree

C6. What is your height?

☐☐ FEET (0-9) ☐☐ Inches (00-11) (Please round to the nearest inch)

C7. How much do you weigh?

☐☐☐ Pounds (Please round to the pound 000-999)

C8. What is your annual income from your job, before taxes?

- ○ $1 - $999
- ○ $1,000 - $1,999
- ○ $2,000 - $2,999
- ○ $3,000 - $3,999
- ○ $4,000 - $4,999
- ○ $5,000 - $5,999
- ○ $6,000 - $6,999
- ○ $7,000 - $7,999
- ○ $8,000 - $8,999
- ○ $9,000 - $9,999
- ○ $10,000 - $10,999
- ○ $11,000 - $11,999
- ○ $12,000 - $12,999
- ○ $13,000 - $13,999
- ○ $14,000 - $14,999
- ○ $15,000 - $15,999
- ○ $16,000 - $16,999
- ○ $17,000 - $17,999
- ○ $18,000 - $18,999
- ○ $19,000 - $19,999
- ○ $20,000 - $24,999
- ○ $25,000 - $29,999
- ○ $30,000 - $34,999
- ○ $35,000 - $39,999
- ○ $40,000 - $44,999
- ○ $45,000 - $49,999
- ○ $50,000 - $74,999
- ○ $75,000 - $99,999
- ○ $100,000 - $149,999
- ○ $150,000 - $199,999
- ○ $200,000 - $299,999
- ○ $300,000 - $499,999
- ○ $500,000 - $999,999
- ○ $1,000,000 or more

C9. We plan to repeat this survey annually to monitor change in the health of respondents. We need you to generate a private identification code that can be used to link survey responses over years. For that purpose, please tell me the following:

C9a. The first four letters of your mother's maiden name:

☐ ☐ ☐ ☐

C9b. The last four number of your social security number:

☐ ☐ ☐ ☐

That completes the survey. Thanks very much for your participation.

Please return the completed survey in the postage-paid envelope to:

DataStat Inc.
3975 Research Park Drive
Ann Arbor, MI 48108

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

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

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

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

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

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

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

Copyright © by American College of Occupational and Environmental Medicine

DOI: 10.1097/01.jom.0000052967.43131.51

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

Development of the HPQ

Background

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

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

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

Work Performance

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

work performance that met our needs.

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

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

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Absenteeism

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

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

Job-Related Accidents

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

The HPQ Calibration Survey

Samples

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 1
Sample Dispositions

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

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 2
Demographic Distributions of the Samples

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

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

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

The Archival Work Performance Measures

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

work performance measures actually used by employers.

The Payroll Record Measures of Absenteeism

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

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

Analysis Methods: Work Performance

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

Analysis methods: absenteeism

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

Results

The Distribution of the Global HPQ Work Performance Ratings

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 4
Distributions of the HPQ Global Work Performance Ratings

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

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

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

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

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

The Associations of HPQ Ratings with Archival and ESM Performance Measures

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

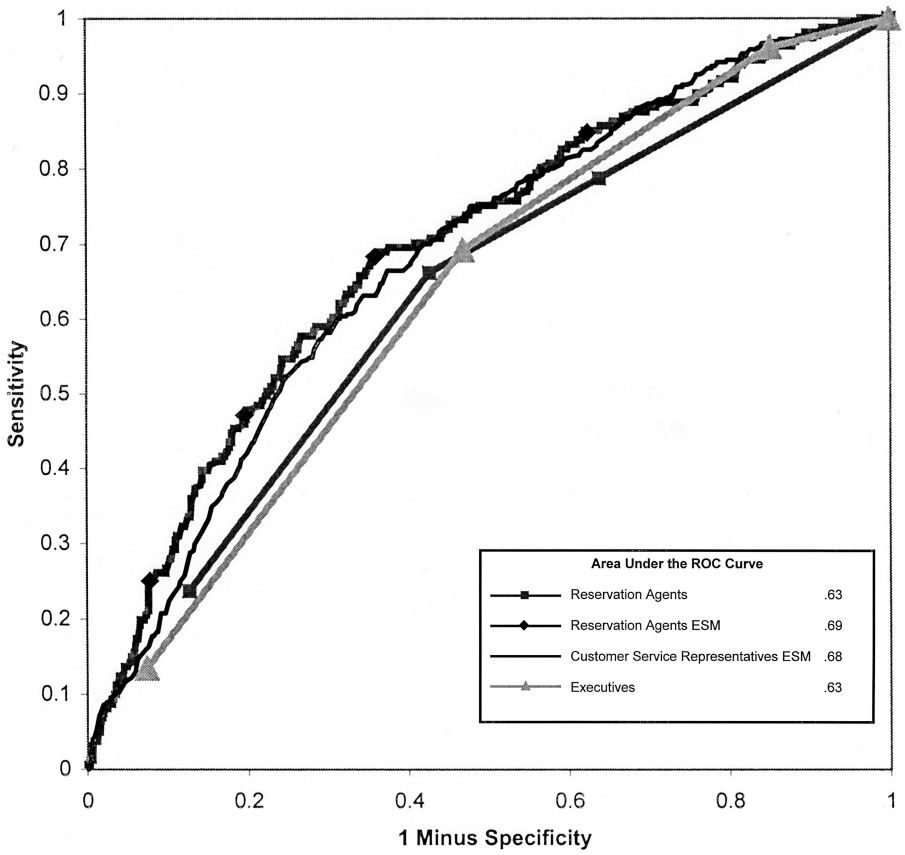


Fig. 1. Receiver operator characteristic curves for HPQ Global Ratings predicting archival and ESM measures of low work performance.

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

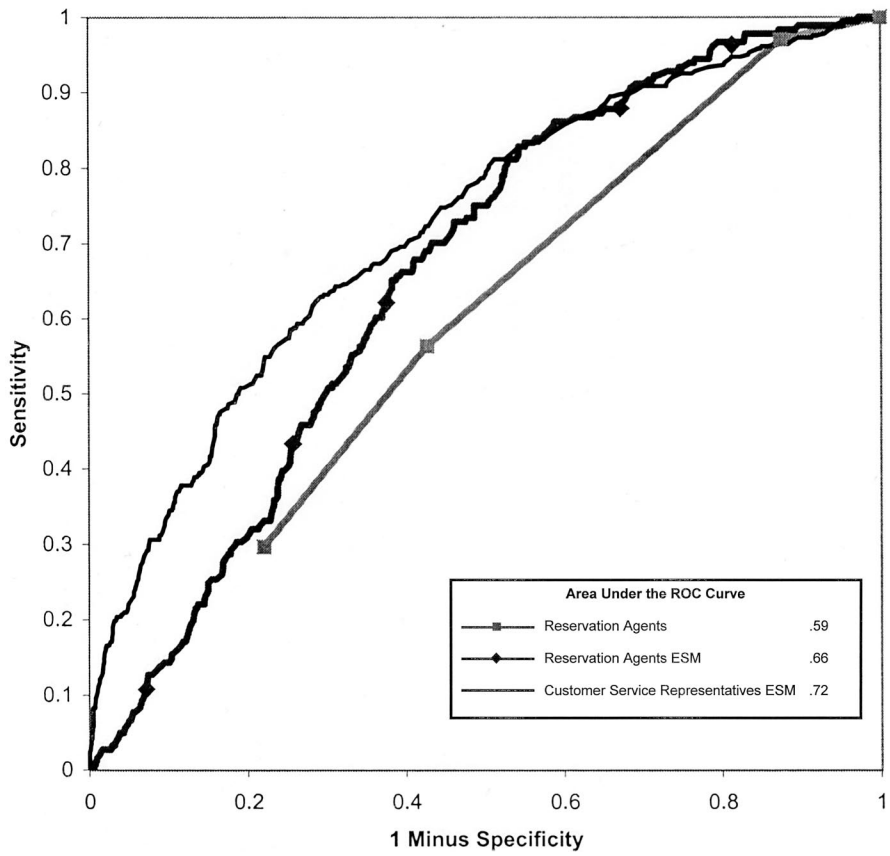


Fig. 2. Receiver operator characteristic curves for HPQ global ratings predicting archival and ESM measures of high work performance.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

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

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

The Associations of HPQ Absenteeism Reports with Payroll Records

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

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

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

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

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

Discussion

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

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

Sensitivity of the HPQ Measures

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

Calibrating HPQ Global Ratings

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

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

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

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

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

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

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

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

Monetizing Absenteeism and Work Performance Ratings

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Future Directions

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

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

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

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

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

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

Acknowledgments

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

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

References

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

Chronic Medical Conditions and Work Performance in the Health and Work Performance Questionnaire Calibration Surveys

Philip S. Wang, MD, DrPH
Arne Beck, PhD
Patricia Berglund, MBA
Joseph A. Leutzinger, PhD
Nico Pronk, PhD
Dennis Richling
Thomas W. Schenk, PhD, MPH
Gregory Simon, MD
Paul Stang, PhD
T. Bedirhan Ustun, MD
Ronald C. Kessler, PhD

Associations between chronic conditions and work performance (absenteeism, presenteeism, and critical incidents) were studied in reservation agents, customer service representatives, executives, and railroad engineers. Conditions and work performance were assessed with the World Health Organization's Health and Work Performance Questionnaire. Analysis of covariance was used to estimate associations. More work performance was lost from presenteeism than absenteeism. However, chronic conditions more consistently had negative impacts on absenteeism than presenteeism. Conditions with significant effects included arthritis, asthma, chronic obstructive pulmonary disease–emphysema, depression, and chronic headaches. Arthritis had the largest aggregate effect on absenteeism–presenteeism. Only depression affected both absenteeism–presenteeism and critical incidents. Some chronic conditions have substantial workplace effects. Disease management programs for these conditions might have a positive return on investment (ROI). Health and productivity tracking surveys are needed to evaluate ROI and provide quality assurance. (J Occup Environ Med. *2003;45:1303–1311*)

From Harvard Medical School, Boston, Massachusetts (Dr Wang); Kaiser-Permanente, Denver, Colorado (Dr Beck); University of Michigan, Ann Arbor, Michigan (Ms Berglund); Wellness Council of America, Omaha, Nebraska (Dr Leutzinger); HealthPartners, Minneapolis, Minnesota (Dr Pronk); Midwest Business Group on Health, Chicago, Illinois (Mr Richling); Oakland University, Detroit, Michigan (Dr Schenk); Group Health Cooperative of Puget Sound, Seattle, Washington (Dr Simon); Galt Associates, Inc., Blue Bell, Pennsylvania, and the University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Dr Stang); World Health Organization, Geneva, Switzerland (Dr Ustun); and Harvard Medical School, Boston, Massachusetts (Dr Kessler).

Address correspondence to: Philip Wang, MD, DrPH, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; E-mail address: wang@hcp.med.harvard.edu.

Copyright © by American College of Occupational and Environmental Medicine

DOI: 10.1097/01.jom.0000100200.90573.df

The ability of society to provide medical treatments for all its citizens for all their illnesses has become increasingly strained as a result of the proliferation of costly therapies for chronic conditions.[1] Triage rules have been developed to deal with this problem based on the principles of cost-effectiveness analysis and cost–benefit analysis.[2] Work impairment is often the largest and most easily monetized indirect societal cost of illness and, as such, plays an important role in these rules, even those developed from the broader societal perspective rather than from the employer perspective.[3] However, absenteeism has generally been the only aspect of work impairment included in these rules because absenteeism is more easily measured than other aspects of work performance.[4,5] This neglect of other workplace costs of illness has led to an underestimation of the indirect costs of illness and to a skewing of focus away from conditions associated with low absenteeism but high rates of impairment while at work. The latter is referred to in this report as presenteeism.

This report presents epidemiologic data designed to evaluate the effects of chronic conditions on work performance by considering both absenteeism and presenteeism. We surveyed workers in 4 occupations to calibrate the World Health Organization's Health and Work Performance Questionnaire (HPQ). The HPQ data

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

were used to examine the associations of common chronic conditions with work performance (absenteeism, presenteeism, and critical workplace incidents) reported to have occurred during the 4 weeks before the survey. This analysis focuses on 2 issues: the extent to which the total workplace costs of illness increase when we include a consideration of presenteeism rather than only absenteeism; and the extent to which the rank ordering of the most costly conditions from the employer perspective changes when we consider presenteeism rather than only absenteeism. Both issues are critical to establish priorities for research, health insurance coverage, and healthcare quality improvement.

These 2 issues are inherently comparative. Comparative analyses of this sort are important because employers have to make comparative decisions when they choose among a wide range of candidate health initiatives, from flu vaccination campaigns to substance dependence treatment programs. However, most research on the workplace costs of illness has focused on a single condition such as asthma,[6,7] cardiovascular disease,[8] depression,[9,10] diabetes,[11] or hypertension.[12] Few studies have examined the comparative effects of several different conditions on work impairment. In recognition of the importance of comparative analysis, the current report focuses on a range of illnesses, with a special focus on the chronic conditions that are most common in the U.S. labor force.

Methods

Samples

As described in more detail previously in this journal,[13] 4 occupations were included in the HPQ calibration surveys: reservation agents working for a major airline, customer service representatives working for a large telecommunications company, executives working for a major automobile manufacturer, and railroad engineers working for a large railroad company. An advance letter or e-mail was sent to each predesignated respondent by the medical director of the participating company to inform recipients that the company was collaborating with researchers from Harvard Medical School (HMS) in a survey of employee health and work performance. The letter made it clear that participation was completely voluntary and anonymous. The letter also provided an 800 number for recipients who had questions or who wanted to opt out of the survey. Professional telephone interviewers then made 20 attempts to contact each of those who did not initially opt out. Verbal informed consent was obtained before administering the survey. The survey was carried out by telephone. These recruitment, consent, and administration procedures were approved by the Human subjects Committee of Harvard Medical School.

The telephone lists provided by the participating companies had substantial proportions of invalid numbers, ranging from 15.0% to 37.6% across samples, making it impossible to determine the precise response rates of the surveys. However, the cooperation rate (the percent of completed interviews among people who were contacted) was in the range 64.0% to 85.6% across occupations, with interviews completed by 441 reservation agents, 505 customer service representatives, 554 executives, and 850 railroad engineers. Reservation agents and customer service representatives were largely women with a modal age of 30 to 44, whereas executives and railroad engineers were largely men with a modal age of 45 to 59. Railroad engineers had the lowest education (50.6% had no more than a high school education), whereas executives had the highest levels of education (98.7% were college graduates). More details on the sample design, administration procedures, and respondent characteristics are reported elsewhere.[13]

Measures

The interviewer version of the HPQ was administered to all respondents in the telephone surveys. Two series of HPQ questions are of particular interest for the current report: those that assessed the presence of chronic conditions and those that assessed work performance. Chronic physical conditions were assessed in a checklist based on the larger list used in the U.S. National Health Interview Survey (NHIS). The NHIS list was reduced by focusing on conditions that are consistently found in the NHIS to be commonly occurring among working people and have consistently been found in a number of national surveys to be significantly related to impaired work performance.[14,15] These conditions include arthritis, asthma, chronic back or neck pain, cancer, chronic obstructive pulmonary disease (COPD) or emphysema, diabetes, gastrointestinal ulcers, chronic headaches, heart disease, hypertension, and seasonal allergies. Methodologic research has shown that these conditions are self-reported with good validity when they either have been brought to medical attention or have caused significant activity limitation.[16–19] Illustrative levels of agreement between self-reports and medical records based on the κ statistic[20] are 0.41 for arthritis, 0.55 for asthma, 0.62 for heart disease, and 0.73 for hypertension.[16] These should be interpreted as lower-bound validity estimates because medical records are imperfect (eg, some chronic conditions such as arthritis and headaches often are not brought to medical attention). Consistent with this, there is lower agreement between self-reports and medical records for mild symptom-based conditions that have not been treated, presumably reflecting failure to detect or record these conditions on the part of healthcare professionals.[21,22] Two chronic recurrent mental conditions are also included in the HPQ: major depression (MD) and generalized anxiety

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

disorder (GAD). Both were assessed with the Composite International Diagnostic Interview Short-Form (CIDI-SF) screening scales, a set of brief scales with good validity in reproducing diagnostic assessments based on in-depth psychiatric evaluations.[23] The concordance of the CIDI-SF measures of MD and GAD with more in-depth diagnostic assessments based on the κ statistic are 0.86 for MD and 0.62 for GAD.

Work performance was assessed in the HPQ with questions about work impairments in the past 30 days resulting from physical or emotional health problems. Three performance domains are considered in the current report: absenteeism, presenteeism, and critical incidents. Absenteeism was assessed with a series of questions about both full days and partial days of missed work in relation to expected time at work. Because workers in 3 of the 4 occupations (all except executives) often work either part-time or in rotating shifts, absenteeism was assessed in terms of hours rather than days and was converted to a day-equivalent by dividing total absenteeism hours by 8. HPQ validation studies show good concordance between measures of self-reported absenteeism and payroll records over a 30-day recall period, with Pearson correlations in the range 0.66 to .79 and evidence of some downward bias in mean self-reported absenteeism levels.[13]

Presenteeism was assessed with a series of questions that required respondents to rate their performance while at work in both absolute and relative (ie, compared with other workers in the same occupation) measures of quality, quantity, and overall performance. The aggregate score across all reports was converted to metric with a theoretical range between 0 and 100 and treated as a ratio scale for purposes of calculating the performance of each respondent in relation to the perceived performance of other workers in the same occupation. The additive inverse of this ratio was used as a raw measure of relative presenteeism (ie, the percentage decrement in work performance of the respondent compared with the average worker in the same occupation) and then normed to have a mean of zero in the total sample of each occupation. Relative presenteeism was multiplied by hours actually worked to create a summary hour-equivalent measure of presenteeism that can be interpreted as the excess number of hours the respondent worked in comparison to the number of hours the average worker would have had to work to complete the same amount of overall work. Presenteeism hours were converted to day-equivalents by dividing by 8; absenteeism and presenteeism days were also added to calculate the number of absenteeism/presenteeism days combined. HPQ validation studies document fair to good concordance between measures of self-reported presenteeism and independent measures of work performance based on either supervisor or peer ratings. Concordance is much better at the tails than in the middle of the distribution.[13] Area under the receiver-operating characteristic (ROC) curve in predicting inadequate work performance, as defined by the lowest 20th percentile of the supervisor or peer rating distributions, is in the range 0.63 to .69. These numbers should be interpreted as lower-bound estimates of the validity of self-reported presenteeism because supervisor and peer ratings are imperfect measures of true work performance.

Critical incidents, finally, were assessed with 4 yes–no questions that asked respondents to report the presence of job-related accidents, injuries, special failures, and special successes in the past 30 days. Open-ended information was obtained about the specifics of each reported critical incident for purposes of coding and estimating financial costs or benefits to the employer. For purposes of the current report, we focus on a single summary measure that assigns 1 point for each negative critical incident and subtracts 1 point for each positive critical incident reported by each respondent. No archival information was available to validate the accuracy or completeness of critical incident reports. However, evaluation of the open-ended responses shows that the reported incidents are, in fact, noteworthy events that employers would either want to reduce or, in the case of successes, increase.

Analysis Procedures

Descriptive analyses were undertaken to calculate the prevalence of chronic conditions and the distribution of the work performance measures. The effects of individual chronic conditions on work performance were estimated using analysis of covariance pooled across the 4 occupations to maximize statistical power. Each condition was treated as a separate dummy predictor variable in a multivariate equation that included controls for age, sex, education, and occupation. Variation in effect sizes across occupations was evaluated with statistical tests of interaction between conditions and occupations. Finally, the prevalence of each condition and the independent effect of the condition on the work performance measures were combined to calculate the projected annual numbers of combined absenteeism–presenteeism days and negative critical incidents as a result of each statistically significant condition per 100 workers in each occupation. Statistical significance was evaluated using 0.05-level 2-sided tests.

Results

Prevalence of Chronic Conditions

The prevalence of chronic conditions in each of the 4 occupations is shown in Table 1. The most common conditions across all samples combined are, in order, seasonal allergies, chronic back–neck pain, chronic headaches, hypertension, arthritis, and depression. Prevalence differs signifi-

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 1
The Prevalences of Chronic Conditions in Four Occupations

	All Companies Pooled		Reservation Agents		Customer Service Representatives		Executives		Railroad Engineers			
	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	Percent	(SE)	$F_{3,2359}$*	*P* Value†
Arthritis/rheumatism	13.9	0.7	12.7	(1.6)	12.2	(1.4)	9.4	(1.2)	18.4	(1.3)	8.9	<0.001
Asthma	6.3	0.5	5.7	(1.1)	10.4	(1.3)	4.0	(0.8)	5.5	(0.8)	7.2	<0.001
Chronic back or neck pain	24.9	0.9	27.7	(2.1)	24.9	(1.9)	15.5	(1.5)	29.5	(1.6)	12.7	<0.001
Cancer	1.3	0.2	1.6	(0.6)	1.0	(0.4)	1.6	(0.5)	1.2	(0.4)	0.4	0.735
COPD/emphysema	1.7	0.3	3.0	(0.8)	1.4	(0.5)	0.4	(0.3)	2.0	(0.5)	3.7	0.011
Depression	9.5	0.6	12.0	(1.6)	14.5	(1.5)	3.9	(0.8)	8.8	(1.0)	12.9	<0.001
Diabetes	3.8	0.4	3.2	(0.8)	3.1	(0.8)	3.1	(0.7)	5.1	(0.8)	1.9	0.129
General anxiety disorder	2.5	0.3	3.4	(0.9)	2.7	(0.7)	1.3	(0.5)	2.8	(0.6)	1.8	0.151
GI ulcers	1.9	0.3	0.7	(0.4)	3.3	(0.8)	0.7	(0.4)	2.6	(0.5)	4.9	0.002
Chronic headaches	15.1	0.7	21.3	(2.0)	20.3	(1.8)	4.9	(0.9)	15.3	(1.2)	23.8	<0.001
Heart disease	2.9	0.3	3.2	(0.8)	1.4	(0.5)	3.4	(0.8)	3.4	(0.6)	1.9	0.120
Hypertension	14.9	0.7	11.1	(1.5)	9.5	(1.3)	18.6	(1.7)	15.5	(1.2)	7.8	<0.001
Seasonal allergies	40.7	1.8	49.0	(2.4)	41.3	(2.2)	31.4	(2.0)	42.0	(1.7)	11.2	<0.001
(no.)	(2363)		(441)		(518)		(554)		(850)			

* F-tests evaluated the variation in condition prevalences across the 4 occupations.
† *P* values from 2-sided tests.
SE, standard error; COPD, chronic obstructive pulmonary disease; GI, gastrointestinal.

TABLE 2
Mean Absenteeism Days and Net Negative Critical Incidents Per Month in Four Occupations

	Mean Absenteeism Days Per Month		Mean Net Negative Critical Incidents Per Month	
	Percent	(SE)	Percent	(SE)
Reservation agents	4.3	0.27	−0.5	0.3
Customer service representatives	1.9	0.16	−0.5	0.2
Executives	1.9	0.11	−0.7	0.2
Railroad engineers	4.9	0.19	−0.1	0.2
$F_{3,2359}$*	58.0		141.8	
P value†	<0.001		<0.001	

* F-tests evaluated the variation in work performance measures across the 4 occupations.
† *P* values from 2-sided tests.
SE, standard error.

cantly across occupations at the 0.05 level for 9 of the 13 conditions, the exceptions being cancer, diabetes, generalized anxiety disorder, and heart disease. Railroad engineers have the highest prevalence of several conditions (arthritis, chronic back–neck pain, diabetes, and allergies). Reservation agents have the highest prevalence of chronic headaches. Customer service representatives have the highest prevalence of asthma, depression, and gastrointestinal ulcers. Executives have the lowest prevalence of most conditions but the highest prevalence of hypertension. The higher prevalence of hypertension among executives and railroad engineers could reflect a difference in hypertension screening and diagnosis because executives and railroad engineers received routine physical examinations at their employers' expense, whereas the reservation agents and service representatives did not.

Work Performance

Table 2 presents means for the measures of absenteeism and critical incidents. (The means of presenteeism are, by definition, zero in each sample.) Both outcomes differ significantly across occupations ($F_{3,2359} = 58.0, 141.8, P < 0.001$). Mean absenteeism days in the past month range between 1.9 days (customer service representatives) and 4.9 days (railroad engineers). The means of critical incidents, in comparison, are all negative, indicating that workers report more special successes than failures, accidents, or injuries. The range of mean critical incidents is between -0.1 (railroad engineers) and -0.7 (executives).

Individual-Level Associations Between Chronic Conditions and Work Performance

The individual-level associations between conditions and work performance are presented in Table 3. The table entries are based on analysis of covariance coefficients that were transformed to represent annual excess number of days (for absenteeism, presenteeism, and their combination) or of critical incidents (the combination of excess negative incidents and fewer positive incidents) associated with each condition independent of the effects of other con-

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 3
Individual-Level Associations Between Chronic Conditions and Annual Excess Absenteeism Days, Presenteeism Days, and Net Negative Critical Incidents

	Annual Excess Absenteeism Days*		Annual Excess Presenteeism Days*†		Annual Excess Absenteeism and Presenteeism Days Combined*†		Annual Excess Net Negative Critical Incidents†	
	Days	(SE)	Days	(SE)	Days	(SE)	No.	(SE)
Arthritis/rheumatism	8.5	(3.4)	15.6	(5.8)	17.8	(6.2)	0.4	(0.4)
Asthma	10.6	(4.8)	18.3	(8.2)	19.4	(8.6)	−0.1	(0.5)
Chronic back or neck pain	3.7	(3.6)	−4.5	(4.8)	2.3	(5.0)	0.4	(0.3)
Cancer	2.4	(10.1)	12.3	(17.3)	20.5	(18.7)	1.0	(1.1)
COPD/emphysema	19.4	(8.9)	27.5	(15.6)	42.9	(17.0)	0.3	(1.0)
Depression	12.4	(4.1)	3.8	(6.7)	15.1	(7.4)	1.0	(0.4)
Diabetes	6.4	(6.0)	7.3	(10.3)	16.0	(11.0)	1.0	(0.7)
General anxiety disorder	11.1	(7.2)	−3.0	(12.7)	5.4	(13.7)	−0.4	(0.8)
GI ulcers	2.6	(8.4)	−7.8	(14.4)	−11.1	(15.4)	−0.3	(0.9)
Chronic headaches	13.3	(3.4)	−6.7	(5.8)	6.3	(6.2)	1.1	(0.4)
Heart disease	8.8	(7.0)	8.9	(11.8)	16.3	(12.7)	0.8	(0.8)
Hypertension	4.2	(3.4)	−7.7	(5.8)	1.6	(6.2)	−0.4	(0.4)
Seasonal allergies	1.2	(2.2)	−6.4	(4.1)	0.1	(4.3)	0.1	(0.3)
$F_{13,2340}$§	6.0		1.7		2.8		1.3	
P value‖	<0.001		0.05		0.00		0.22	
$F_{12,2336}$¶	1.2		2.3		1.8		18.0	
P value‖	0.26		0.00		0.5		<0.001	

* Effects of individual conditions on work performance measures were derived from analyses of covariance adjusted for age, sex, education, occupation, and other conditions.

† Annual excess presenteeism days were calculated by first multiplying a summary measure of relative presenteeism at work by the actual number of hours worked; resulting presenteeism hours were divided by 8 to produce day-equivalents.

‡ Annual excess absenteeism and presenteeism days were added to calculate absenteeism/presenteeism days combined.

§ F-tests evaluated whether the set of conditions, as a whole, is related to work performance measures.

‖ *P* values from 2-sided tests.

¶ F-tests evaluated the variation in effects on the work performance measure across the 13 conditions.

SE, standard error; COPD, chronic obstructive pulmonary disease; GI, gastrointestinal.

ditions and control variables. Approximately 80% of coefficients are positive (ie, there are poorer work outcomes among workers with those specific chronic conditions) and 25% are significant at the 0.05 level. All significant coefficients are positive. Taken as a whole, these results show that the set of chronic conditions considered here is associated with elevated rates of work impairment. However, chronic conditions more consistently predicted decrements in absenteeism (100% of coefficients positive, 38.4% statistically significant) than either presenteeism (53.8% positive, 15.4% significant) or critical incidents (69.2% positive, 15.4% significant). Effects on presenteeism are significant and larger, although not significantly so, than effects on absenteeism for only 2 conditions: arthritis and asthma.

Before examining individual coefficients, significance tests were carried out to evaluate whether the set of conditions taken as a whole is significantly related to the outcomes. Significant overall associations were found for 3 of the outcomes ($F_{13,2340} = 1.7–6.0$, $P = 0.047–< 0.001$), the exception being critical incidents. Tests were also carried out to determine whether the effects differ significantly across conditions. Significant differences of this sort were found for 2 of the outcomes ($F_{12,2336} = 1.1–2.3$, $P = 0.35–0.00$), with the exceptions being absenteeism and critical incidents. When we combine absenteeism with presenteeism, 4 conditions are significant at the 0.05 level: arthritis, asthma, depression, and COPD–emphysema. The first 2 of these 4 significantly predict both absenteeism and presenteeism, whereas the latter 2 predict only absenteeism. The effect sizes of arthritis, asthma, and depression are in the range from 15.1 to 19.4 days per year. Even for workers with entry-level salaries, effects as large as these translate into annual salary-equivalent costs of \$1200 to \$1500, whereas the salary-equivalent costs are many times higher for executives. The effect size of COPD–emphysema is much higher (42.9 days per year) than the other conditions, but the standard error is large the result of the comparatively small number of respondents with COPD–emphysema. As a result, the coefficient of COPD–emphysema is not significantly different from the coefficients of the other statistically significant conditions ($z = 1.2–1.5$, $P = 0.131–0.219$).

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 4
Annual Excess Absenteeism Days, Presenteeism Days, and Net Negative Critical Incidents From Chronic Conditions Per 100 Workers in Four Occupations

	Reservation Agents		Customer Service Representatives		Executives		Railroad Engineers	
	No.	(SE)	No.	(SE)	No.	(SE)	No.	(SE)
Annual Excess Absenteeism and Presenteeism Days Per 100 Workers*								
Arthritis	226.4	(79.2)	217.5	(76.1)	167.6	(58.7)	328.0	(114.8)
Asthma	110.6	(49.2)	201.8	(89.9)	77.6	(34.6)	106.7	(47.5)
COPD/emphysema	128.6	(51.1)	60.0	(33.9)	17.1	(6.9)	85.7	(34.1)
Depression	181.3	(89.3)	219.1	(107.9)	58.9	(29.0)	133.0	(65.5)
Annual Excess Net Negative Critical Incidents Per 100 Workers†								
Depression	12.2	(5.3)	14.7	(6.4)	4.0	(1.7)	8.9	(3.9)
Chronic headaches	21.5	(7.9)	20.5	(7.7)	4.9	(1.8)	15.4	(5.7)

* Aggregate effects on annual absenteeism and presenteeism days combined are shown for the 4 conditions statistically significantly associated with this work performance measure.
† Aggregate effects on annual net negative critical incidents are shown for the 2 conditions statistically significantly associated with this work performance measure.
SE, standard error; COPD, chronic obstructive pulmonary disease.

Depression and chronic headaches are the only conditions that significantly predict critical incidents. The effect sizes are 1.0 to 1.1 critical incidents per year. It is impossible to put a dollar value on these effects without knowing specifics about the value to the company of a foregone incident defined as a "big success" or of the occurrence of an incident defined as a "big failure." The costs of these effects to the employer might be as great as, or even greater than, the costs associated with absenteeism and presenteeism.

Statistical tests were carried out to evaluate whether the 13 significant coefficients in Table 3 vary by occupation. Global tests failed to find significant between-occupation variation in these effects for any of the 4 outcomes ($F_{39,2301} = 1.1$, $P = 0.308$ for absenteeism; $F_{39,2301} = 1.0$, $P = 0.426$ for presenteeism; $F_{39,2301} = 0.9$, $P = 0.634$ for absenteeism–presenteeism; and $F_{39,2301} = 0.9$, $P = 0.614$ for critical incidents). More specific tests for each of the 13 coefficients found 2 individual coefficients that vary significantly across occupations. Seasonal allergies predict presenteeism differently across the occupations ($F_{3,2301} = 3.0$, $P = 0.029$), largely as a result of a less powerful effect among railroad engineers than respondents in the other occupations. Chronic headaches predict absenteeism–presenteeism differently across the occupations ($F_{3,2301} = 2.6$, $P = 0.052$), largely as a result of a more powerful effect among customer service representatives than respondents in the other occupations. Caution is needed in interpreting these specifications, however, because the insignificance of the global tests means that the 2 specifications could be the result of chance variation in the large number of tests.

Aggregate Associations Between Chronic Conditions and Work Performance

Although the effect sizes in Table 3 do not differ across the significant conditions, the conditions vary significantly in prevalence. As a result, the aggregate effects of the conditions differ. This variation is shown in Table 4, which presents the annual number of excess absenteeism–presenteeism days and critical incidents associated with each significant condition per 100 workers in the labor force. It is important to recognize that the results are per 100 overall workers, not per 100 workers with the condition, allowing differences in the prevalence of the condition to be taken into consideration in calculating the effects. The individual-level effect size estimates are assumed to be equal across occupations based on the results reported in the previous paragraph.

Part I of the table shows effects on combined absenteeism–presenteeism. Arthritis is associated with the largest aggregate effects in 3 of the 4 occupations. The exception is customer service representatives, among whom arthritis, asthma, and depression have very similar effects. The aggregate effects of arthritis range between 167.6 (executives) and 328.0 (railroad engineers) excess absenteeism–presenteeism days per year for every 100 workers. COPD–emphysema, in comparison, is associated with the smallest aggregate effects in 3 of the 4 occupations. The exception is reservation agents, among whom asthma and COPD–emphysema have similar effects. The aggregate effects of COPD–emphysema range between 17.1 (executives) and 128.6 (reservation agents) excess absenteeism–presenteeism days per year for every 100 workers. Part II of the table shows effects on critical incidents. Chronic headaches are consistently associated with larger aggregate effects than depression. The aggregate effects of chronic headaches range between 4.9 (executives) and 21.5 (reservation agents) excess critical incidents per year for every 100 workers. The aggregate effects of depression, in comparison, range between 4.0 (executives) and 14.7 (customer service representatives) excess critical incidents per year for every 100 workers.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

Discussion

Three limitations are important to keep in mind when interpreting these results. The surveys had comparatively low response rates and self-reports could be only imperfectly related to true conditions and true work performance; both issues could have introduced bias and adversely affected the accuracy of our estimates. These 2 limitations should lead to caution in interpreting results, albeit caution that is tempered by the results of methodologic research on the magnitude of measurement error. In an earlier series of HPQ workplace surveys, we found that variation in response rate resulting from differences in intensity of recruitment effort is not significantly related to condition prevalence or to the estimated effects of conditions on work performance.[24] It is likely, based on this finding, that the results reported here, are not importantly affected by the survey response rates.

With regard to measurement of chronic conditions, methodologic research reviewed previously has shown that chronic conditions are generally self-reported with good accuracy.[16–19] This means that any bias introduced by this type of measurement error should be minimal. To the extent that this bias exists, methodologic research suggests that prevalence will be underestimated, making the aggregate estimates conservative. However, because conditions that cause significant activity limitation are less likely to be unreported than those that do not cause activity limitation, individual-level estimates are likely to be overestimated. With regard to errors in self-reports about work impairments, the HPQ calibration studies showed that there is good concordance of self-reports both with payroll records and with archival supervisor and peer performance ratings.[13] These results suggest that any bias in the estimated effects of conditions on work performance is likely to be minimal.

The third limitation of the results in this article is that they focus on the additive effects of separate individual conditions. Although such information is useful, it is important to recognize that many people with chronic conditions suffer from more than 1 condition,[25] and that co-occurring conditions are, in general, more impairing than pure conditions.[26] We were unable to evaluate the joint effects of co-occurring conditions in the current report because a very large sample is required for statistically powerful analysis of these interactive effects. Such a sample is currently being assembled by the HPQ data-pooling consortium (www.hpq.org). Replication and extension of the current results to evaluate the effects of co-occurring conditions will be carried out in this database once it is assembled and ready for analysis.

Within the context of these limitations, the results reported here are broadly consistent with previous research in finding a high prevalence of chronic conditions among working people[14,15] as well as substantial effects of chronic conditions on work performance.[27,28] These results also extend beyond previous research, which focused largely on absenteeism and show that absenteeism is generally associated with smaller decrements in work performance than presenteeism in the occupations considered here. They also show that the chronic conditions considered here more consistently have adverse effects on absenteeism than presenteeism. Two exceptions are arthritis and asthma, both of which have significant effects on both presenteeism and absenteeism and effect sizes that are larger (although not significantly so) for the former than the latter. It is noteworthy that the presenteeism coefficient effect sizes are also substantial in magnitude, although not statistically significant, for cancer, heart disease, and COPD–emphysema. All 3 of these are low-prevalence conditions, which means that a larger sample might show that they have meaningful effects on presenteeism. It is also noteworthy that the surveys were not carried out during allergy season, which means that the effects of seasonal allergies were almost certainly underestimated.

We are not aware of any previous studies that have examined the effects of chronic conditions on critical incidents, although research has documented effects of some chronic conditions, most notably seasonal allergies[29] and substance abuse,[30] on workplace accidents. Only 2 of the conditions examined here, depression and chronic headaches, were found to be significant predictors of critical incidents. Although it is conceivable that these could be chance associations resulting from replication across a large number of conditions, the consistency in their effects in all 4 occupations suggests that they are genuine rather than the result of chance.

Our finding about comparative effects is consistent with the few previous studies that compared the relative effects of individual chronic conditions on work performance. Specifically, arthropathies, asthma, COPD, and mood disorders have been independently reported to be among the top conditions responsible for annual work-loss days in the nationally generalizable Medical Expenditure Panel Survey (MEPS).[4] Among specific conditions, arthritis and depression stand out as especially important in the current study. Arthritis is important because it uniquely combines high prevalence with significant effects on both absenteeism and presenteeism, resulting in the highest aggregate absenteeism–presenteeism effects in the study. This finding is consistent with independent evidence that arthritis has enormous workplace costs.[31] Depression is important because it is the only condition that has significant effects on both absenteeism–presenteeism and critical incidents. This finding is consistent with evidence from previous studies that depression causes significant decrements in a

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

wide range of functional domains.[32,33] Reasons for depression's impairment across a variety of aspects of performance could include its multifaceted adverse effects on attention, memory, other aspects of cognition, motivation, relations with peers and superiors, self-efficacy, and even its ability to cause purposeful self-harm.[34]

These results have implications for a variety of stakeholders interested in reducing work impairments from chronic illness. Perhaps the most important is that cost-effectiveness analyses should move beyond simply enumerating absenteeism to include other aspects of work performance to avoid underestimating the workplace costs of illness. It is interesting to note that none of the conditions found in the current study to be significant predictors of either combined absenteeism–presenteeism or critical incidents would have been missed if absenteeism had been the only aspect of work performance considered. However, the magnitude of the effects would have been dramatically underestimated. For example, the estimated individual-level effect of arthritis on absenteeism (8.5 days per year) is less than half as large as the estimated effect on absenteeism and presenteeism combined (17.8 days per year). Differences as large as this can dramatically change the conclusions of cost-effectiveness analyses.

Another important implication is that policymakers should consider information about the workplace costs of illness in current debates and planning activities regarding insurance coverage and healthcare expenditures. It is striking that no meaningful relationship exists in the United States between the magnitude of impairments caused by different conditions and the amount spent on these conditions in direct treatment costs.[4] Equally striking is the fact that treatment costs in the United States for patients with conditions that have the greatest role disabilities relative to direct treatment costs (mood disorders, chronic pulmonary diseases, and arthropathies) largely go to treat comorbid problems rather than the disabling conditions themselves.[5] These results clearly show that allocation of treatment resources is currently not being made in a way designed to maximize return on investment (ROI) with respect to role performance.

To develop and target successful interventions to improve this state of affairs, employers and health plans need better information than they currently have about the mechanisms through which specific conditions impair work performance. For example, interventions can be designed to enhance case finding and aggressive outreach to untreated workers in an effort to increase the proportion of ill workers who receive treatment, to help providers improve the intensity or quality of existing treatments for patients already in treatment, or to help patients adhere to their regimens. Not all of these interventions make equal sense for all conditions. Disease management programs that incorporate all of these features exist for all the conditions found in the current study. These programs have been found to have significant effects on work performance.[35–40] Optimal allocation of resources across these features and conditions requires information on the particular needs for specific conditions as well as the cost-effectiveness of individual program components.

Available evidence from economic analyses suggests that interventions aimed at some combination of outreach, best practices, and adherence can lead to a favorable ROI for the employer–purchaser of health care. For example, simulations based on secondary analyses of clinical trials for the treatment of depression suggest that the direct costs of outreach and treatment of depression among employed people could be totally offset by the indirect cost savings associated with the decreased work impairment that occurs when depression remits.[41] Rational economic evaluations of the likely ROI of disease management interventions for chronic conditions should be factored into employer cost–benefit calculations regarding employer-sponsored health insurance coverage. Furthermore, once an investment in employer-sponsored health care is made, longitudinal workplace health and productivity tracking surveys should be used to help provide quality assurance and guide subsequent ROI evaluations. The availability of inexpensive standardized procedures for such ongoing surveys (www.hpq.org) makes this approach to employer healthcare investment all the more feasible.

Acknowledgments

Research supported by The World Health Organization, the John D. and Catherine T. MacArthur Foundation Program on Health and Work Performance, and by unrestricted educational grants from Glaxo-Wellcome, Pfizer, Schering-Plough, and Searle. Preparation of this report was also supported by NIMH grants K01 MH01651 (Dr Wang), R01 MH61941 (Dr Wang), and K05 MH00507 (Dr Kessler). The complete text of the HPQ is posted at the http://www.hcp.med.harvard.edu/hpq/.

The authors gratefully acknowledge the assistance of Bill Whitmer and his colleagues at the Health Enhancement Research Organization (HERO) in recruiting companies to participate in the study, the staff of the participating companies in facilitating the surveys, and the staff of DataStat, Inc. in carrying out the surveys.

References

1. Burner ST, Waldo DR, McKusick DR. National health expenditures projections through 2030. *Health Care Financing Review*. 1992;14:1–29.
2. Weinstein MC, Fineberg HV, Elstein AS, et al. *Clinical Decision Analysis*. Philadelphia: WB Saunders; 1980.
3. Gold MR, Siegel JE, Russell LB, et al. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press; 1996.
4. Druss BG, Marcus SC, Olfson M, et al. The most expensive medical conditions in America. *Health Aff*. 2002;21:105–111.
5. Druss BG, Marcus SC, Olfson M, et al. Comparing the national economic burden of five chronic conditions. *Health Aff*. 2001;20:233–241.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

6. Krahn MD, Berka C, Langlois P, et al. Direct and indirect costs of asthma in Canada. *Canadian Medical Association Journal*. 1996;154:821–831.
7. Weiss KB, Gergen PJ, Hodgson TA. An economic evaluation of asthma in the United States. *N Engl J Med*. 1992;326: 862–866.
8. Guico-Pabia CJ, Murray JF, Teutsch SM, et al. Indirect costs of ischemic heart disease to employers. *American Journal of Managed Care*. 2001;7:27–34.
9. Greenberg PE, Stiglin LE, Finkelstein SN, et al. The economic burden of depression in 1990. *J Clin Psychiatry*. 1993;54:405–418.
10. Simon GE, Barber C, Birnbaum HG, et al. Depression and work productivity: the comparative costs of treatment versus nontreatment. *J Occup Environ Med*. 2001;43:2–9.
11. Testa MA, Simonson DC. Health economic benefits and quality of life during improved glycemic control in patients with type 2 mellitus. *JAMA*. 1998;280: 1490–1496.
12. Harlan WR, Parsons PE, Thomas JW, et al. *Health Care Utilization and Costs of Adult Cardiovascular Conditions, United States, 1980.* National Center for Health Statistics, Public Health Service, U.S. Government Printing Office, DHHS Pub. No. 89-20407; Washington, DC; 1989.
13. 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.
14. Eisenberg DM, Kessler RC, Foster C, et al. Unconventional medicine in the United States. Prevalence, costs, and patterns of use. *N Engl J Med*. 1993;328: 246–252.
15. Eisenberg DM, Davis RB, Ettner SL, et al. Trends in alternative medicine use in the United States, 1990–1997: results of a follow-up national survey. *JAMA*. 1998;280:1569–1575.
16. National Center for Health Statistics. Evaluation of National Health Interview Survey diagnostic reporting. *Vital Health Stat 2*. 1994;120:1–116.
17. Heliovaara M, Aromaa A, Klaukka T, et al. Reliability and validity of interview data on chronic diseases. The Mini-Finland Health Study. *J Clin Epidemiol*. 1993;46:181–191.
18. Halabi S, Zurayk H, Awaida R, et al. Reliability and validity of self and proxy reporting of morbidity data: a case study from Beirut, Lebanon. *Int J Epidemiol*. 1992;21:607–612.
19. Kriegsman DM, Penninx BW, van Eijk JT, et al. 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.
20. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960;20:37–46.
21. Gross R, Bentur N, Elhayany A, et al. The validity of self-reports on chronic disease: characteristics of underreporters and implications for the planning of services. *Public Health Rep*. 1996;24:167–182.
22. 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*. 1996;86: 706–711.
23. Kessler RC, Andrews G, Mroczek D, et al. The World Health Organization Composite International Diagnostic Interview Short Form (CIDI-SF). *International Journal of Methods in Psychiatric Research.* 1998;7:171–185.
24. Wang PS, Beck AL, McKenas DK, et al. Effects of efforts to increase response rates on a workplace chronic condition screening survey. *Med Care*. 2002;40: 752–760.
25. Dewa CS, Lin E. Chronic physical illness, psychiatric disorder and disability in the workplace. *Soc Sci Med*. 2000;51: 41–50.
26. Ormel J, Von Korff N, Ustun TB, et al. Common mental disorders and disability across cultures: results from the WHO collaborative study on psychological problems in general health care. *JAMA*. 1994;272:1741–1748.
27. Centers for Disease Control and Prevention. Health-related quality of life and activity limitation-eight states, 1995. *MMWR Morb Mortal Wkly Rep.* 1998;47: 134–139.
28. Hoffman C, Rice D, Sung HY. Persons with chronic conditions: their prevalence and costs. *JAMA*. 1996;276:1473–1479.
29. Kay GG, Quig ME. Impact of sedating antihistamines on safety and productivity. *Allergy Asthma Proc*. 2001;22:281–283.
30. Veazie MA, Smith GS. Heavy drinking, alcohol dependence, and injuries at work among young workers in the United States labor force. *Alcohol Clin Exp Res*. 2000;24:1811–1819.
31. Leigh JP, Seavey W, Leistikow B. Estimating the costs of job related arthritis. *J Rheumatol*. 2001;28:1647–1654.
32. Hays RD. Functioning and wellbeing outcomes of patients with depression compared with chronic general medical illnesses. *Arch Gen Psychiatry*. 1995;52: 11–19.
33. 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.
34. Klerman GL, Weissman MM. The course, morbidity and costs of depression. *Arch Gen Psychiatry*. 1992;49:831–834.
35. Katon W, Von Korff M, Lin E, et al. Collaborative management to achieve treatment guidelines: impact on depression in primary care. *JAMA*. 1995;273: 1026–1031.
36. Katon W, Robinson P, VonKorff M, et al. A multifaceted intervention to improve treatment of depression in primary care. *Arch Gen Psychiatry*. 1996;53:924–932.
37. Wells KB, Sherbourne C, Schoenbaum M, et al. Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *JAMA*. 2000;283:212–220.
38. Legorreta AP, Leung KW, Berkbigler D, et al. Outcomes of a population-based asthma management program: quality of life, absenteeism, and utilization. *Ann Allergy Asthma Immunol*. 2000;85:28–34.
39. Sobkowski DA, Maquera V. Critical path case management: the headache clinic. *Best Practices and Benchmarking in Healthcare*. 1996;1:198–202.
40. Fries JF, Carey C, McShane DJ. Patient education in arthritis: randomized controlled trial of a mail-delivered program. *J Rheumatol*. 1997;24:1378–1383.
41. Kessler RC, Barber CB, Birnbaum HG, et al. Depression in the workplace: effects on short-term work disability. *Health Aff*. 1999;18:163–171.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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.

Copyright © by American College of Occupational and Environmental Medicine

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,

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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 oc-

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

curs 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?"

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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	—
χ^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	—
χ^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.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

(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:

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

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

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

$$= [(acd/c]^{1/2} = (ad)^{1/2} = (a^2)^{1/2} = (d^2)^{1/2} \quad (2a)$$

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.

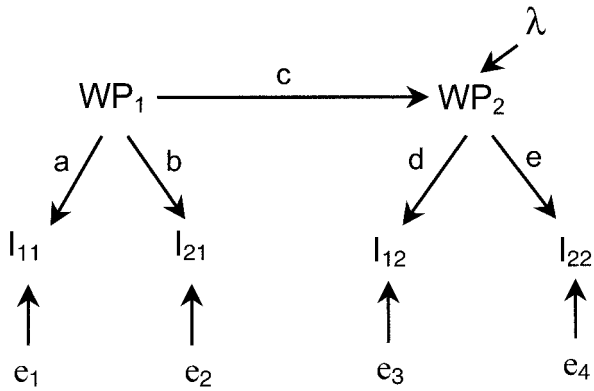


Fig. 1. A two-indicator two-time model of the stability of true presenteeism (WP_t) and the reliability of measured presenteeism indicators (I_{nt}). As in more conventional exploratory factor analysis models, measurement errors (e_{1-4}) are assumed to be independent of true presenteeism and of each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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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.

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.

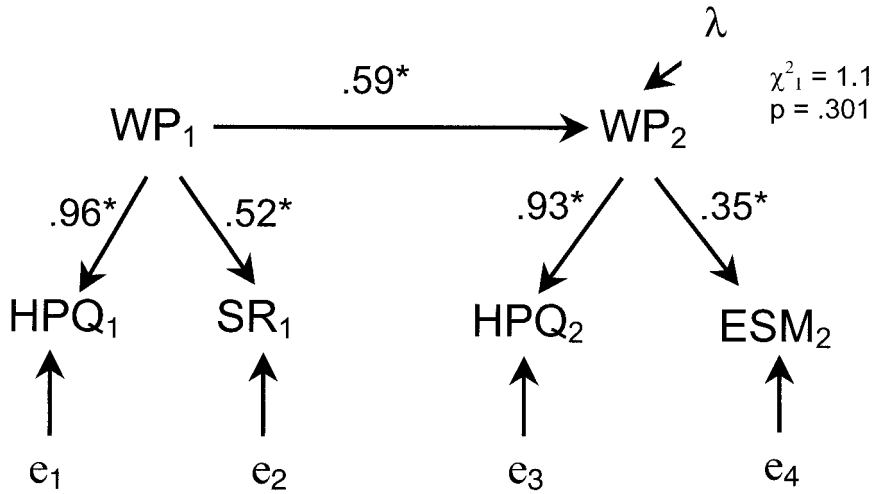


Fig. 2. Parameter estimates of the two-indicator two-time model of the stability of true presenteeism (WP_t) and the reliability of the HPQ presenteeism scale (HPQ_t). True presenteeism is included by the HPQ at both times by supervisor ratings at time 1 (SR_1), and by an aggregate performance measure 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.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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):

$$HPQ\mathrm{t} = b_{0\mathrm{t}} + b_1\mathrm{C_t} + b_2\mathrm{V_t} + b_3\mathrm{U_t} \qquad (3)$$

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 as-

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

sociation 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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

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 under-reporters 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 non-specific 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.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

Health & Productivity Management, 3(2), pp. 4-8

Rationalizing Employer Health Care Spending: The Atlanta and Midwest Business Group on Health HPQ Demonstration Project

Ronald C. Kessler, Ph.D.
Professor, Department of Health Care Policy, Harvard Medical School

Dennis E. Richling, M.D.
President, Midwest Business Group on Health

Sean Sullivan
President & Chief Executive Officer,
Institute for Health and Productivity Management

The research reported here was 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. The authors gratefully acknowledge the help of questionnaire wording experts Stephanie Chardoul, Lou Magilavy, Beth Ellen Pennell, Kent Peterson, Tom Wilkinson, and Debbie Zivan in developing and revising the HPQ, the assistance of Bill Whitmer and his colleagues at the Health Enhancement Research Organization (HERO) in recruiting companies to participate in the calibration survey, the staff of the participating companies in facilitating the sample selection and abstraction of objective performance measures and absenteeism records, and Joe Myer, Tom Wilkinson, and their staff at DataStat, Inc. in carrying out the calibration survey. The complete text of the HPQ is posted at the http://www.hcp.med.harvard.edu/ncs/hpq. Address comments to RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; 617-432-3587 (phone); 617-432-3588 (fax); kessler@hcp.med.harvard.edu (email).

The health and productivity movement has now come to a critical juncture at which we either have to put up or shut up with regard to our claim that employer investments in the health care of workers can be conceptualized, at least in part, as human capital investments and can be evaluated in terms of return on investment (ROI). We now have a number of different players in the health and productivity research arena. Interest in the social consequences of illness has broadened in the past decade as epidemiologists have joined health economists and health services researchers to develop methods aimed at rationalizing the allocation of health care resources (Gold, Siegel et al. 1996; Sloan 1996; Tarlov, Ware et al. 1989; Patrick and Erickson 1993; Kessler, Greenberg et al. 2001; Pauly, Nicholson et al. 2002). These studies show clearly that the "indirect" workplace costs of some untreated illnesses are enormous.

Depression, for example, has been estimated to cause an annual loss of $33 billion in work absenteeism in the U.S. (Greenberg, Kessler et al. 1996). Given the low rate of depression treatment (Kessler, Barber et al. 1999) and the fact that treatment of depression substantially improves work performance (Coulehan, Schulberg et al. 1997; Wells, Sherboume et al. 2000), these data suggest that the expansion of employer-sponsored depression outreach and treatment programs might be cost-effective for American business. (Kessler, Barber et al. 1999; Simon, Barber et al. 2001). And the same could be true for prevention, treatment, and disability management programs associated with a number of other illnesses (Kessler, Almeida et al. 2001; Boonen, van-der Heijde et al. 2002; Reginster and Khaltaev 2002).

Even when the focus of intervention is on potentially cost-effective programs, though, employers have difficulty thinking of employee health care as an investment.

Health care costs have been increasing so rapidly in recent years that most corporations are trying to reduce, not expand, their health care budget. Yet, rational management of health care investments is even more important now than in the past because employers must be able to determine what to keep, reduce, or eliminate in their current health plans. Reducing health programs that effectively address commonly occurring health problems with major impacts on work performance can be penny-wise and pound-foolish (Rosenheck, Druss et al. 1999). This means that employers need that clearly differentiates the relative workplace effects of individual health plan components for expansion, as well as reduction, of employer-sponsored health care benefits.

The information gap that thwarts efforts at rationalization

Most employers, however, lack any of the three critical types of information needed to calculate the ROI for investments in employee health care. These include information about: (1) the number of workers with specific health problems (included both those who are being treated and those not being treated); (2) the effects of each specific health problem on workplace functioning; and (3) the cost-effectiveness of available interventions by health care professionals who treat their employees. As a result, medical directors and benefits managers are unable to make a convincing business case for rationalizing employee health care benefits.

Even though few employers can assemble databases with the integrated information of these three sorts that are needed to rationalize purchasing decisions about health care for their employees, proxy versions of these data can be collected easily and inexpensively in employee health-and-productivity surveys. When these surveys are repeated annually, they can be used to track linked information about the prevalence of

employee illnesses, treatment of these illnesses, and the effects of these illnesses on work functioning. In cases where archival data are available (e.g., medical claims or performance reviews) and can be linked to the survey data, the self-reports can be evaluated for accuracy as well.

Such surveys can provide employers with answers to several key questions: How many of my workers have health problems of various sorts? How many of these problems are currently being treated? What are the "indirect" costs of untreated conditions measured by reduced worker performance and increased absence from sickness, workplace accidents, and disability? What are the comparable costs of conditions that are treated? If an intervention is put into place to improve treatment of a condition, changes measured by re-administering the survey periodically can help answer two other questions: How effective is the intervention in reducing the workplace costs of the condition? What is the ROI on the intervention? And the logic of these evaluations is identical to that of test market and tracking studies routinely done by the market research departments of most major corporations.

The Harvard Health and Work Performance Initiative

Recognizing the need for answers to these questions in order to help employers rationalize health care purchasing decisions, the John D and Catherine T MacArthur Foundation funded a group of researchers from Harvard Medical School to develop a system that would allow this information to be collected and reported quickly and inexpensively based on employee health-and-productivity surveys and electronic report-generating software. Subsequent funding of this Harvard Health and Work Performance

Initiative also came from the Robert Wood Johnson Foundation and a number of industry sponsors.

The first phase of the Harvard initiative focused on developing the Health and Work Performance Questionnaire (HPQ). The HPQ is a short self-administered questionnaire to collect valid information among workers about the prevalence of treated and untreated health problems and about the effects of these health problems on work performance. The HPQ was created in collaboration with the Assessment and Classification Group of the World Health Organization and the medical directors of several major corporations. A complex and comprehensive series of methodological studies described elsewhere (Kessler et al., 2003; in press) was used to develop and validate the HPQ. Based on the impressive results of these methodological studies, the HPQ was selected by WHO for inclusion in a series of international health surveys they carried out to assess the prevalence and societal burdens of mental versus physical disorders. These surveys include interviews with nearly 200,000 respondents in 28 countries around the world. In addition, the HPQ is now being used in workplace health surveys in a number of large corporations in the US, Europe, and Asia. HPQ Administration.

The HPQ can be administered either electronically or by paper and pencil, either as a stand-alone survey or in conjunction with an existing HRA survey in collaboration with their developers. A clinical trials version of the HPQ has also been developed and is now being used by several pharmaceutical companies to measure secondary outcomes in their experimental treatment trials. The use of this common metric in controlled clinical studies as well as in workplace surveys facilitates efforts to crosswalk results from the

former to the latter in the service of making projections regarding potential workplace effects of treatment enhancements.

The second phase of the Harvard initiative focused on developing an electronic HPQ report-generating system. This is an on-line system that can be used by employers to interpret the results of their HPQ surveys and to compare their results either with the results of nationally and internationally representative HPQ benchmark surveys, with surveys carried out by other companies in their same markets, or with their own HPQ surveys from past years. The HPW report-generating software was developed in close collaboration with an advisory group of medical directors, benefits managers, and health care consultants to optimize the ease and flexibility of the tool for generating useful survey results. A demonstration version of the HPQ report-generating system along with a copy of the questionnaire and useful background information on the HPQ Consortium are posted at www.HPQ.org.

The third phase of the Harvard initiative focused on implementing workplace health care experiments in conjunction with HPQ surveys. The largest of these is a $6 million workplace depression treatment experiment funded by the National Institute of Mental Health and carried out in collaboration with United Behavioral Health in a number of large national corporations. The intervention features expanded employee outreach and best-practices treatment of workers who suffer from major depression. The outcomes are HPQ measures of absenteeism, work performance, workplace accidents, and disability. The goal is to determine whether a positive ROI can be demonstrated for the intervention over follow-up periods of one, two, and three years.

The Atlanta and Midwest Business Group on Health HPQ Demonstration Project

The fourth phase of the Harvard Initiative is just now starting. In this phase, annual HPQ surveys are being carried out in the workforces of collaborating workplaces in two major health care markers, Atlanta and Chicago, as part of comprehensive demonstration programs aimed at rationalizing employer health care purchasing and quality control monitoring. These programs are developing integrated databases that combine pharmacy and medical claims data from health care providers, workplace performance and disability data from employers, and self-report health and productivity data from the HPQ surveys for the purposes of generating data on the workplace costs of illness and the cost-effectiveness of current treatment approaches, sharing these data with employers, working with employers to discuss these data and their implications with their health care providers, and evaluating the ROI of any modifications health plans make in their services as a result of this feedback. We also hope to use these demonstration sites and the integrated databases we develop in them as a natural laboratory to test a number of innovative ideas about targeted disorder-specific disease management programs. To the extent that our experience in these sites shows that we can cost-effectively rationalize health care spending and quality assurance and achieve a positive ROI, the systems and procedures developed in these demonstration sites will subsequently be disseminated to other markets throughout the country. The remainder of this article review the three main data acquisition areas associated with the HPQ in this demonstration project.

The prevalence of health problems among working people

The prevalence of individual conditions in the workforce must be known in order to target disease-specific health care investments. This information is available most

directly and accurately from annual physical examinations that some employers make available for all their workers, but few employers capture this information and study the prevalence of individual untreated health problems. In the absence of such data, medical claims can be used to study treated illnesses. These records tell us little, however, about the prevalence of untreated conditions because of wide variation in the proportion of workers with a given health problem who seek treatment. This is a critical issue, as untreated conditions often have substantial adverse effects on the workplace.

Indirect estimates of the prevalence of untreated health problems can be obtained from archival sources such as the U.S. National Health Interview Survey (NHIS), which includes a chronic conditions checklist that can be used to generate national estimates of the prevalence of individual conditions. The NHIS data can also be disaggregated to produce refined estimates of prevalence among workers in specific occupations and regions of the country. Similar indirect estimates can be obtained from various data warehouses, such as HPQ data warehouse. Indirect estimates cannot be used, though, to evaluate the effects of workplace interventions. As a result, the Atlanta and Midwest Business Group on Health demonstration project is collecting direct information about the prevalence of untreated conditions by administering HPQ surveys that systematically ask employees about their health.

These surveys can be carried out inexpensively, especially in companies where a large proportion of workers have e-mail addresses and can be surveyed electronically. Self-report surveys are, of course, limited to information that workers are able to report – such as information about symptom-based conditions and silent-conditions that have been diagnosed in health examinations. The majority of the commonly occurring health

problems that cause serious impairment in the workplace are symptom-based, including such conditions as arthritis, asthma, emotional disorders, migraine, neck/back pain, and seasonal allergies. In addition, a high proportion of the workers with untreated silent conditions (e.g. hypertension, hypercholesterolemia) are aware that they have these conditions based on previous medical examinations. As a result, self-report surveys can provide very useful data on the prevalence of treated and untreated conditions when employers are not able to obtain data from physical examinations for a large part of the workforce.

The effects of health problems on workplace functioning

Even the most common health problems will not be targets for employer investment unless they have a serious impact on work functioning – on sickness absence, quality of performance on the job, work-related accidents, and disability (Blum and Roman 1993). However, because previous studies show that the effects of particular health problems on work functioning vary substantially across occupations (Kessler, Mickelson et al. 2001), it is necessary to tailor estimates of likely effects to the composition of the workforce. The ideal way to assess work performance, of course, is objective performance-based assessment. Some employers have developed assessment systems of this sort for at least some of their workers (Grote 1996; Harbour 1997), but these systems vary enormously in coverage and sophistication. As a result, most employers lack individual-level work performance data for the majority of their employees. Even basic data on sickness absence are generally available only for blue-collar and pink-collar workers (but not for white-collar workers), while data on job

performance either are nonexistent, superficial, or very difficult to obtain in machine-readable form in most companies.

In the absence of objective performance data for a large proportion of the workforce, self-report measures are usually the most feasible tools for obtaining individual-level information on work functioning. The most compelling of these focus on single occupations and include questions that are tailored to the unique demands of those occupations (Sawyer, Latham et al. 1999; Warr and Conner 1999; Neal, Griffen et al. 2000), but these are not useful for general-purpose studies. As a result, a number of health researchers have developed self-report measures of work performance that are designed for use across a wide range of occupations (Lynch and Riedel 2001). The HPQ is now the most widely-used and thoroughly validated of these measures.

Although self-report survey data can be analyzed to generate estimates of the effects of health problems on work functioning, the results are much more meaningful to employers if they are translated into monetary terms from raw performance metrics (e.g., average number of lost work days associated with a particular illness). The appropriate translation rules can be quite complex, though, and are not necessarily proportional to salary. The unscheduled sickness absence of an airline reservation agent, for example, might lead to customers waiting longer to speak to an agent and agents on duty feeling more pressure then usual. But the costs of these inconveniences to the employer are probably minimal unless the delays are so long and persistent that customers go to a different airline to purchase their tickets. The unscheduled sickness absence of a flight attendant, in comparison, can cause a flight departure delay (due to FAA staffing requirements for number of personnel needed for a flight to depart) that costs the airline

as much as $5000 per hour in additional gate fees. Similarly, the illness of a salesman, if it leads to the loss of a major contract, can cost a corporation millions of dollars even though the salesman's salary is only a fraction of that amount. The rules used to convert information about the impact of health problems on work functioning into a dollar metric should capture these differences. An important challenge in monetizing the HPQ results in the demonstration markets will be to create realistic transformation rules that capture these kinds of complexities making use of the best current thinking about this problem (Berger, in press; Pauly, Nicholson et al., 2002; Berger, Murray et al., 2001).

The cost-effectiveness of health-related interventions

Information about the cost-effectiveness of available interventions is obviously important to making good decisions about health care investments. The word "interventions" is used broadly here to include both conventional medical treatments as well as other things that affect worker health, such as ergonomic chairs and on-site yoga classes for workers with low back pain. Interventions can be targeted at prevention, early detection and treatment, improvement in the quality of treatment, disability management or some combination. Ideally, cost-effectiveness would be evaluated first in controlled trials before an employer would invest in the intervention. The settings of such trials, however, bear little resemblance to real clinical settings. Consequently, employers need a method for monitoring the results from new health care interventions in their own companies.

This is especially the case when health care is being proposed as an investment opportunity rather than a necessary cost of doing business – inviting comparisons with other investment opportunities. The same standards for evaluating ROI in other areas of

the business will then be applied to health care, and data of equivalent quality will be demanded to make the business case for health care investments. The same methods used to evaluate the quality of materials purchased, the cost-savings created by new production technologies or the effects of direct to consumer ad campaigns are needed to evaluate the effects of investments in employee health care on work functioning.

In most large corporations, these methods involve longitudinal tracking studies in test markets. Measures of output are defined for each potential kind of investment (e.g., widgets produced per hour per worker, sales or market share per month in each Nielsen market). Scores on these measures are plotted on a monthly or quarterly basis within production or sales units (e.g., for each plant or each company region). Changes resulting from the interventions (e.g., new advertisement campaigns, new computer systems, new sales force incentive structures) are evaluated by charting the trend line after their introduction. Where the intervention is tried out in test markets, the evaluation compares changes in the trend line in these markets with changes in other markets.

Exactly these same methods will be used to evaluate the effects of health care interventions in the Atlanta and Midwest Business Group on Health demonstration project using HPQ trend data available on individual-level relationships between health problems and work functioning. Such data will make it possible to answer questions like the following: What is the prevalence of, say, arthritis in the work force? What is the total "indirect" cost of arthritis in the monetized cost of sickness absence, diminished work performance, workplace accidents, and disability? How did this cost change when a new arthritis disease management program was put into place in two test markets? What was the ROI from this intervention measured by comparing the indirect cost savings with the

direct costs? Where integrated data on health care expenditures are available, information about direct medical treatment off-set savings (i.e., changing other health care costs) can be added to the indirect savings.

This same logic can be used to evaluate the effects of reducing investments in health care. Rosenheck et al. (1999) showed this in a before-and-after market tracking study of a large company that restricted benefits for treating mental health problems. The tracking data showed that, as expected, the direct costs of outpatient mental health treatment decreased dramatically as a result of this intervention. Other costs increased, however, for the workers who were previously in mental health treatment. These included the costs of inpatient mental health treatment, outpatient primary care treatment, and sickness absence. When all of these costs were taken into consideration, this intervention had a large negative ROI. The company had launched its intervention using just the data showing that the direct costs of mental health treatment were on the rise. They lacked information about the effects of mental health problems on work functioning or on other health care costs. Had such data been available, the intervention might never have been implemented. Even when well-conceived interventions are launched, they might prove to have a negative ROI for reasons that could not be foreseen. Having systems in place to monitor the changes that Rosenheck et al. were able to track can minimize the damage by canceling these interventions quickly.

The Atlanta and Midwest Business Group on Health HPQ demonstration project is designed to capture exactly the same type of information as in the Rosenheck evaluation without requiring the participating employers to have the same state-of-the-art integrated database that Rosenheck and his colleagues had in their evaluation. However,

unlike the Rosenheck study and other claims-based evaluations, we will also have information on untreated health problems that can be used to go beyond evaluating the ROI of externally derived interventions (i.e., interventions, like the one evaluated by Riosenheck, that were put in place as cost-savings measures without consideration of possible effects on work performance) to pinpoint potentially important intervention targets. The latter can be associated either with under-treated conditions that have substantial workplace costs (where the intervention would involve expanded outreach and treatment) or with conditions that are expensive to treat that have low workplace costs (where the intervention would involve restricting benefits). Initial HPQ monitoring after implementation of the interventions will be used to determine ROI, while ongoing HPQ monitoring in subsequent years will be used for ongoing quality assure auditing to guarantee that any positive ROI found in the initial evaluation is maintained over time.

Next steps

As the principals of the undertaking in Atlanta (SS), the Midwest Business Group on Health (DR), and the Harvard Medical School data coordinating center in Boston (RK), we hope to keep the readership of this magazine informed as the HPQ demonstration project unfolds over the next five years. We will have information on the indirect costs of illness within a year, information on the ROI of early responses to the baseline results a year later, and more detailed data on successively more refined modifications of plan design and services in subsequent years. This is the first major initiative in a new generation of health and productivity research that has enormous potential value for our field. We need all the encouragement and good wishes you can give us. We look forward to updating you periodically as the project unfolds.

References

Berger, M. L., Howell, R., Nicholson, S. & Sharda, C. (2003). Investing in healthy human capital. *J Occup Environ Med* **45,** 1213-1225.

Berger, M. L., Murray, J. F., Xu, J. & Pauly, M. (2001). Alternative valuations of work loss and productivity. *J Occup Environ Med* **43,** 18-24.

Blum, T. C. & Roman, P. M. (1993). Alcohol consumption and work performance. *Journal of Studies on Alcohol* **54,** 61-70.

Boonen, A., van der Heijde, D., Landewe, R., Spoorenberg, A., Schouten, H., Rutten-van Molken, M., Guillemin, F., Dougados, M., Mielants, H., de Vlam, K., van der Tempel, H. & van der Linden, S. (2002). Work status and productivity costs due to ankylosing spondylitis: comparison of three European countries. *Annals of the Rheumatic Diseases* **61,** 429-437.

Coulehan, J. L., Schulberg, H. C., Block, M. R., Madonia, M. J. & Rodriguez, E. (1997). Treating depressed primary care patients improves their physical, mental, and social functioning. *Archives of Internal Medicine* **157,** 1113-1120.

Gold, M. R., Siegel, J. E., Russell, L. B. & Weinstein, M. C. (1996). *Cost-Effectiveness in Health and Medicine.* Oxford University Press: Oxford.

Greenberg, P. E., Kessler, R. C., Nells, T. L., Finkelstein, S. N. & Berndt, E. R. (1996). Depression in the Workplace: An Economic Perspective. In *Selective Serotonin Reuptake Inhibitors: Advances in Basic Research and Clinical Practice* (eds, Feighner, J. P. & Boyer, W. F.), pp. John Wiley & Sons Ltd.: New York, NY.

Grote, D. (1996). *The Complete Guide to Performance Appraisal.* AMACOM: New York.

Harbour, J. L. (1997). *The Basics of Performance Measurement.* Productivity, Inc.: New York.

Kessler, R. C., Almeida, D. M., Berglund, P. & Stang, P. (2001a). Pollen and mold exposure impairs the work performance of employees with allergic rhinitis. *Annals of Allergy, Asthma & Immunology* **87,** 289-295.

Kessler, R. C., Barber, C., Beck, A., Bergland, P., Cleary, P. D., McKenas, D., Pronk, N. P., Simon, G. E., Stang, P., Usten, T. B. & Wang, P. S. (2003). The World Health Organization Health and Work Performance Questionnaire (HPQ). *Journal of Occupational and Environmental Medicine* **45,** 156-174.

Kessler, R. C., Barber, C. B., Birnbaum, H. G., Frank, R. G., Greenberg, P. E., Rose, R. M., Simon, G. E. & Wang, P. S. (1999). Depression in the workplace: effects on short-term work disability. *Health Affairs* **18,** 163-171.

Kessler, R. C., Greenberg, P. E., Mickelson, K. D., Meneades, L. M. & Wang, P. S. (2001b). The effects of chronic medical conditions on work loss and work cutback. *Journal of Occupational and Environmental Medicine* **43,** 218-225.

Kessler, R. C., Mickelson, K. D., Barber, C. B. & Wang, P. S. (2001c). The association between chronic medical conditions and work impairment. In *Caring and Doing for Others: Social Responsibility in the Domains of Family, Work, and Community* (ed, Rossi, A. S.), pp. 403-426. University of Chicago Press: Chicago, IL.

Lynch, W. & Riedel, J. E. (2001). *Measuring Employee Productivity: A Guide to Self-Asessment Tools.* The Institute for Health and Productivity Management: Scottsdale, AZ.

Neal, A., Griffen, M. A., Paterson, J. & Bordia, P. (2000). Development of measures of situational awareness, task performance, and contextual performance in air traffic control. In *Aviation Resource Management* (eds, Lowe, A. R. & Hayward, B. J.), pp. 241-267. Ashgate: Aldershot, UK.

Patrick, D. L. & Erickson, P. (1993). *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resourced Allocation.* Oxford Univerity Press: New York.

Pauly, M. V., Nicholson, S., Xu, J., Polsky, D., Danzon, P. M., Murray, J. F. & Berger, M. L. (2002a). A general model of the impact of absenteeism on employers and employees. *Health Economics* **11,** 221-231.

Pauly, M. V., Nicholson, S., Xu, J., Polsky, D., Danzon, P. M., Murray, J. F. & Berger, M. L. (2002b). A general model of the impact of absenteeism on employers and employees. *Health Econ* **11,** 221-231.

Reginster, J. Y. & Khaltaev, N. G. (2002). Introduction and WHO perspective on the global burden of musculoskeletal conditions. *Rheumatology* **41,** 1-2.

Rosenheck, R. A., Druss, B., Stolar, M., Leslie, D. & Sledge, W. (1999). Effect of declining mental health service use on employees of a large corporation. *Health Aff* **18,** 193-203.

Sawyer, J. E., Latham, W. R., Pritchard, R. D. & Benett, W. R. (1999). Analysis of work group productivity in an applied setting: Application of a time series panel design. *Personnel Psychology* **52,** 927-967.

Simon, G. E., Barber, C., Birnbaum, H. G., Frank, R. G., Greenberg, P. E., Rose, R. M., Wang, P. S. & Kessler, R. C. (2001). Depression and work productivity: the

comparative costs of treatment versus nontreatment. *Journal of Occupational and Environmental Medicine* **43,** 2-9.

Sloan, F. A. (ed.) (1996). *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies.* Cambridge University Press: New York.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology.* Jossey-Bass: San Francisco, CA.

Tarlov, A. R., Ware, J. E., Jr., Greenfield, S., Nelson, E. C., Perrin, E. & Zubkoff, M. (1989). The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association* **262,** 925-930.

Warr, P. B. & Conner, M. T. (1999). *The Measurement of Personal Effectiveness for Review and Guidance.* Department of Education and Employment: London.

Wells, K. B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., Miranda, J., Carney, M. F. & Rubenstein, L. V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association* **283,** 212-220.

Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist Instructions

The questions on the back page are designed to stimulate dialogue between you and your patients and to help confirm if they may be suffering from the symptoms of attention-deficit/hyperactivity disorder (ADHD).

Description: The Symptom Checklist is an instrument consisting of the eighteen DSM-IV-TR criteria. Six of the eighteen questions were found to be the most predictive of symptoms consistent with ADHD. These six questions are the basis for the ASRS v1.1 Screener and are also Part A of the Symptom Checklist. Part B of the Symptom Checklist contains the remaining twelve questions.

Instructions:

Symptoms

1. Ask the patient to complete both Part A and Part B of the Symptom Checklist by marking an X in the box that most closely represents the frequency of occurrence of each of the symptoms.

2. Score Part A. If four or more marks appear in the darkly shaded boxes within Part A then the patient has symptoms highly consistent with ADHD in adults and further investigation is warranted.

3. The frequency scores on Part B provide additional cues and can serve as further probes into the patient's symptoms. Pay particular attention to marks appearing in the dark shaded boxes. The frequency-based response is more sensitive with certain questions. No total score or diagnostic likelihood is utilized for the twelve questions. It has been found that the six questions in Part A are the most predictive of the disorder and are best for use as a screening instrument.

Impairments

1. Review the entire Symptom Checklist with your patients and evaluate the level of impairment associated with the symptom.

2. Consider work/school, social and family settings.

3. Symptom frequency is often associated with symptom severity, therefore the Symptom Checklist may also aid in the assessment of impairments. If your patients have frequent symptoms, you may want to ask them to describe how these problems have affected the ability to work, take care of things at home, or get along with other people such as their spouse/significant other.

History

1. Assess the presence of these symptoms or similar symptoms in childhood. Adults who have ADHD need not have been formally diagnosed in childhood. In evaluating a patient's history, look for evidence of early-appearing and long-standing problems with attention or self-control. Some significant symptoms should have been present in childhood, but full symptomology is not necessary.

Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist

Patient Name		Today's Date	

Please answer the questions below, rating yourself on each of the criteria shown using the scale on the right side of the page. As you answer each question, place an X in the box that best describes how you have felt and conducted yourself over the past 6 months. Please give this completed checklist to your healthcare professional to discuss during today's appointment.	Never	Rarely	Sometimes	Often	Very Often
1. How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done?					
2. How often do you have difficulty getting things in order when you have to do a task that requires organization?					
3. How often do you have problems remembering appointments or obligations?					
4. When you have a task that requires a lot of thought, how often do you avoid or delay getting started?					
5. How often do you fidget or squirm with your hands or feet when you have to sit down for a long time?					
6. How often do you feel overly active and compelled to do things, like you were driven by a motor?					
					Part A
7. How often do you make careless mistakes when you have to work on a boring or difficult project?					
8. How often do you have difficulty keeping your attention when you are doing boring or repetitive work?					
9. How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly?					
10. How often do you misplace or have difficulty finding things at home or at work?					
11. How often are you distracted by activity or noise around you?					
12. How often do you leave your seat in meetings or other situations in which you are expected to remain seated?					
13. How often do you feel restless or fidgety?					
14. How often do you have difficulty unwinding and relaxing when you have time to yourself?					
15. How often do you find yourself talking too much when you are in social situations?					
16. When you're in a conversation, how often do you find yourself finishing the sentences of the people you are talking to, before they can finish them themselves?					
17. How often do you have difficulty waiting your turn in situations when turn taking is required?					
18. How often do you interrupt others when they are busy?					

The Value of Screening for Adults With ADHD

Research suggests that the symptoms of ADHD can persist into adulthood, having a significant impact on the relationships, careers, and even the personal safety of your patients who may suffer from it.[1-4] Because this disorder is often misunderstood, many people who have it do not receive appropriate treatment and, as a result, may never reach their full potential. Part of the problem is that it can be difficult to diagnose, particularly in adults.

The Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist was developed in conjunction with the World Health Organization (WHO), and the Workgroup on Adult ADHD that included the following team of psychiatrists and researchers:

- **Lenard Adler, MD**
  Associate Professor of Psychiatry and Neurology
  New York University Medical School

- **Ronald C. Kessler, PhD**
  Professor, Department of Health Care Policy
  Harvard Medical School

- **Thomas Spencer, MD**
  Associate Professor of Psychiatry
  Harvard Medical School

As a healthcare professional, you can use the ASRS v1.1 as a tool to help screen for ADHD in adult patients. Insights gained through this screening may suggest the need for a more in-depth clinician interview. The questions in the ASRS v1.1 are consistent with DSM-IV criteria and address the manifestations of ADHD symptoms in adults. Content of the questionnaire also reflects the importance that DSM-IV places on symptoms, impairments, and history for a correct diagnosis.[4]

The checklist takes about 5 minutes to complete and can provide information that is critical to supplement the diagnostic process.

References:

1. Schweitzer JB, et al. Med Clin North Am. 2001;85(3):10-11, 757-777.
2. Barkley RA. Attention Deficit Hyperactivity Disorder: A Handbook for Diagnosis and Treatment. 2nd ed. 1998.
3. Biederman J, et al. Am J Psychiatry.1993;150:1792-1798.
4. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision. Washington, DC, American Psychiatric Association. 2000: 85-93.

CME Available for this Article at ACOEM.org

The Prevalence and Effects of Adult Attention Deficit/Hyperactivity Disorder on Work Performance in a Nationally Representative Sample of Workers

Ronald C. Kessler, PhD
Lenard Adler, MD
Minnie Ames, PhD
Russell A. Barkley, PhD
Howard Birnbaum, PhD
Paul Greenberg, MA
Joseph A. Johnston, MD, MSc
Thomas Spencer, MD
T. Bedirhan Üstün, MD

Learning Objectives

Recall the estimated prevalence of attention-deficit/hyperactivity disorder (ADHD) in this national household survey of 3198 workers 18 to 44 years of age, and any demographic factors associated with a relatively high risk of this disorder.

Point out the impact of ADHD on lost work performance, and any association between type of work and the risk of this disorder.

Summarize how often affected workers had sought professional treatment in the past year for manifestations of ADHD and for other emotional problems.

Abstract

Objective: *The prevalence and workplace consequences of adult attention deficit/hyperactivity disorder (ADHD) are unknown.* ***Methods:*** *An ADHD screen was included in a national household survey (n = 3198, ages 18–44). Clinical reinterviews calibrated the screen to diagnoses of Diagnostic and Statistical Manual of Mental Disorders, 4th edition ADHD. Diagnoses among workers were compared with responses to the WHO Health and Work Performance Questionnaire (HPQ).* ***Results:*** *A total of 4.2% of workers had ADHD. ADHD was associated with 35.0 days of annual lost work performance, with higher associations among blue collar (55.8 days) than professional (12.2 days), technical (19.8 days), or service (32.6 days) workers. These associations represent 120 million days of annual lost work in the U.S. labor force, equivalent to $19.5 billion lost human capital.* ***Conclusions:*** *ADHD is a common and costly workplace condition. Effectiveness trials are needed to estimate the region of interest of workplace ADHD screening and treatment programs.* (J Occup Environ Med. *2005;47:565–572*)

From the Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts (Drs Kessler, Ames); the Departments of Psychiatry and Neurology, New York University School of Medicine and Psychiatry Service, New York Harbor Veterans Administration Medical Center, New York, NY (Dr Adler); the University of Massachusetts Medical School, Boston, Massachusetts (Dr Barkley); the Analysis Group–Economics (Dr Birnbaum, Mr Greenberg); Eli Lilly and Company, US Health Outcomes (Dr Johnson); the Department of Psychiatry, Massachusetts General Hospital (Dr Spencer); and Classifications, World Health Organization (Dr Üstün).

Ronald C. Kessler received support for this research from Eli Lilly and Company.

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.

Copyright © by American College of Occupational and Environmental Medicine

DOI: 10.1097/01.jom.0000166863.33541.39

Although it has long been suspected that attention deficit/hyperactivity disorder (ADHD) often persists into adulthood,[1,2] adult ADHD has only recently become the focus of clinical attention.[3,4] As a result of this neglect, little is known about the prevalence or correlates of adult ADHD. Attempts to estimate prevalence by extrapolating from childhood prevalence estimates linked with adult persistence estimates[5–8] or by directly estimating prevalence in small samples of adults[9] or college students[10] have yielded prevalence estimates in the range of 1% to 6%. These studies suggest that adult ADHD is often seriously impairing but less prevalent among employed people than the rest of the population. However, nationally representative surveys of the prevalence and consequences of ADHD for work performance have never before been undertaken.

In an effort to obtain estimates of adult ADHD prevalence and correlates, a screen for ADHD among adults was included in the National Comorbidity Survey Replication (NCS-R),[11] a recently completed face-to-face, nationally representative household survey of the U.S. adult population. Clinical reappraisal interviews to evaluate the presence of ADHD among adult respondents were then carried out in a subsample of NCS-R respondents that oversampled screened positives to assess the validity of the ADHD screen. Extrapolation to the population from

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

the patterns in the reappraisal interviews led to an estimate that 4.4% of the U.S. adult population in the age range of the screen meet criteria for current ADHD,[12] one of the highest point prevalence estimates of any mental disorder.[13]

The current report extends these results to consider the prevalence and workplace consequences of ADHD among NCS-R respondents in the active labor force. Because the World Health Organization's (WHO) Health and Work Performance Questionnaire (HPQ)[14,15] was included in the NCS-R, statistical analysis linking clinical diagnoses of ADHD with HPQ results is used to examine the associations of the disorder with work performance in the total sample as well as in major occupational categories. Past research on this topic, which was based on the analysis of medical claims data for a large corporation, showed that patients in treatment for adult ADHD had higher rates of work loss[16] and accidents.[17] We extend these earlier results in several ways: to a nationally representative sample of employed people rather than to workers in only one corporation; to all workers with ADHD rather than only to those in treatment; and to consider associations of ADHD with work performance while on the job as well as with sickness absence.

Materials and Methods

Sample

As detailed elsewhere,[18] the NCS-R is a nationally representative survey of mental disorders among English-speaking household residents ages 18 and older in the coterminous United States. Face-to-face interviews were carried out with 9282 respondents between February 2001 and April 2003. The response rate was 70.9%. Sample recruitment began by mailing an advance letter and study fact brochure to each preselected sample household. This was followed by an in-person interviewer visit to answer questions before obtaining informed consent and scheduling an interview. Respondents were given a $50 incentive to participate in the survey. In addition, a probability subsample of hard-to-reach and initially reluctant predesignated respondents was selected to participate in a brief telephone nonrespondent survey in an effort to obtain responses to marker questions that were used to weight the main sample for nonresponse bias. Participants in the nonrespondent survey were given a $100 incentive. The human subjects committees of Harvard Medical School and the University of Michigan both approved these recruitment and consent procedures.

The NCS-R interview was administered in two parts. Part I included a core psychiatric diagnostic assessment that was administered to all respondents (n = 9282), whereas part II included questions about correlates and additional disorders that was administered to all part I respondents who met lifetime criteria for any core disorder plus a roughly one-in-three probability subsample of other respondents (n = 5692). ADHD was included in the part II assessment. Because ADHD is a childhood disorder, which requires age of onset before age 8, respondents were required to provide retrospective reports about childhood symptoms before reporting whether these symptoms continued to persist in adulthood. Based on concern about the accuracy of recall about childhood symptoms among older adults, ADHD was assessed only among part II respondents in the age range 18 to 44 (n = 3198). These respondents were weighted to be representative of the entire U.S. population in this age range on a profile of 2000 Census sociodemographic and geographic variables. A subsample of 2399 of these 3198 respondents was either employed or self-employed in the month before the interview and had valid data on all the measures used in the analyses presented here. This is the sample considered in the current report. A more detailed discussion about NCS-R sampling and weighting is presented elsewhere.[18]

The NCS-R respondents assessed for ADHD were divided into four sampling strata for purposes of selecting a clinical reappraisal subsample: those who reported no symptoms of childhood ADHD; those who reported some symptoms, but did not meet full criteria for childhood ADHD; those who were classified as childhood cases, but who denied any adult symptoms; and those who were both classified as childhood cases and who also reported adult symptoms. An attempt was made to contact by telephone and administer a semistructured adult ADHD clinical interview to 30 respondents in each of the first three strata and 60 in the fourth stratum. The final quota sample included 154 respondents (slightly more than the target because more predesignated respondents kept their appointments to be interviewed than expected). These cases were weighted to be representative of the U.S. population in the age range of the ADHD subsample. A more detailed description of the ADHD clinical reappraisal sample design is reported elsewhere.[19]

Adult Attention Deficit/Hyperactivity Disorder

The retrospective assessment of childhood ADHD in the NCS-R was based on the Diagnostic Interview Schedule for the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV).[20] Respondents classified retrospectively as having had ADHD symptoms in childhood were then asked a single question about whether they continued to have any current problems with attention or hyperactivity–impulsivity. The clinical reappraisal interviews obtained much more detailed information based on responses to the open-ended questions in the Adult ADHD Clinical Diagnostic Scale (ACDS) version

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

1.2.[21,22] The ACDS is a semistructured interview that includes the ADHD Rating Scale (ADHD-RS)[23] to assess childhood ADHD and an adaptation of the ADHD-RS to assess current ADHD among adult respondents. The ACDS is widely used in clinical trials of adult ADHD.[24,25]

Four experienced clinical interviewers (all PhD clinical psychologists) carried out the clinical reappraisal interviews. Each interviewer received 40 hours of training from two board-certified psychiatrists who specialize in adult ADHD (LA, TS) and successfully completed five practice interviews. All clinical interviews were tape-recorded and reviewed by a clinical supervisor. Weekly group interviewer calibration meetings and one-on-one feedback meetings were used to prevent drift. A clinical diagnosis of adult ADHD required six symptoms of either inattention or hyperactivity–impulsivity during the 6 months before the interview (DSM-IV criterion A), at least two criterion A symptoms before age 7 (criterion B), some impairment in at least two areas of living during the past 6 months (criterion C), and clinically significant impairment in at least one of these areas (criterion D). Because no attempt was made to operationalize DSM-IV diagnostic hierarchy rules (criterion E), the clinical classification was an assessment of the syndrome rather than the diagnosis of adult ADHD.

Other Correlates of Adult Attention Deficit/Hyperactivity Disorder

We examined associations of adult ADHD with sociodemographic variables and with measures of work performance assessed in the WHO Health and Work Performance Questionnaire.[11,14,18] The sociodemographic variables included age (18–24, 25–34, 35–44), sex, race–ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, other), education (less than high school, high school graduate, some college, college graduate), and broad occupational category (professional, white collar technical, service, blue collar). The HPQ used self-reports about absenteeism (missed days of work) and presenteeism (low performance while at work transformed to lost workday-equivalents) to generate a summary measure of overall lost workdays in the month before the interview. Information about salary was used to transform the measures of lost work performance from a time metric to a salary metric for purposes of estimating human capital costs associated with ADHD. Salary was incremented by 25% as an estimate of fringe benefits.

Analysis Methods

The method of multiple imputation (MI)[26] was used to assign predicted probabilities of adult ADHD clinical diagnoses to respondents in the main sample based on the associations between the ADHD sampling strata and the clinical diagnoses in the clinical reappraisal subsample. MI is a simulation method that adjusts significance tests when one or more variables are based on imputation rather than on direct measurement. A more detailed description of MI is presented elsewhere.[12] Because a strong monotonic relationship was found between clinical reappraisal sampling strata and blind clinical diagnoses of adult ADHD in the weighted clinical reappraisal subsample, with an area under the receiver operator characteristic curve (AUC) of 0.86, the precision of the MI estimates is quite good. This means that the standard errors of estimates will not be markedly greater based on the use of MI than if clinical interviews for ADHD had been carried out in the complete sample.

Simple subgroup comparison of prevalence was used to study sociodemographic correlates of ADHD, whereas linear regression analysis was used to estimate the associations of ADHD with work performance. ADHD was coded as a yes–no dummy predictor variable in the linear regression analysis, whereas sociodemographic variables (age, sex, race–ethnicity, education, occupation) were included as controls. The dependent variables in the two main linear regression equations were composite measures of overall lost work performance in the metrics of day-equivalents and salary plus fringe benefits. We also disaggregated both regression equations to examine separate effects of ADHD on absenteeism and presenteeism. The individual-level estimates were annualized by multiplying the unstandardized linear regression coefficients (ie, days or dollars of lost performance associated with ADHD in the past month) by 12 (ie, number of months in a year), whereas the individual-level estimates were projected to the total U.S. civilian labor force by adjusting for the prevalence estimate and for the fact that the seasonally adjusted number of workers in the U.S. civilian labor force in the age range of the ADHD sample at the time the NCS-R was fielded was 82.3 million people. More complex linear regression equations with interactions between ADHD and broadly defined occupational category were used to evaluate whether the associations of ADHD with the outcomes differ significantly by occupational category.

Because the NCS-R sample design features weighting and clustering of observations, significance tests based on the assumption of simple random sampling are inaccurate. This problem was corrected by using a design-based method to evaluate statistical significance. The Taylor series linearization method[27] was used for this purpose as implemented in the SUDAAN software system.[28] The MI adjustments of significance tests were based on Taylor series estimates of coefficient standard errors. Significance tests of set of coefficients in the logistic regression equations were made using Wald chi-squared tests based on design-corrected MI coefficient variance–covariance matrices.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

Results

Prevalence and Sociodemographic Correlates

The MI estimate of current ADHD prevalence (standard error in parentheses) among employed NCS-R respondents is 4.2% (0.7) (Table 1). Prevalence does not differ significantly by respondent sex, age, education, or occupation, but is significantly higher among non-Hispanic whites (5.2) than those in other race–ethnic groups (0.7–2.5).

Effects of Attention Deficit/Hyperactivity Disorder on Work Performance

The MI estimate of ADHD is a significant predictor of overall lost work performance in the regression analysis, with annualized individual-level regression slopes equivalent to 35.0 days and $5661 of salary-equivalent lost performance per worker with ADHD per year (Table 2). Disaggregation shows that absenteeism (13.6 days and $2351) and presenteeism (21.6 days and $3404) both play important parts in the aggregate regression coefficient. Projections of these individual-level associations to the total U.S. civilian labor force yield estimates that 120.8 million lost days of work per year and $19.6 billion salary-equivalent per year are associated with ADHD.

TABLE 1
The Associations of Sociodemographics With Multiply Imputed Estimates of Current DSM-IV ADHD Among Employed NCS-R Respondents Ages 18–44 (n = 2399)

	Prevalence						
	Percent	(SE)	OR	(95% CI)	χ^2	df	*P*
Total	4.2	(0.7)	—	—	—	—	—
Sex					0.7	1	0.4
Male	4.9	(0.9)	1.0	—			
Female	3.3	(0.8)	0.8	(0.4–1.4)			
Age					1.7	1	0.2
18–29	3.4	(0.8)	1.0	(0.8–2.5)			
30–44	4.7	(1.0)	1.4	—			
Race					16.7	3	0.0
Non-Hispanic white	5.2	(0.9)	1.0	—			
Non-Hispanic black	1.4	(0.5)	0.2	(0.1–0.6)			
Hispanic	2.5	(1.1)	0.5	(0.2–1.1)			
Other	0.7	(0.8)	0.1	(0.0–0.8)			
Education					1.4	3	0.7
Less than high school	4.2	(1.3)	1.3	(0.4–4.1)			
High school graduate	4.3	(1.3)	1.3	(0.4–4.2)			
Some college	4.9	(1.1)	1.6	(0.6–4.4)			
College graduate	3.0	(1.2)	1.0	—			
Occupation					5.2	3	0.2
Professional	3.1	(1.1)	1.0	—			
White collar technical	6.5	(1.7)	2.1	(0.8–6.0)			
Service	3.0	(0.8)	1.0	(0.4–3.0)			
Blue collar	4.2	(1.5)	1.7	(0.8–3.9)			

DSM-IV indicates Diagnostic and Statistical Manual of Mental Disorders, 4th edition; ADHD, attention deficit/hyperactivity disorder; NCS-R, National Comorbidity Survey Replication; SE, standard error; OR, odds ratio; CI, confidence interval.

TABLE 2
The Estimated Annual Effects of DSM-IV ADHD on Total Lost Work Performance and Its Components (Absenteeism, Presenteeism) Among Employed NCS-R Respondents Ages 18–44 (n = 2399)

	Individual Level				National Projections (in millions)			
	Days	(SE)	$	(SE)	Days	(SE)	$	(SE)
Total	35.0	(14.1)	5661	(5661)	120.8	(48.9)	19,561	(11,974)
Absenteeism	13.6	(7.5)	2351	(1560)	47.1	(25.9)	8122	(5391)
Presenteeism	21.6	(11.3)	3404	(3384)	74.6	(38.9)	11,762	(11,694)

DSM-IV indicates Diagnostic and Statistical Manual of Mental Disorders, 4th edition; ADHD, attention deficit/hyperactivity disorder; NCS-R, National Comorbidity Survey Replication; SE, standard error.

Variation in Effects Across Occupational Categories

The linear regression equations were elaborated to test the possibility that the associations of ADHD with decrements in work performance vary significantly by broad occupational categories. Overall estimated individual-level associations were found to be largest among blue collar workers and smallest among professionals (Table 3), although these differences were not statistically significant. Individual-level associations among blue collar workers were 55.8 days and $6410 salary-equivalent lost performance per worker with ADHD per year.

Treatment

NCS-R respondents who screened positive for current ADHD were asked whether they received any professional treatment for their problems with concentration, inattention, or impulsivity at any time in the 12 months before the interview. Only 16.4% of clinically confirmed cases (4.6% standard error) reported that they received any such treatment. All part II NCS-R respondents were also queried in a separate section of the interview about whether they received treatment at any time in the past 12 months for any problem with their emotions, nerves, or mental health. As one might expect in light of the high comorbidity known to

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

TABLE 3
Variation in the Estimated Annual Effects of DSM-IV ADHD on Total Lost Work Performance and Its Components (Absenteeism, Presenteeism) by Occupational Category Among Employed NCS-R Respondents Ages 18–44 Who Are Either Professionals (n = 609), White Collar Technical Workers (n = 301), Service Workers (n = 796), or Blue Collar Workers (n = 693)

	Individual Level				National Projections (in millions)			
	Days	(SE)	$	(SE)	Days	(SE)	$	(SE)
Total lost work performance								
Professional	12.2	(44.5)	7198	(7613)	7.3	(26.6)	4295	(4542)
White collar technical	19.8	(25.7)	5981	(4858)	12.6	(16.3)	3811	(3096)
Service	32.6	(21.7)	2168	(6600)	25.9	(17.3)	1724	(5248)
Blue collar	55.8	(31.5)	6410	(7038)	77.0	(43.5)	8853	(9721)
χ^2_3	0.8		0.3					
P	0.839		0.957					
Absenteeism								
Professional	3.4	(14.0)	−66	(1808)	2.0	(8.3)	−40	(1079)
White collar technical	7.1	(11.0)	1195	(2352)	4.5	(7.0)	761	(1499)
Service	18.2	(15.2)	5011	(4081)	14.4	(12.1)	3984	(3245)
Blue collar	21.6	(16.8)	2517	(3115)	29.8	(23.2)	3476	(4302)
χ^2_3	0.8		1.7					
P	0.844		0.631					
Presenteeism								
Professional	8.3	(35.1)	7210	(6935)	5.0	(20.9)	4302	(4138)
White collar technical	16.1	(22.1)	4719	(5248)	10.2	(14.1)	3007	(3344)
Service	13.4	(15.4)	−2972	(8229)	10.6	(12.3)	−2363	(6543)
Blue collar	35.7	(23.2)	4389	(4997)	49.3	(32.1)	6061	(6901)
χ^2_3	0.9		1.0					
P	0.830		0.803					

DSM-IV indicates Diagnostic and Statistical Manual of Mental Disorders, 4th edition; ADHD, attention deficit/hyperactivity disorder; NCS-R, National Comorbidity Survey Replication; SE, standard error.

exist between ADHD with other adult mental and substance disorders,[12] a substantial proportion of workers with ADHD reported that they received 12-month treatment for other emotional problems (32.1%). The difference between 32.1% and 16.4% (ie, 15.7%) represents the proportion of workers with ADHD who received 12-month treatment for an emotional problem but without ADHD being a focus of that treatment.

Discussion

Several limitations of this analysis are noteworthy. The most important of them is that the DSM-IV criteria for ADHD were developed with children in mind and offer only limited guidance regarding the diagnosis of ADHD among adults. This is of considerable concern because clinical studies make it clear that symptoms of ADHD are more heterogeneous and subtle in adults than children,[29,30] leading some clinical researchers to suggest that assessment of adult ADHD might require an increase in the variety of symptoms assessed,[31] a reduction in the severity threshold for diagnosis,[32] or a reduction in the DSM-IV six-of-nine symptom requirement.[33] To the extent that such changes would lead to a more valid assessment than in the clinical interviews carried out in the current study, our prevalence estimate is conservative.

A second limitation is that the clinical assessment of ADHD was carried out only in the reappraisal interview subsample and imputed in the larger sample. Although the imputation equation was quite strong (AUC = 0.86), the use of imputation led to an increase in the standard error of the estimates when we used the MI method. Statistical power to detect meaningful associations was consequently reduced. This might explain our failure to document statistically significant differences in the impact of ADHD across types of occupation despite the substantively higher estimated days of lost productivity associated with ADHD among blue collar respondents. Because of this problem with precision, it would be useful to replicate and extend the results reported here regarding differences in associations across occupations in an independent and larger sample. Several replications of this sort are currently underway in a number of workplace HPQ surveys.

An additional limitation concerns the fact that the clinical assessment was based entirely on respondent self-report. It is noteworthy in this regard that childhood ADHD is diagnosed largely based on parent and teacher reports rather than on patient self-reports as a result of the fact that children with ADHD are notoriously

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

unaware of their symptoms.[34] Informant assessment is much more difficult for adults, however, making it necessary to base assessment largely on self-report.[30] Methodological studies comparing adult self-reports versus informant reports of adult ADHD symptoms document the same general pattern of underestimation in adult self-reports as in child self-reports.[35,36] This suggests that the prevalence estimates reported here are probably conservative, although the only study of adult self versus informant assessment of adult ADHD in a nonclinical sample found fairly strong associations between the two reports and no self versus informant difference in reported symptom severity.[37]

Within the context of these limitations, the results reported here document that adult ADHD is a commonly occurring disorder in the U.S. civilian labor force that is associated with substantial lost work performance, especially among blue collar workers. The NCS-R prevalence estimate of 4.2% is likely to be conservative based on the limitations described in the last few paragraphs. The prevalence estimate is within the range of estimates obtained in less comprehensive studies.[5–10] The finding that prevalence is higher among non-Hispanic whites than other racial–ethnic groups was unexpected in that studies of children and adolescents do not find race–ethnic differences in ADHD.[38] The same pattern was found in the full NCS-R sample, however, arguing against the otherwise plausible possibility that ADHD creates a higher barrier to employment among minorities than nonminorities.[12] The finding that adult ADHD is not related to age in the range considered here (ie, 18–29 vs. 30–44) extends the broader finding that the disorder does not spontaneously remit with age after early adulthood. The failure to find a sex difference in prevalence is also noteworthy because ADHD is known to be significantly more common among males than females in the total population both among children[39] and among adults.[12] More detailed analysis (results not shown, but available on request) shows that our failure to find a comparable sex difference among employed people in the NCS-R is the result of ADHD being significantly associated with unemployment among men but not women. This difference could be the result of either a sex difference in work requirements that causes work termination resulting from ADHD to occur more often among men than women or a sex difference in other consequences of adult ADHD (eg, secondary alcohol or drug abuse) that lead to unemployment. We have no way of distinguishing between these possibilities in the NCS-R data.

The finding that adult ADHD is associated with significant decrements in work performance is consistent both with clinical observations of role impairment[22] and with neuropsychologic evidence of impaired performance[40] among adult patients with ADHD. The estimated association of ADHD with absenteeism is quite large in relation to comparable estimates for other chronic physical and mental disorders reported in the literature.[41–43] It is interesting that we found a trend for impairment to be greatest among blue collar workers. One might have expected that professionals would be most affected, because the inattentive subtype of ADHD is more common than the impulsive subtype among adults.[44] One possible explanation is that selection into occupation based on childhood ADHD leads to the most severe cases being blue collar workers. Another possibility is that the less flexible job demands of blue collar than white collar workers results in ADHD being associated with especially high decrements in work performance. Replication of this specification in independent surveys is needed, however, before alternative interpretations should be examined empirically.

The most important question for future research raised by this study is whether effective outreach and treatment would improve the performance of workers with ADHD sufficiently to have a positive return on investment for employers. This question can be divided into two parts. The first part is whether the estimated associations of ADHD with work performance are accurate reflections of the causal effect of ADHD rather than common causes or mediators that will not resolve with the successful treatment of ADHD. The second part is whether these effects, assuming they are causal and reversible, would in fact be reversed by treatment.

Regarding the first part of this question, the estimated associations of ADHD with work performance reported here could be biased upward to the extent that unmeasured common causes of ADHD and work performance account for the observed associations. Because ADHD starts very early in life, however, any such common causes would have to be in existence since childhood. A more plausible possibility, then, is that ADHD has causal effects on work performance that are mediated by more proximate causes that would not remit with the treatment of ADHD although their onset was significantly influenced by ADHD. Such mediators could be behavioral (eg, alcohol abuse), psychologic (eg, low commitment to work), or some combination of the two. Comorbid mental disorders could be among the mediators. It is impossible to evaluate these effects nonexperimentally in the NCS, although we have measures of these comorbid disorders, because we have no way of knowing which of these presumed mediators would be affected by successful treatment of ADHD. However, the fact that experimental intervention studies document that successful treatment of ADHD among adults is associated with substantial gains in neuropsychologic task performance and cognition[45–47] argues that at

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

least some part of the associations documented here are likely to be the result of the reversible effects of ADHD on work performance, possibly mediated through effects of ADHD on other comorbid conditions that are positively affected by successful treatment of ADHD.

Regarding the second part of the question, whether effective outreach and treatment would improve the performance of workers with ADHD sufficiently to have a positive return on investment for employers, the evidence from experimental treatment trials cited at the end of the last paragraph leads to the expectation that presenteeism, and perhaps also absenteeism, would decrease as a result of the successful treatment of workers with ADHD. However, there is no way to know definitively if these decreases in indirect workplace costs would exceed the direct costs of treatment in the absence of an experimental effectiveness study. It seems plausible that this would be the case, however, based on the fact that indirect costs are as high as they are estimated to be here (ie, more than $6000 per year for technical and blue collar workers and more than $8000 per year for professionals). The fact that these costs are high means that treatment would be cost-effective from the employer perspective even if it only resulted in a 15% reduction in the work performance decrement. Indeed, given that separate research has documented effects of adult ADHD on workplace accidents,[17] an outcome not considered in the current report, an improvement in work performance even less than 15% might have a positive region of interest (ROI) if it helped reduce accidents associated with workers' compensation claims. Given that treatment effects considerably larger than 15% seem plausible based on the available evidence from treatment efficacy trials, the next logical step is to carry out effectiveness trials to evaluate the ROI of workplace screening and best-practices treatment of ADHD. We are currently in the process of designing such a trial.

Acknowledgments

The National Comorbidity Survey Replication (NCS-R) is supported by the National Institute of Mental Health (NIMH; U01-MH60220) with supplemental support from the National Institute of Drug Abuse, the Substance Abuse and Mental Health Services Administration, the Robert Wood Johnson Foundation (grant 044780), and the John W. Alden Trust. Additional support for preparation of this paper was provided by the Eli Lilly Company. Collaborating NCS-R investigators include Ronald C. Kessler (Principal Investigator, Harvard Medical School), Kathleen Merikangas (Co-Principal Investigator, NIMH), James Anthony (Michigan State University), William Eaton (The Johns Hopkins University), Meyer Glantz (NIDA), Doreen Koretz (Harvard University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Greg Simon (Group Health Cooperative), T. Bedirhan Ustun (World Health Organization), Michael Von Korff (Group Health Cooperative), Philip Wang (Harvard Medical School), Kenneth Wells (UCLA), Elaine Wethington (Cornell University), and Hans-Ulrich Wittchen (Max Planck Institute of Psychiatry). The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or the U.S. Government. A complete list of NCS publications and the full text of all NCS-R instruments can be found at http://www.hcp.med.harvard.edu/ncs. Send correspondence to NCS@hcp.med.harvard.edu. The NCS-R is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. The authors thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. These activities were supported by the National Institute of Mental Health (R01 MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the U.S. Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Pan American Health Organization, Eli Lilly and Company, and GlaxoSmithKline. A complete list of WMH publications and instruments can be found at http://www.hcp.med.harvard.edu/wmh.

References

1. Mannuzza S, Klein RG, Bessler A, et al. Adult outcome of hyperactive boys. Educational achievement, occupational rank, and psychiatric status. *Arch Gen Psychiatry*. 1993;50:565–576.
2. Weiss G, Hechtman L. *Hyperactive Children Grown Up: ADHD in Children, Adolescents, and Adults.* New York: Guilford Press; 1993.
3. Pary R, Lewis S, Matuschka PR, et al. Attention deficit disorder in adults. *Ann Clin Psychiatry*. 2002;14:105–111.
4. Wilens TE, Biederman J, Spencer TJ. Attention deficit/hyperactivity disorder across the lifespan. *Annu Rev Med*. 2002; 53:113–131.
5. Barkley RA, Fischer M, Smallish L, et al. The persistence of attention-deficit/hyperactivity disorder into young adulthood as a function of reporting source and definition of disorder. *J Abnorm Psychol*. 2002;111:279–289.
6. Biederman J, Mick E, Faraone SV. Age-dependent decline of symptoms of attention deficit hyperactivity disorder: impact of remission definition and symptom type. *Am J Psychiatry*. 2000;157:816–818.
7. Mannuzza S, Klein RG, Bessler A, et al. Adult psychiatric status of hyperactive boys grown up. *Am J Psychiatry*. 1998; 155:493–498.
8. Weiss G, Hechtman L, Milroy T, et al. Psychiatric status of hyperactives as adults: a controlled prospective 15-year follow-up of 63 hyperactive children. *J Am Acad Child Psychiatry*. 1985;24: 211–220.
9. Murphy K, Barkley RA. Attention deficit hyperactivity disorder adults: comorbidities and adaptive impairments. *Compr Psychiatry*. 1996;37:393–401.
10. Heiligenstein E, Conyers LM, Berns AR, et al. Preliminary normative data on DSM-IV attention deficit hyperactivity disorder in college students. *J Am Coll Health*. 1998;46:185–188.
11. Kessler RC, Merikangas KR. The National Comorbidity Survey Replication (NCS-R): background and aims. *Int J Methods Psychiatr Res*. 2004;13:60–68.
12. Kessler RC, Adler L, Barkley R, et al. The prevalence and correlates of adult ADHD in the United States: Results from the National Comorbidity Survey Replication. *Am J Psychiatry.* In press.
13. Kessler RC, Chiu WT, Demler O, et al. Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the National Comorbidity Survey Replication (NCS-R). *Arch Gen Psychiatry.* In press.
14. Kessler RC, Barber C, Beck A, et al. The World Health Organization Health and Work Performance Questionnaire (HPQ).

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

J Occup Environ Med. 2003;45:156–174.
15. Kessler RC, Ames M, Hymel PA, et al. Using the WHO Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. *J Occup Environ Med*. 2004;46:s23–s27.
16. Swensen AR, Birnbaum HG, Secnik K, et al. Attention-deficit/hyperactivity disorder: increased costs for patients and their families. *J Am Acad Child Adolesc Psychiatry*. 2003;42:1415–1423.
17. Swensen A, Birnbaum HG, Ben Hamadi R, et al. Incidence and costs of accidents among attention-deficit/hyperactivity disorder patients. *J Adolesc Health*. 2004; 35:346–347.
18. Kessler RC, Berglund P, Chiu WT, et al. The US National Comorbidity Survey Replication (NCS-R): design and field procedures. *Int J Methods Psychiatr Res*. 2004;13:69–92.
19. Kessler RC, Adler L, Ames M, et al. The World Health Organization Adult ADHD Self-Report Scale (ASRS). *Psychol Med*. 2005;35:245–256.
20. Robins LN, Cottler L, Bucholz K, et al. *Diagnostic Interview Schedule for DSM-IV*. St. Louis: Washington University; 1995.
21. Adler L, Spencer T. *The Adult ADHD Clinical Diagnostic Scale (ACDS) V 1.2*. New York: New York University School of Medicine; 2004.
22. Adler L, Cohen J. Diagnosis and evaluation of adults with attention-deficit/hyperactivity disorder. *Psychiatr Clin North Am*. 2004;27:187–201.
23. DuPaul GJ, Power TJ, Anastopoulos AD, et al. *ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretation.* New York: Guilford Press; 1998.
24. Michelson D, Adler L, Spencer T, et al. Atomoxetine in adults with ADHD: two randomized, placebo-controlled studies. *Biol Psychiatry*. 2003;53:112–120.
25. Spencer T, Biederman J, Wilens T, et al. Efficacy of a mixed amphetamine salts compound in adults with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry*. 2001;58:775–782.
26. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons; 1987.
27. Wolter KM. *Introduction to Variance Estimation.* New York: Springer-Verlag; 1985.
28. Research Triangle Institute. *SUDAAN: Professional Software for Survey Data Analysis* [computer program]. Research Triangle Park, NC: Research Triangle Institute; 2002.
29. DeQuiros GB, Kinsbourne M. Adult ADHD: analysis of self-ratings in a behavior questionnaire. *Ann N Y Acad Sci*. 2001;931:140–147.
30. Wender PH, Wolf LE, Wasserstein J. Adults with ADHD. An overview. *Ann N Y Acad Sci.* 2001;931:1–16.
31. Barkley RA. ADHD behavior checklist for adults. *The ADHD Report*. 1995;3:16.
32. Ratey J, Greenberg S, Bemporad JR, et al. Unrecognized attention-deficit hyperactivity disorder in adults presenting for outpatient psychotherapy. *J Child Adolesc Psychopharmacol.* 1992;4:267–275.
33. McBurnett K. Attention-deficit/hyperactivity disorder: a review of diagnostic issues. In: Widiger TA, Francis AJ, Pincus HA, et al., eds. *DSM-IV Sourcebook.* Washington, DC: American Psychiatric Association; 1997:111–143.
34. Jensen PS, Rubio-Stipec M, Canino G, et al. Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *J Am Acad Child Adolesc Psychiatry*. 1999;38: 1569–1579.
35. Gittelman R, Mannuzza S. Diagnosing ADD-H in adolescents. *Psychopharmacol Bull*. 1985;21:237–242.
36. Zucker M, Morris MK, Ingram SM, et al. Concordance of self- and informant ratings of adults' current and childhood attention-deficit/hyperactivity disorder symptoms. *Psychol Assess*. 2002;14: 379–389.
37. Murphy P, Schachar R. Use of self-ratings in the assessment of symptoms of attention deficit hyperactivity disorder in adults. *Am J Psychiatry*. 2000;157:1156–1159.
38. Lahey BB, Applegate B, McBurnett K, et al. DSM-IV field trials for attention deficit hyperactivity disorder in children and adolescents. *Am J Psychiatry*. 1994;151: 1673–1685.
39. Scahill L, Schwab-Stone M. Epidemiology of ADHD in school-age children. *Child Adolesc Psychiatr Clin N Am.* 2000;9:541–555, vii.
40. Hervey AS, Epstein JN, Curry JF. Neuropsychology of adults with attention-deficit/hyperactivity disorder: a meta-analytic review. *Neuropsychology*. 2004;18:485–503.
41. Wang PS, Beck A, Berglund P, et al. Chronic medical conditions and work performance in the health and work performance questionnaire calibration surveys. *J Occup Environ Med*. 2003;45: 1303–1311.
42. Kessler RC, Greenberg PE, Mickelson KD, et al. The effects of chronic medical conditions on work loss and work cutback. *J Occup Environ Med*. 2001;43: 218–225.
43. Druss BG, Marcus SC, Olfson M, et al. Comparing the national economic burden of five chronic conditions. *Health Aff (Millwood)*. 2001;20:233–241.
44. Wilens TE, Faraone SV, Biederman J. Attention-deficit/hyperactivity disorder in adults. *JAMA*. 2004;292:619–623.
45. Schweitzer JB, Lee DO, Hanford RB, et al. Effect of methylphenidate on executive functioning in adults with attention-deficit/hyperactivity disorder: normalization of behavior but not related brain activity. *Biol Psychiatry*. 2004;56:597–606.
46. Turner DC, Clark L, Dowson J, et al. Modafinil improves cognition and response inhibition in adult attention-deficit/hyperactivity disorder. *Biol Psychiatry*. 2004;55:1031–1040.
47. Simpson D, Plosker GL. Spotlight on atomoxetine in adults with attention-deficit hyperactivity disorder. *CNS Drugs*. 2004;18:397–401.

Copyright © American College of Occupational and Environmental Medicine. Unauthorized reproduction of this article is prohibited.

NEURASTHENIA SECTION (N)

*N1. (READ SLOWLY) Did you ever in your life have a period lasting several months or longer when you became very tired, weak, or exhausted either while performing minor everyday physical tasks like working, shopping, housekeeping, and walking, or while performing everyday mental tasks like reading, writing, and doing paperwork?

YES 1
NO ..... 5 **GO TO *FD1, NEXT SECTION**
DON'T KNOW ..... 8 **GO TO *FD1, NEXT SECTION**
REFUSED ..... 9 **GO TO *FD1, NEXT SECTION**

---

*N2. What would happen when you tried to rest or relax — would you fully regain your energy and strength? Or would you still feel tired or weak?

FULLY REGAIN ..... 1 **GO TO *FD1, NEXT SECTION**
STILL FEEL TIRED/WEAK..... 2
DON'T KNOW 8
REFUSED..... 9

---

*N3. During the months or years when this problem was most severe, how often did you get tired — nearly every day, most days, about half the days, or less than half the days?

NEARLY EVERY DAY..... 1
MOST DAYS 2
ABOUT HALF THE DAYS..... 3 **GO TO *FD1, NEXT SECTION**
LESS THAN HALF THE DAYS .... 4 **GO TO *FD1, NEXT SECTION**
DON'T KNOW ..... 8 **GO TO *FD1, NEXT SECTION**
REFUSED..... 9 **GO TO *FD1, NEXT SECTION**

---

*N4. How much did your tiredness ever interfere with either your work, your social life, or your personal relationships – not at all, a little, some, a lot, or extremely?

NOT AT ALL ..... 1 **GO TO *FD1, NEXT SECTION**
A LITTLE ..... 2 **GO TO *FD1, NEXT SECTION**
SOME..... 3
A LOT 4
EXTREMELY 5
DON'T KNOW 8
REFUSED..... 9 **GO TO *FD1, NEXT SECTION**

*N4a. How often were you too tired to carry out your daily activities – often, sometimes, rarely, or never?

OFTEN.....1
SOMETIMES 2
RARELY.....3
NEVER 4
DON'T KNOW 8
REFUSED.....9

***CC20**. The next few questions are about problems with your sleep. Did you have a period lasting two weeks or longer in the past 12 months when you had any of the following problems with your sleep:

	YES (1)	NO (5)	DK (8)	RF (9)
*CC20a. Problems getting to sleep, when nearly every night it took you two hours or longer before you could fall asleep?	1	5	8	9
*CC20b. Problems staying asleep, when you woke up nearly every night and took an hour or more to get back to sleep?	1	5	8	9
*CC20c. Problems waking too early, when you woke up nearly every morning at least two hours earlier than you wanted to?	1	5	8	9
*CC20d. Problems feeling sleepy during the day?	1	5	8	9

*CC21. INTERVIEWER CHECKPOINT: (SEE ***CC20 SERIES**)

ONE OR MORE "YES" RESPONSES IN ***CC20 SERIES**.....1
ALL OTHERS.....2 **GO TO *CC28.1**

---

*CC22. About how many weeks in the past 52 did you have problems with your sleep?

_____ WEEKS

DON'T KNOW 998
REFUSED 999

---

*CC23. INTERVIEWER CHECKPOINT: (SEE ***CC20d**)

***CC20d** EQUALS "YES" 1
ALL OTHERS.....2 **GO TO *CC26**

*CC24. Think of the time during the past 12 months when your sleep problems were most severe and frequent. During that time, how often did you have each of the following experiences:						
(IF NEC: How often did you…/ Often, sometimes, rarely, or never?)	**OFTEN (1)**	**SOME (2)**	**RARE (3)**	**NEVER (4)**	**DK (8)**	**RF (9)**
*CC24a. …fall asleep while watching TV, listening to the radio, or reading – often, sometimes, rarely, or never?	1	2	3	4	8	9
*CC24b. …get drowsy within 10 minutes of sitting still?	1	2	3	4	8	9
*CC24c. …doze off when you relaxed?	1	2	3	4 **GO TO *CC24f**	8	9
*CC24d. …fall asleep during conversations or while visiting friends?	1	2	3	4	8	9
*CC24e. …feel fatigued during the day because of poor sleep?	1	2	3	4	8	9
***CC24f**. …wake up more than 3 times per night?	1	2	3	4	8	9
*CC24g. …wake up feeling rested?	1	2	3	4	8	9
*CC24h. …have difficulty getting up in the morning?	1	2	3	4	8	9
*CC24j. …feel you had not slept long enough even after having enough time in bed?	1 **GO TO *CC28.1**	2 **GO TO *CC28.1**	3 **GO TO *CC28.1**	4 **GO TO *CC28.1**	8 **GO TO *CC28.1**	9 **GO TO *CC28.1**

***CC26**. Think of the time during the past 12 months when your sleep problems were most severe and frequent. During that time, how often did you have each of the following experiences:						
(IF NEC: How often did you…/ Often, sometimes, rarely, or never?)	**OFTEN (1)**	**SOME (2)**	**RARE (3)**	**NEVER (4)**	**DK (8)**	**RF (9)**
*CC26a. …wake up more than 3 times per night?	1	2	3	4	8	9
*CC26b. …wake up feeling rested?	1	2	3	4	8	9
*CC26c. …have difficulty getting up in the morning?	1	2	3	4	8	9
*CC26d. …feel refreshed after sleeping?	1	2	3	4	8	9
*CC26e …feel you had not slept long enough even after having enough time in bed?	1	2	3	4	8	9

***CC28.1.** The next questions are about "medically unexplained chronic pain". This is defined as pain lasting six months or longer that is (READ SLOWLY) severe enough either to interfere a lot with your normal activities or to cause a lot of emotional distress and that a doctor cannot find a physical cause to explain. With that definition in mind, did you ever have "medically unexplained chronic pain"?

YES.....1
NO .....5 **GO TO *CC36**
DON'T KNOW .....8 **GO TO *CC36**
REFUSED .....9 **GO TO *CC36**

*CC28.1a. In what part of your body did the pain occur?

RECORD ALL MENTIONS

NECK OR BACK.....1
STOMACH OR ABDOMEN 2
JOINT LIKE ARMS, HANDS, LEGS, OR FEET 3
FACE OR JAW OR THE JOINT JUST BELOW THE EAR 4
CHEST.....5
ANY OTHER TYPE OF CHRONIC PAIN (SPECIFY).....6

_____

DON'T KNOW.....8
REFUSED.....9

---

*CC28.2. Thinking of a time in your life when the pain was most consistent and severe, how much emotional distress did you experience because of your pain – none, mild, moderate, severe, or very severe distress?

NONE 1
MILD 2
MODERATE.....3
SEVERE 4
VERY SEVERE 5
DON'T KNOW 8
REFUSED.....9

*CC28.3. How much did your pain ever interfere with either your work, your social life, or your personal relationships – not at all, a little, some, a lot, or extremely?

NOT AT ALL..... 1 **GO TO *CC28.4**
A LITTLE..... 2 **GO TO *CC28.4**
SOME 3
A LOT..... 4
EXTREMELY..... 5
DON'T KNOW 8
REFUSED ..... 9 **GO TO *CC28.4**

*CC28.3a. How often were you unable to carry out your daily activities because of your pain – often, sometimes, rarely, or never?

OFTEN.....1
SOMETIMES 2
RARELY.....3
NEVER 4
DON'T KNOW 8
REFUSED.....9

***CC28.4**. INTERVIEWER CHECKPOINT: (SEE ***CC28.2, *CC28.3, *CC28.3a**)

***CC28.**2 EQUALS '3' – '5' OR ***CC28.**3 EQUALS '3' – '5' OR
***CC28.3a** EQUALS '1' – '3' 1
ALL OTHERS.....2 **GO TO *CC36**

---

*CC29.1. How old were you the first time you had "medically unexplained chronic pain"?

_____ YEARS OLD

DON'T KNOW 998
REFUSED 999

---

*CC29.2. How many years altogether did you have medically unexplained pain for at least six months of the year?

_____ YEARS

DON'T KNOW 998
REFUSED 999

---

*CC29.3. Did you have medically unexplained chronic pain for six months or longer in the past 12 months?

YES.....1
NO.....5 **GO TO *CC34.1**
DON'T KNOW .....8 **GO TO *CC34.1**
REFUSED.....9 **GO TO *CC34.1**

---

*CC29.4. How many months in the past 12 months did you have this pain?

_____ MONTHS

DON'T KNOW 998
REFUSED 999

Under review. Cited with the permission of the author (Ronald C. Kessler).

Word count: 3805
Abstract: 199
Tables: 5
Figures: 1

Sleep problems, comorbid mental disorders, and role functioning in the National Comorbidity Survey Replication (NCS-R)

Running title: Sleep problems in the NCS-R

Keywords: insomnia, nonrestorative sleep, epidemiology, comorbidity, NCS-R, sleep

Thomas Roth, Ph.D.

Savina Jaeger, Ph.D.

Robert Jin, M.A.

Anupama A. Krishnan, M.S.

Paul E. Stang, Ph.D.

Ronald C. Kessler, Ph.D.

October 2005

From the Sleep Disorders Center, Department of Psychiatry and Behavioral Neurosciences, Henry Ford Hospital, Detroit, MI (Roth), the Department of Health Care Policy, Harvard Medical School (Jaeger, Jin, Kessler), Global Health Outcomes Research, Eli Lilly and Company (Krishnan), West Chester University of Pennsylvania (Stang), and Galt Associates (Stang). Direct comments to RC Kessler, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, USA 02115. Voice: 617-432-3587; Fax: 617-432-3588; E-mail: kessler@hcp.med.harvard.edu.

Sleep problems, comorbid mental disorders, and role functioning in the National Comorbidity Survey Replication (NCS-R)

ABSTRACT

Background. Little is known about the prevalence of sleep problems in the US general population and whether the associations of sleep problems with role impairment are due to comorbid mental disorders.

Methods. The associations of 12-month sleep problems (difficulty initiating or maintaining sleep, early morning awakening, nonrestorative sleep) with role impairment were analyzed in Part II of the National Comorbidity Survey Replication (n=5692) controlling for 12-month DSM-IV anxiety, mood, impulse-control, and substance disorders. The WHO Composite International Diagnostic Interview was used to assess sleep problems and DSM-IV disorders. The WHO Disability Schedule-II (WHO-DAS) was used to assess role impairment.

Results. Prevalence estimates of the four sleep problems are 16.4-25.0%, with 36.3% reporting at least one such problem. Mean duration of sleep problems in the past 12 months was 24.4 weeks, with 32.1% of cases reporting short durations (2-4 weeks) and 28.0% long durations (51-52 weeks). The sleep problems were all significantly comorbid with all the 12-month DMS-IV disorders (median OR: 3.4; 25th-75th percentile: 2.8-3.9) and related to substantial role impairment that generally remained significant after controlling comorbid mental disorders.

Conclusions. The four sleep problems considered here are of great public health significance because of their high prevalence and significant associations with role impairment.

INTRODUCTION

DSM-IV defines insomnia as a syndrome characterized by problems in one or more of four sleep domains – difficulty initiating sleep (DIS), difficulty maintaining sleep (DMS), early morning awakening (EMA), and not feeling rested even after ample time in bed (nonrestorative sleep or NRS) -- associated with impairment in daytime functioning. General population surveys consistently find that roughly one-third of the US adult population report sleep problems (Ancoli-Israel and Roth 1999; Grandner and Kripke 2004; National Sleep Foundation 2005) and that 10-15% of the population meet diagnostic criteria for DSM-IV insomnia (Breslau et al 1996; Cirignotta et al 1985; Costa e Silva et al 1996; Ohayon 1996; Ohayon 1997; Ohayon 2002).

The impaired daytime functioning required for a DSM-IV diagnosis of insomnia has been documented across a wide range of role functioning domains (Leger et al 2001) with a severity distribution comparable to that of many chronic physical disorders (Katz and McHorney 2002). Given the high prevalence of insomnia, this evidence of substantial impairment could have considerable public health significance if it was known to be due to insomnia. As insomnia is strongly comorbid with a number of mental disorders (Breslau et al 1996; Ford and Kamerow 1989), an unanswered question is whether the impairment associated with DIS, DMS, EMA or NRS is due to the sleep problem itself or to comorbid mental disorders. The current report addresses this issue by analyzing data from the National Comorbidity Survey Replication (NCS-R; Kessler and Merikangas 2004).

METHODS

Sample

The NCS-R is a nationally representative, face-to-face household survey of adults (ages 18+) based on a multi-stage clustered area probability sampling design (Kessler et al 2004b). A total of 9282 respondents participated in the survey (February 2001 - December 2003). The response rate was 70.9 %. Participants received a $50 honorarium. After complete description of the study to the subjects, verbal informed consent was obtained. Consent was verbal rather than written in order to be consistent with the recruitment procedures used in the baseline NCS (Kessler et al 1994). The Human Subjects Committees of Harvard Medical School and the University of Michigan both approved these procedures.

All NCS-R respondents completed a Part I diagnostic interview using the WHO Composite International Diagnostic Interview (CIDI 3.0; Kessler and Ustun 2004), while a probability sub-sample of 5692 Part I respondents also received a Part II interview that assessed additional disorders and correlates. Sleep problems were assessed in Part II. Part II respondents included all who met criteria for any Part I disorder plus a probability sub-sample of others. The sample was weighted for differential probabilities of selection within households, differential recruitment intensity, over-sampling of Part I respondents with disorders into Part II, and residual discrepancies with the 2000 Census on socio-demographic and geographic variables. More complete information on NCS-R sampling and weighting is reported elsewhere (Kessler et al 2004b).

Sleep problems

Part II NCS-R respondents were asked using a yes-no response format whether they had "periods lasting *two weeks or longer* in the past 12 months" when they experienced "problems

getting to sleep, when nearly every night it took you *two hours or longer* before you could fall asleep" (DIS), "problems *staying* asleep, when you woke up nearly every night and took an hour or more to get back to sleep" (DMS), or "problems waking *too early*, when you woke up nearly every morning at least two hours earlier than you wanted to" (EMA). (Italics indicate emphasis in the original questions.)

Nonrestorative sleep (NRS) was assessed with a scale based on responses to four questions about the frequency of sleep problems during the time in the past 12 months when such problems were most severe and frequent. These questions asked respondents about waking up not feeling rested, having difficulties getting up in the morning, feeling as if they had not slept long enough even after having enough time in bed, and not feeling refreshed after sleep. Response options were often, sometimes, rarely, and never (0-3). Principal axis factor analysis documented a clear one-factor structure, with factor loadings in the range .79-.86. A summary factor-based scale with a range of 0-12 was creating by summing responses to the individual items. Nonlinear regression analysis was used to study the associations of scores on this scale with the measures of role impairment described below, controlling for basic socio-demographic variables. A dichotomous cut-point on the scale to define the clinical threshold for nonrestorative sleep was selected to discriminate optimally among responses to the outcome measures. (More details available on request.) This dichotomization was carried out to facilitate comparison of NRS with the other dichotomously assessed sleep problems.

Respondents who reported any of the above four sleep problems were asked how many weeks out of 52 in the past 12 months they had any of these problems.

Comorbid DSM-IV disorders

Core DSM-IV disorders assessed in the NCS-R with the CIDI include anxiety disorders (panic disorder with or without agoraphobia, agoraphobia without panic disorder, specific phobia, social phobia, generalized anxiety disorder, post-traumatic stress disorder, separation anxiety disorder), mood disorders (major depressive disorder, dysthymia, bipolar disorder I or II), impulse-control disorders (intermittent explosive disorder, attention-deficit/hyperactivity disorder, oppositional-defiant disorder, conduct disorder), and substance use disorders (alcohol and illicit drug abuse and dependence with abuse). Organic exclusion rules and diagnostic hierarchy rules were used in making diagnoses. As detailed elsewhere (Kessler et al 2004a; Kessler et al 2005), blinded clinical re-interviews using the non-patient version of the Structured Clinical Interview for DSM-IV (SCID; First et al 2002) with a probability sub-sample of NCS-R respondents found generally good concordance between DSM-IV diagnoses of anxiety, mood, and substance disorders based on the CIDI and the clinical assessments. Impulse-control disorder diagnoses were not validated, as the SCID clinical reappraisal interviews did not include an assessment of these disorders.

Role impairment

Role impairment was assessed with five measures. The first four of these five were taken from the World Health Organization Disability Assessment Schedule-II (WHO-DAS; Chwastiak and Von Korff 2003). and asked respondents to report the number of days out of the past 30 when they had each of the following experiences: totally unable to work or carry out their normal daily activities (days out of role); able to carry out their normal activities but had to cut down on what they did or not get as much done as usual (reduced quantity); cut back on the quality of your work or how carefully you worked (reduced quality); and needed to make an extreme effort to

perform up to your usual in carrying out their normal daily activities (extreme effort). Each of the four responses is in the range 0-30 days.

The fifth measure of role impairment was based on responses to five questions about the frequency of daytime sleepiness assessed in conjunction with the assessment of nonrestorative sleep. The daytime sleepiness questions asked about falling asleep while watching TV or listening to the radio or reading, getting drowsy within ten minutes of sitting down, dozing off while relaxing, falling asleep during conversations or while visiting friends, and feeling fatigued during the day because of poor sleep. As with the assessment of nonrestorative sleep, response options were often, sometimes, rarely, and never. Principal axis factor analysis documented a clear one-factor structure (factor loadings in the range .78-.87). A summary factor-based scale with a range of 0-100 was creating by summing responses to the individual items and transformed to a 0-100 metric for ease of interpretation. The highest score (100) on the scale was assigned to respondents who reported often having all five indicators of daytime sleepiness while the lowest score (0) was assigned to respondents who reported never having any of these experiences.

Socio-demographic controls

Socio-demographic control variables included gender, age (18-29, 30-44, 45-59, 60+), race-ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic, Other), education (less than high school graduation, high school, some post high school education without a college degree, and college degree or more), marital status (married-cohabiting, never married, separated-divorced, widowed), employment status (employed or self-employed, student, homemaker, retired, other), family income (in quartiles of the population distribution), and number of pre-school children (0, 1, 2+).

Analysis methods

Odds-ratios (ORs) and tetrachoric correlations were used to examine associations among the four dichotomous measures of DIS, DMS, EMA and, NRS. Associations with socio-demographics and comorbid disorders were examined using logistic regression analysis in which the four sleep problems were treated as dichotomous outcomes. Logistic regression coefficients and their standard errors were exponentiated and are presented as ORs with 95% confidence intervals (CIs). Associations with role impairment were estimated using linear regression analysis in which sleep problems were treated as predictors, controlling for socio-demographics and for comorbid DSM-IV disorders. Unstandardized regression coefficients and their standard errors are presented. Because the NCS-R sample design features weighting and clustering, all parameter estimates were estimated using the Taylor series linearization method (Wolter 1985) implemented in the SUDAAN software system (Research Triangle Institute 2002). The significance of set of coefficients was assessed with design-corrected Wald χ^2 tests. Significance was evaluated using two-sided .05-level design-based tests.

RESULTS

Prevalence, inter-correlations and duration

Twelve-month prevalence estimates are 16.4% for DIS, 19.9% for DMS, 16.7% for EMA, 25.0% for NRS, and 36.3% for any of the four sleep problems. (Table 1) The four are strongly interrelated, with tetrachoric correlations in the range .65-.76. Mean duration in the twelve months before interview was 22.4 weeks among respondents who reported any of the four problems and in the range 25.2-28.7 weeks among respondents for any given individual sleep problems. The mean is somewhat deceptive, as duration shows a bimodal distribution in which

one-third of cases (32.1%) reported a short duration (2-4 weeks), 28.0% reported persistence throughout the year, and the remainder reporting intermediate durations. (Figure 1)

(Table 1 and Figure 1 about here)

Although the four sleep problems are strongly inter-correlated, all logically possible combinations exist among them. (Table 2) The proportion of the total sample that reported only one of the three types of classic sleep problems -- DIS, DMS or EMA -- (12.8%) is higher than the proportion that reported either two (9.0%) or all three (7.4%). Furthermore, while the conditional probability of nonrestorative sleep is highest among respondents with the other three sleep problems (73.4%), lower among those with two (59.2-66.4%) or one (50.2-55.7%) and lowest among those with none of the three (10.0%), nearly one-third of the respondents with nonrestorative sleep reported not having DIS, DMS or EMA. Mean duration of sleep problems in the twelve months before interview was highest among respondents who reported all four sleep problems (34.5 weeks) and lowest among those who reported only one (11.2-24.8 weeks).

(Table 2 about here)

Socio-demographic correlates

Women have modestly elevated odds of all four sleep problems compared to men (1.1-1.6). (Table 3) Age is inversely related to DIS and NRS (1.9-2.7 elevated odds among respondents ages 18-29 vs. 60+), while middle-aged respondents have the highest odds of DMS and EMA (1.3-1.7 elevated odds among respondents 30-44 and 45-59 vs. 60+). Non-Hispanic Blacks have a somewhat lower odds of NRS (0.7), but not of any other sleep problem (0.8-1.1), than Non-Hispanic Whites. Education is inversely related to DIS, DMS and EMA (1.5-2.2 elevated odds among respondents with 0-11 vs. 16+ years of education), but not to NRS. Previously married people have consistently elevated odds of all four sleep problems compared

to the married (1.4-1.8), while never married people have elevated odds only of DIS (1.4). Disabled people have elevated odds of all four sleep problems compared to the employed (1.8-3.2), while homemakers have elevated odds only of DIS (1.7) and retired people only of EMA (2.1). Somewhat surprisingly, the number of young children in the home and family income are unrelated to any of the sleep problems.

(Table 3 about here)

Comorbidity with DSM-IV disorders

The four sleep problems are all significantly and positively related to each of the 12-month DMS-IV disorders considered here, resulting in roughly half (47.8-53.7) of respondents with the separate sleep problems meeting criteria for one or more of these disorders. Odds-ratios are in the range 1.6-6.1, with a median OR of 3.4 and an inter-quartile range (25th-75th percentiles) of 2.8-3.9. No consistent variation can be seen in the magnitude of ORs either across the four sleep problems for particular classes of DSM-IV disorders or across classes of these disorders for individual sleep problems. The only clear pattern of variation involves comorbidity among DSM-IV disorders, with respondents who have high comorbidity (defined as meeting criteria for three or more 12-month DSM-IV disorders) also having much higher odds of sleep problems (ORs in the range 4.6-6.3 compared to respondents with no 12-month DSM-IV disorders) than respondents with either low comorbidity (defined as meeting criteria for exactly two 12-month DSM-IV disorders, with ORs in the range 2.2-3.2) or pure DSM-IV disorders (ORs in the range 1.5-2.0).

(Table 4 about here)

Role impairment

The associations of sleep problems in statistically predicting role impairment were examined in three different models. The first was a model in which the associations of the four sleep problem with each measure of role impairment were considered at the same time with controls for socio-demographic variables. All 20 of the coefficients for the sleep problems (i.e., each of four sleep problems predicting each of five measures of role impairment) in this gross effects model are significant at the .05 level. (Table 5) Coefficients to predict daytime sleepiness range from a low of 19.3 (associated with DIS) to a high of 34.4 (associated with NRS) in the 0-100 metric of the daytime sleepiness scale. Coefficients to predict number of days out of role in the past 30 days range between 3.2 days (associated with EMA) and 4.0 days (associated with DIS). Coefficients to predict number of days in the past 30 when respondents reported either reduced quantity or work, reduced quality of work, or extreme effort to complete work range between 1.0 days (EMA predicting days of reduced work quality) and 2.0 days (NRS predicting days of reduced work quantity and days of extreme effort).

(Table 5 about here)

The second model added controls for all the DSM-IV disorders examined in the analysis of comorbidity in order to partial out the effects of mental disorders in examining the associations of sleep problems with role impairment. All 20 coefficients in this net effects model are lower than those in the gross effects model, although 13 of the 20 coefficients remain significant at the .05 level. The three coefficients that significantly predict daytime sleepiness are in the range 6.1-29.8. The three coefficients that significantly predict days out of role are in the range 1.3-2.2 days. The seven coefficients that significantly predict days with reduced quantity

or work, reduced quality of work, or extreme effort are in the range 0.4-1.5 days. NRS sleep is the only sleep problem that significantly predicts all five measures of role impairment, while DIS predicts four, DMS three, and EMA one.

The third model controls comorbid DSM-IV in a more structural way: by focusing only on the subset of respondents who had none of the 12-month DSM-IV disorders. Ten of the 20 coefficients in this pure effects model are positive and significant at the .05 level. As in the net effects model, NRS is the only sleep problem that significantly predicts all five measures of role impairment, while DIS predicts three, DMS one, and EMA one. The significant coefficients are generally lower than those in the gross effects model but higher than those in the net effects model. One negative coefficient is significant (DMS predicting days of extreme effort).

DISCUSSION

These results have to be interpreted with three limitations in mind. First, sleep problems were assessed in the NCS-R using self-report measures with ad hoc cut-points that were not validated against a clinical assessment. Second, because of the comparatively low survey response rate, people with sleep disorders might be represented in a different proportion than in the population. Third, the four WHO-DAS impairment measures were assessed for the past 30 days while sleep problems were assessed for the past 12 months, leading to a likely conservative bias in the estimated effects of sleep problems on role functioning.

Within the context of these limitations, we estimate that the sleep problems considered here are highly prevalent, that these problems often persist throughout the year, that they often co-occur with DSM-IV mental disorders, and that they are associated with substantial role impairment that cannot be explained by comorbid mental disorders.

Consistent with previous reports, approximately one-third of NCS-R respondents reported one or more of these four sleep problems (Ancoli-Israel and Roth 1999; Grandner and Kripke 2004; National Sleep Foundation 2005). We also examined associations among the four problems and found that NRS is the most common of the four (25%), with the others in the range 16.4-19.9%. This higher prevalence of nonrestorative sleep presumably reflects the fact that it can occur as a result of DIS, DMS or EMA as well as in the absence of and of the three classic sleep problems. All four sleep problems were highly inter-correlated, with roughly two-thirds of respondents who reported any having more than one. About one-third of respondents with these problems reported that the problems persisted throughout the entire past 12 months. Chronicity is much more strongly associated with number than type of sleep problems.

The fact that roughly one-third of people with NRS report neither DIS, DMS, nor EMA implies that nonrestorative sleep is sometimes indicative of poor sleep quality or continuity rather than short sleep duration. This is plausible in light of the fact that several sleep disorders (e.g., sleep apnea) have their primary effects on sleep quality and continuity rather than on sleep duration.

The socio-demographic correlates of sleep problems in the NCS-R are similar to those found in previous surveys (Roth and Roehrs 2003), although our more fine-grained analysis showed that these correlates vary by type of sleep problem. The most important of these specifications involves age. DMS and EMA are most common among the elderly, while nonrestorative sleep is most common among the young, and DIS is inversely related to age. These results clearly suggest that the focus of previous reports on increasing age as a risk factor for difficulties with sleep (Griffiths and Peerson 2005; Ohayon 2005) need to be revised to recognize the heterogeneity of the different sleep problems.

Major depression is the only DSM-IV mental disorder that includes any of the sleep problems considered here in its diagnostic criteria. Consistent with this fact, several previous epidemiological studies have shown that major depression is more strongly related to a diagnosis of insomnia than are other mental disorders (Breslau et al 1996; Ford and Kamerow 1989; Ohayon and Roth 2003). This finding is superficially inconsistent with the NCS-R finding that the sleep problems considered here are not more strongly related to major depression than to the other DSM mental disorders evaluated in the survey. It is important to remember, though, that a DSM diagnosis of insomnia requires not only sleep problems but also daytime impairment associated with these problems. It might be, then, that the seeming inconsistency between findings regarding depression the NCS-R compared to earlier surveys is due to major depression being more strongly related to the daytime impairment caused by sleep problems rather than to these sleep problems themselves. Although exploration of this possibility goes beyond the bounds of the current report, it should be included in future investigations of sleep disturbance in depression.

The consistency and strength of the associations between sleep problems and role impairment are striking. This is especially true with regard to the excess numbers of days out of role associated with sleep problems, which exceed the numbers found in previous studies to be associated with most chronic physical and mental disorders (Kessler et al 2003). This is consistent with previous reports of strong associations between sleep problems and days out of role (Simon and Von Korff 1997). The net effects and sub-group analyses argue against the possibility that the associations of sleep problems with role impairment are due to comorbid mental disorders, although the possibilities remain that unmeasured comorbid physical disorders or systematic response bias might be involved.

It is important to note that NRS is the most consistent correlate of role impairment among the sleep problems. NRS is also the only sleep problem significantly related to all the measures of role impairment among respondents without comorbid DSM-IV mental disorders. These are not unexpected results in light of the fact that nonrestorative sleep is the only type of insomnia that involves wake functioning rather than sleep functioning.

As the threshold used to define NRS was high, out estimate of the prevalence of NRS is likely a conservative. Furthermore, the respondents classified as having NRS represent a more severe NRS group with higher levels of role impairment than if the definition were broadened. These same conclusions would likely hold for DIS, DMS and EMA, all of which were defined using conservative criteria in this study. For example, in order to be classified as a DIS sufferer, a patient had to experience "*two hours or longer period in bed* before falling asleep", compared to the more widely used 30 minute sleep latency criterion used in other epidemiologic studies and clinical trials (Ohayon, 2005). Future research should examine the sensitivity of results to variation in cut-points on these dimensions.

The suggestion that a substantial proportion of people with DIS, DMS and EMA might have NRS has significant implications for therapeutics. For example, in the cognitive-behavioral treatment of sleep disorders, one could speculate that sleep restriction therapy (which has the goal of enhancing sleep drive) would be more effective in cases of NRS than stimulus control therapy (which has the goal of hastening sleep onset). In the pharmacological treatment of sleep disorders, one could speculate that NRS might best be treated with medications that enhance slow wave sleep, thereby preventing awakening and arousals, than with medications that hasten sleep onset or that decrease the duration of night time awakenings.

The results reported here leave a number of issues unresolved that could be addressed in the NCS-R data, but go beyond the boundaries of this first report. We already noted the need to explore the differential associations of depression and other DSM-IV mental disorders with sleep problems versus DSM-IV insomnia. In addition, future research should investigate the joint effects of multivariate sleep problem profiles, the effects of chronic versus intermittent sleep problems, the role of comorbid physical disorders, and the extent to which the associations of sleep problems with role impairment are due to daytime sleepiness. In addition, the recent report of the NIH consensus panel on insomnia noted that more research is needed on the associations of sleep problems with work performance and disability (National Institutes of Health 2005). All of these issues will be examined in ongoing analyses of the NCS-R data.

ACKNOWLEDGEMENTS

The National Comorbidity Survey Replication (NCS-R) is supported by NIMH (U01-MH60220) with supplemental support from NIDA, the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044708), and the John W. Alden Trust. Preparation of this report was additionally supported by a grant from Eli Lilly and Company. Collaborating investigators include Ronald C. Kessler (Principal Investigator, Harvard Medical School), Kathleen Merikangas (Co-Principal Investigator, NIMH), James Anthony (Michigan State University), William Eaton (The Johns Hopkins University), Meyer Glantz (NIDA), Doreen Koretz (Harvard University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Greg Simon (Group Health Cooperative), Michael Von Korff (Group Health Cooperative), Philip Wang (Harvard Medical School), Kenneth Wells (UCLA), Elaine Wethington (Cornell University), and Hans-Ulrich Wittchen (Max Planck Institute of Psychiatry). The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or U.S. Government. A complete list of NCS publications and the full text of all NCS-R instruments can be found at http://www.hcp.med.harvard.edu/ncs**.** Send correspondence to NCS@hcp.med.harvard.edu. The NCS-R is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative. We thank the staff of the WMH Data Collection and Data Analysis Coordination Centres for assistance with instrumentation, fieldwork, and consultation on data analysis. These activities were supported by the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13-MH066849, R01-MH069864, and R01 DA016558), the Pan American Health Organization, Eli

Lilly and Company, GlaxoSmithKline, and Ortho-McNeil Pharmaceutical, Inc. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmhcidi.

REFERENCES

Ancoli-Israel S, Roth T (1999): Characteristics of insomnia in the United States: results of the 1991 National Sleep Foundation Survey. I. *Sleep* 22 Suppl 2:S347-S353.

Breslau N, Roth T, Rosenthal L, Andreski P (1996): Sleep disturbance and psychiatric disorders: a longitudinal epidemiological study of young adults. *Biol Psychiatry* 39:411-418.

Chwastiak LA, Von Korff M (2003): Disability in depression and back pain: evaluation of the World Health Organization Disability Assessment Schedule (WHO DAS II) in a primary care setting. *J Clin Epidemiol* 56:507-514.

Cirignotta F, Mondini S, Zucconi M, Lenzi PL, Lugaresi E (1985): Insomnia: an epidemiological survey. *Clin Neuropharmacol* 8 Suppl 1:S49-S54.

Costa e Silva JA, Chase M, Sartorius N, Roth T (1996): Special report from a symposium held by the World Health Organization and the World Federation of Sleep Research Societies: an overview of insomnias and related disorders--recognition, epidemiology, and rational management. *Sleep* 19:412-416.

First MB, Spitzer RL, Gibbon M, Williams JBW (2002): *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-Patient Edition (SCID-I/NP)*. New York, NY: Biometrics Research, New York State Psychiatric Institute.

Ford DE, Kamerow DB (1989): Epidemiologic study of sleep disturbances and psychiatric disorders. An opportunity for prevention? *JAMA* 262:1479-1484.

Grandner MA, Kripke DF (2004): Self-reported sleep complaints with long and short sleep: a nationally representative sample. *Psychosom Med* 66:239-241.

Griffiths MF, Peerson A (2005): Risk factors for chronic insomnia following hospitalization. *J Adv Nurs* 49:245-253.

Katz DA, McHorney CA (2002): The relationship between insomnia and health-related quality of life in patients with chronic illness. *J Fam Pract* 51:229-235.

Kessler RC, Abelson J, Demler O, Escobar JI, Gibbon M, Guyer ME, et al (2004a): Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *Int J Methods Psychiatr Res* 13:122-139.

Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al (2003): The World Health Organization Health and Work Performance Questionnaire (HPQ). *J Occup Environ Med* 45:156-174.

Kessler RC, Berglund P, Chiu WT, Demler O, Heeringa S, Hiripi E, et al (2004b): The US National Comorbidity Survey Replication (NCS-R): design and field procedures. *Int J Methods Psychiatr Res* 13:69-92.

Kessler RC, Berglund P, Demler O, Jin R, Walters EE (2005): Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* 62:593-602.

Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al (1994): Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Arch Gen Psychiatry* 51:8-19.

Kessler RC, Merikangas KR (2004): The National Comorbidity Survey Replication (NCS-R): background and aims. *Int J Methods Psychiatr Res* 13:60-68.

Kessler RC, Ustun TB (2004): The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res* 13:93-121.

Leger D, Scheuermaier K, Philip P, Paillard M, Guilleminault C (2001): SF-36: evaluation of quality of life in severe and mild insomniacs compared with good sleepers. *Psychosom Med* 63:49-55.

National Institutes of Health (2005): State of the Science Conference Statement: Manifestations and management of chronic insomnia in adults. *Sleep* 28:1049-1057.

National Sleep Foundation (2005): Summary of Findings: 2005 Sleep in America Poll, Vol 2005. Washington, DC: National Sleep Foundation.

Ohayon M (1996): Epidemiological study on insomnia in the general population. *Sleep* 19:S7-S15.

Ohayon MM (1997): Prevalence of DSM-IV diagnostic criteria of insomnia: distinguishing insomnia related to mental disorders from sleep disorders. *J Psychiatr Res* 31:333-346.

Ohayon MM (2002): Epidemiology of insomnia: what we know and what we still need to learn. *Sleep Med Rev* 6:97-111.

Ohayon MM (2005): Prevalence and correlates of nonrestorative sleep complaints. *Arch Intern Med* 165:35-41.

Ohayon MM, Roth T (2003): Place of chronic insomnia in the course of depressive and anxiety disorders. *J Psychiatr Res* 37:9-15.

Research Triangle Institute (2002): SUDAAN: Professional Software for Survey Data Analysis, 8.01 ed. Research Triangle Park, NC: Research Triangle Institute.

Roth T, Roehrs T (2003): Insomnia: epidemiology, characteristics, and consequences. *Clin Cornerstone* 5:5-15.

Simon GE, Von Korff M (1997): Prevalence, burden, and treatment of insomnia in primary care. *Am J Psychiatry* 154:1417-1423.

Wolter K (1985): *Introduction to Variance Estimation*. New York, NY: Springer-Verlag.

Table 1. Prevalence, inter-correlations, and mean duration (weeks in the past year) of 12-month sleep problems in the Part II NCS-R (n=5692)

	Prevalence		Inter-correlations[1]				Duration		
			DIS	DMS	EMA	NRS			
	%	(se)	OR	OR	OR	OR	Mean	(se)	(n)
Difficulty Initiating Sleep (DIS)	16.4	(0.6)	--	--	--	--	28.7	(0.9)	(1191)
Difficulty Maintaining Sleep (DMS)	19.9	(0.8)	0.76*	--	--	--	28.7	(0.7)	(1436)
Early Morning Awakening (EMA)	16.7	(0.7)	0.66*	0.84*	--	--	28.1	(0.9)	(1203)
Non-Restorative Sleep (NRS)	25.0	(0.8)	0.68*	0.70*	0.65*	--	25.2	(0.8)	(1901)
Any	36.3	(1.0)					24.4	(0.7)	(2578)

*Significant at the .05 level, two-sided test.
[1]Tetrachoric correlations

Table 2. The multivariate associations and mean duration of 12-month sleep problem profiles in the Part II NCS-R

Type of sleep problem	Multivariate distribution of DIS, DMS and EMA		Conditional prevalence of NRS		Mean Duration NRS+		Mean Duration NRS-		
	%	(se)	%	(se)	%	(se)	%	(se)	(n)
I. Only one									
DIS	4.7	(0.4)	55.7	(4.1)	24.0	(2.6)	24.8	(3.7)	(284)
DMS	4.5	(0.3)	53.7	(3.3)	24.1	(2.6)	20.9	(2.6)	(302)
EMA	3.6	(0.3)	50.2	(4.4)	18.9	(3.1)	11.2	(1.7)	(231)
II. Exactly two									
DIS-DMS	3.3	(0.3)	66.4	(4.4)	28.2	(1.9)	19.8	(3.9)	(245)
DIS-EMA	1.0	(0.2)	63.2	(6.5)	27.9	(3.4)	22.9	(5.6)	(83)
DMS-EMA	4.7	(0.5)	59.2	(4.8)	31.4	(1.4)	27.2	(3.6)	(310)
III. Other									
None	70.9	(0.9)	10.0	(0.6)	16.7	(1.1)	--	--	(3658)
DIS-DMS-EMA	7.4	(0.4)	73.4	(1.9)	34.5	(1.4)	29.5	(3.1)	(579)

Table 3. Socio-demographic correlates of 12-month sleep problems in the NCS-R (n = 5692)[1]

	DIS		DMS		EMA		NRS		Any 12-month sleep problem	
	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
Sex										
Female	1.1	(0.9-1.3)	1.2*	(1.0-1.5)	1.2	(1.0-1.5)	1.6*	(1.4-1.8)	1.4*	(1.2-1.6)
Male	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_1 (p-value)	0.7 (.410)		6.0* (.014)		3.4 (.064)		39.4* (.000)		25.5* (.000)	
Age										
18-29	1.9*	(1.1-3.2)	0.9	(0.6-1.5)	1.2	(0.7-1.9)	2.7*	(1.7-4.3)	1.8*	(1.2-2.5)
30-44	1.8*	(1.1-2.9)	1.3	(0.9-2.0)	1.6*	(1.1-2.3)	2.2*	(1.5-3.2)	1.7*	(1.3-2.2)
45-59	1.4	(0.9-2.4)	1.4*	(1.0-2.2)	1.7*	(1.3-2.4)	1.6*	(1.1-2.3)	1.5*	(1.1-1.9)
60+	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_3 (p-value)	10.8* (.013)		14.8* (.002)		19.8* (.000)		21.7* (.000)		15.3* (.002)	
Race-ethnicity										
Non-Hispanic White	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
Non-Hispanic Black	0.8	(0.6-1.2)	0.8	(0.6-1.1)	1.1	(0.9-1.5)	0.7*	(0.5-0.9)	0.7*	(0.6-1.0)
Hispanic	1.2	(0.9-1.5)	1.0	(0.7-1.3)	1.2	(0.8-1.7)	0.9	(0.7-1.1)	1.0	(0.8-1.2)
Other	1.5	(1.0-2.5)	1.3	(0.8-2.0)	1.4	(1.0-2.2)	1.0	(0.8-1.4)	1.2	(0.8-1.8)
χ^2_3 (p-value)	7.0 (.074)		2.6 (.460)		7.0 (.070)		7.9* (.049)		6.8 (.079)	
Education (years)										
0-11	2.2*	(1.6-3.1)	1.5*	(1.1-2.1)	2.0*	(1.5-2.8)	1.2	(0.9-1.5)	1.6*	(1.2-2.0)
12	1.9*	(1.4-2.6)	1.5*	(1.2-1.8)	1.7*	(1.3-2.1)	1.3*	(1.1-1.6)	1.4*	(1.2-1.7)
13-15	1.6*	(1.2-2.1)	1.4*	(1.1-1.7)	1.5*	(1.2-1.9)	1.3	(1.0-1.6)	1.4*	(1.2-1.8)
16+	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_3 (p-value)	27.6* (.000)		19.1* (.000)		33.5* (.000)		7.2 (.065)		20.6* (.000)	
Marital status										
Separated/Widowed/Divorced	1.8*	(1.5-2.2)	1.5*	(1.2-1.8)	1.6*	(1.2-2.0)	1.4*	(1.1-1.7)	1.4*	(1.2-1.7)
Never married	1.4*	(1.0-1.9)	0.9	(0.7-1.3)	1.1	(0.8-1.4)	1.0	(0.8-1.2)	1.1	(0.8-1.4)
Married	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_2 (p-value)	34.1* (.000)		21.8* (.000)		13.0* (.002)		11.1* (.004)		14.1* (.001)	
Employment										
Retired	1.4	(0.9-2.2)	1.3	(1.0-1.8)	2.1*	(1.6-2.7)	0.9	(0.6-1.3)	1.4*	(1.1-1.9)
Unemployed-disabled	3.2*	(2.3-4.5)	2.0*	(1.6-2.5)	2.0*	(1.5-2.5)	1.8*	(1.4-2.2)	2.2*	(1.7-2.9)
Student	0.7	(0.4-1.4)	0.6	(0.3-1.1)	0.6	(0.3-1.1)	0.9	(0.6-1.5)	0.8	(0.5-1.4)
Homemaker	1.7*	(1.1-2.7)	1.4	(1.0-2.0)	1.1	(0.8-1.6)	1.0	(0.7-1.3)	1.2	(0.9-1.7)
Employed	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_4 (p-value)	54.5* (.000)		51.7* (.000)		76.2* (.000)		28.5* (.000)		45.6* (.000)	
Number of pre-school children										
None	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
One	0.7	(0.5-1.0)	0.8	(0.6-1.1)	0.8	(0.6-1.2)	1.0	(0.8-1.3)	0.9	(0.7-1.2)
Two or more	0.9	(0.6-1.5)	1.2	(0.8-1.8)	1.0	(0.6-1.6)	0.9	(0.6-1.4)	1.1	(0.7-1.6)
χ^2_2 (p-value)	4.3 (.118)		3.3 (.194)		1.2 (.538)		0.4 (.798)		0.6* (.738)	
Family income										
Low	1.0	(0.8-1.5)	1.1	(0.8-1.4)	1.1	(0.8-1.6)	1.2	(0.9-1.6)	1.0	(0.8-1.3)
Low-average	1.0	(0.8-1.3)	1.1	(0.9-1.4)	1.0	(0.8-1.3)	1.2	(0.9-1.4)	1.0	(0.8-1.3)
High-average	1.0	(0.8-1.4)	1.0	(0.8-1.4)	1.2	(1.0-1.5)	1.2	(1.0-1.5)	1.2	(1.0-1.4)
High	1.0	--	1.0	--	1.0	--	1.0	--	1.0	--
χ^2_3 (p-value)	0.3 (.956)		1.3 (.734)		3.4 (.337)		2.9 (.413)		2.6 (.450)	

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

[1]Among respondents with one or more of the sleep problems, controlling for the multivariate profile of sleep problems

Table 4. Comorbidity of 12-month sleep problems with 12-month DSM-IV disorders in the NCS-R (n = 5692)[1]

	DIS			DMS			EMA			NRS			Any		
	%	OR	(95% CI)	%	OR	(95% CI)	%	OR	(95% CI)	%	OR	(95% CI)	%	OR	(95% CI)
I. Anxiety disorders															
Panic disorder	7.5	3.4*	(2.5-4.7)	7.2	4.0*	(3.0-5.2)	7.1	3.4*	(2.3-4.9)	7.0	4.1*	(3.0-5.5)	6.0	5.6*	(4.2-7.5)
Agoraphobia without panic	2.3	2.9*	(1.6-5.5)	2.2	3.9*	(2.2-7.0)	1.7	2.2*	(1.2-3.8)	1.6	2.4*	(1.3-4.3)	1.6	3.4*	(1.9-6.3)
Specific phobia	17.3	2.4*	(2.0-3.0)	16.8	2.5*	(2.1-3.1)	17.2	2.5*	(2.1-3.0)	18.3	3.2*	(2.7-3.8)	15.4	3.0*	(2.5-3.6)
Social phobia	16.9	3.5*	(3.0-4.1)	14.4	3.1*	(2.6-3.8)	15.2	3.3*	(2.7-4.0)	14.9	3.4*	(2.8-4.0)	12.5	3.4*	(2.8-4.2)
Generalized anxiety disorder	5.9	2.9*	(2.2-3.8)	6.0	3.0*	(2.3-3.9)	6.3	3.2*	(2.2-4.7)	7.2	6.1*	(4.4-8.3)	5.6	5.7*	(4.0-8.1)
Post-traumatic stress disorder	9.6	3.8*	(2.8-5.2)	10.0	5.0*	(3.8-6.5)	9.2	3.6*	(2.7-4.6)	9.3	5.0*	(3.6-6.9)	7.5	4.9*	(3.8-6.4)
Separation anxiety disorder	1.9	3.6*	(1.7-7.6)	1.2	2.5*	(1.0-6.2)	1.5	3.4*	(1.5-7.6)	1.8	8.3*	(4.0-17.2)	1.3	7.2*	(3.1-16.8)
Any anxiety disorder	37.0	3.2*	(2.6-3.9)	35.8	3.4*	(2.9-4.1)	35.8	3.3*	(2.8-3.9)	38.0	4.1*	(3.6-4.7)	32.5	4.0*	(3.4-4.8)
II. Mood disorders															
Major depressive disorder	16.5	3.5*	(2.8-4.4)	15.8	3.9*	(3.2-4.7)	14.0	2.9*	(2.4-3.4)	18.0	6.1*	(5.0-7.5)	14.2	5.8*	(4.5-7.5)
Dysthymia	4.9	5.4*	(3.2-9.0)	4.3	4.9*	(2.8-8.6)	3.8	3.2*	(1.8-5.7)	3.5	3.8*	(2.5-5.9)	3.1	5.2*	(3.5-7.8)
Bipolar I-II disorders	7.8	3.8*	(2.6-5.4)	6.6	4.0*	(2.9-5.4)	7.0	3.7*	(2.8-4.9)	6.2	3.5*	(2.6-4.7)	5.2	3.7*	(2.7-5.1)
Any mood disorder	24.9	4.0*	(3.2-5.1)	22.9	4.4*	(3.7-5.2)	21.2	3.3*	(2.7-4.1)	24.4	5.6*	(4.6-6.6)	19.7	5.4*	(4.4-6.7)
III. Impulse-control disorders															
Intermittent explosive disorder	4.1	1.6*	(1.1-2.5)	4.1	2.0*	(1.4-2.9)	4.0	1.9*	(1.3-2.9)	4.5	2.2*	(1.4-3.3)	4.1	2.4*	(1.7-3.4)
Attention-deficit/hyperactivity disorder	10.1	3.8*	(2.8-5.1)	10.3	3.8*	(2.6-5.7)	12.6	5.3*	(3.7-7.8)	8.6	4.2*	(3.1-5.8)	7.8	4.3*	(2.9-6.4)
Conduct disorder or ODD	4.7	2.8*	(1.5-5.4)	4.6	3.3*	(1.8-5.9)	5.3	3.6*	(1.9-6.9)	3.6	2.8*	(1.6-4.8)	3.4	2.8*	(1.4-5.6)
Any impulse-control disorder	18.0	2.7*	(2.1-3.4)	18.3	2.9*	(2.1-4.0)	20.6	3.4*	(2.5-4.7)	16.6	3.2*	(2.5-4.2)	15.3	3.2*	(2.4-4.3)
IV. Substance use disorders															
Alcohol abuse[2]	6.7	2.7*	(2.0-3.7)	4.8	2.4*	(1.6-3.7)	4.9	2.3*	(1.6-3.2)	4.9	2.0*	(1.4-2.7)	4.7	2.4*	(1.7-3.6)
Alcohol dependence with abuse	4.3	5.0*	(3.4-7.5)	3.4	4.7*	(2.6-8.5)	3.2	3.7*	(2.3-5.8)	2.8	2.8*	(1.7-4.8)	2.7	4.9*	(2.8-8.5)
Drug abuse[2]	3.4	3.2*	(1.8-5.6)	2.3	2.9*	(1.6-5.1)	2.5	2.8*	(1.6-5.2)	2.6	2.7*	(1.6-4.8)	2.3	3.3*	(1.7-6.2)
Drug dependence with abuse	1.2	3.5*	(1.5-8.4)	0.8	3.1*	(1.3-7.8)	1.0	3.7*	(1.7-8.3)	0.8	2.5*	(1.2-5.2)	0.8	4.2*	(1.8-10.2)
Any substance use disorder	16.3	2.8*	(2.3-3.5)	12.4	2.4*	(1.9-3.2)	11.2	1.9*	(1.4-2.4)	13.5	2.6*	(2.0-3.3)	11.8	2.6*	(2.0-3.5)
V. Any disorder															
Any disorder	53.2	3.6*	(2.9-4.4)	49.3	3.7*	(3.1-4.4)	47.8	3.2*	(2.7-3.8)	53.7	4.4*	(3.8-5.0)	46.4	4.1*	(3.4-5.0)
Exactly one	20.6	1.6*	(1.3-2.0)	19.8	1.6*	(1.3-2.0)	18.6	1.5*	(1.2-1.8)	22.6	2.0*	(1.7-2.3)	20.4	1.9*	(1.6-2.4)
Exactly two	12.6	2.2*	(1.7-2.8)	12.5	2.7*	(2.1-3.3)	12.1	2.4*	(1.9-3.0)	14.0	3.2*	(2.6-3.9)	12.2	3.5*	(2.9-4.2)
Three or more	20.0	5.4*	(4.6-6.4)	17.0	5.3*	(4.3-6.6)	17.0	4.6*	(3.7-5.7)	17.1	6.3*	(5.2-7.6)	13.8	6.9*	(5.5-8.7)

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

[1]Controlling for sex, age, race-ethnicity, education, marital status, occupation, poverty line, number of young, middle and old aged children

[2]With or without dependence

Table 5. Associations of 12-month sleep problems with 30-day role functioning in the NCS-R (n = 5692)[1]

	Outcomes									
	Daytime sleepiness		Days out of role		Quantity		Quality		Extreme effort	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
A. DIS										
Gross[2]	19.3*	(1.0)	4.0*	(0.4)	1.9*	(0.3)	1.3*	(0.2)	1.8*	(0.2)
Net[2]	-1.2	(1.2)	2.0*	(0.5)	0.8*	(0.3)	0.6*	(0.3)	0.7*	(0.3)
Pure[2]	10.4*	(2.9)	1.8	(1.2)	1.6*	(0.8)	1.0	(0.7)	1.7*	(0.8)
B. DMS										
Gross[2]	24.0*	(1.1)	3.6*	(0.3)	1.7*	(0.2)	1.2*	(0.1)	1.6*	(0.2)
Net[2]	6.1*	(1.6)	1.3*	(0.4)	0.5	(0.3)	0.4*	(0.2)	0.4	(0.3)
Pure[2]	18.1*	(2.1)	-0.4	(0.4)	0.1	(0.4)	-0.1	(0.2)	-0.5*	(0.2)
C. EMA										
Gross[2]	24.5*	(1.3)	3.2*	(0.3)	1.4*	(0.2)	1.0*	(0.2)	1.6*	(0.2)
Net[2]	7.3*	(1.3)	0.6	(0.4)	0.1	(0.3)	0.0	(0.3)	0.4	(0.2)
Pure[2]	12.4*	(2.6)	-0.5	(0.6)	0.1	(0.6)	0.4	(0.6)	0.0	(0.3)
D. NRS										
Gross[2]	34.4*	(1.0)	3.7*	(0.3)	2.0*	(0.3)	1.4*	(0.1)	2.0	(0.2)
Net[2]	29.8*	(1.2)	2.2*	(0.4)	1.5*	(0.3)	1.0*	(0.2)	1.5*	(0.2)
Pure[2]	42.2*	(1.7)	2.0*	(0.3)	1.6*	(0.3)	1.2*	(0.2)	1.4*	(0.2)

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

[1]Controlling for sex, age, race-ethnicity, education, marital status, occupation, number of young children

[2]Gross models include controls only for the socio-demographic variables listed in footnote 1. Net models additionally control for all the 12-month DSM-IV disorders included in Table 4. Pure models are estimated in the sub-sample of respondents who met criteria for none of the 12-month DSM-IV disorders.

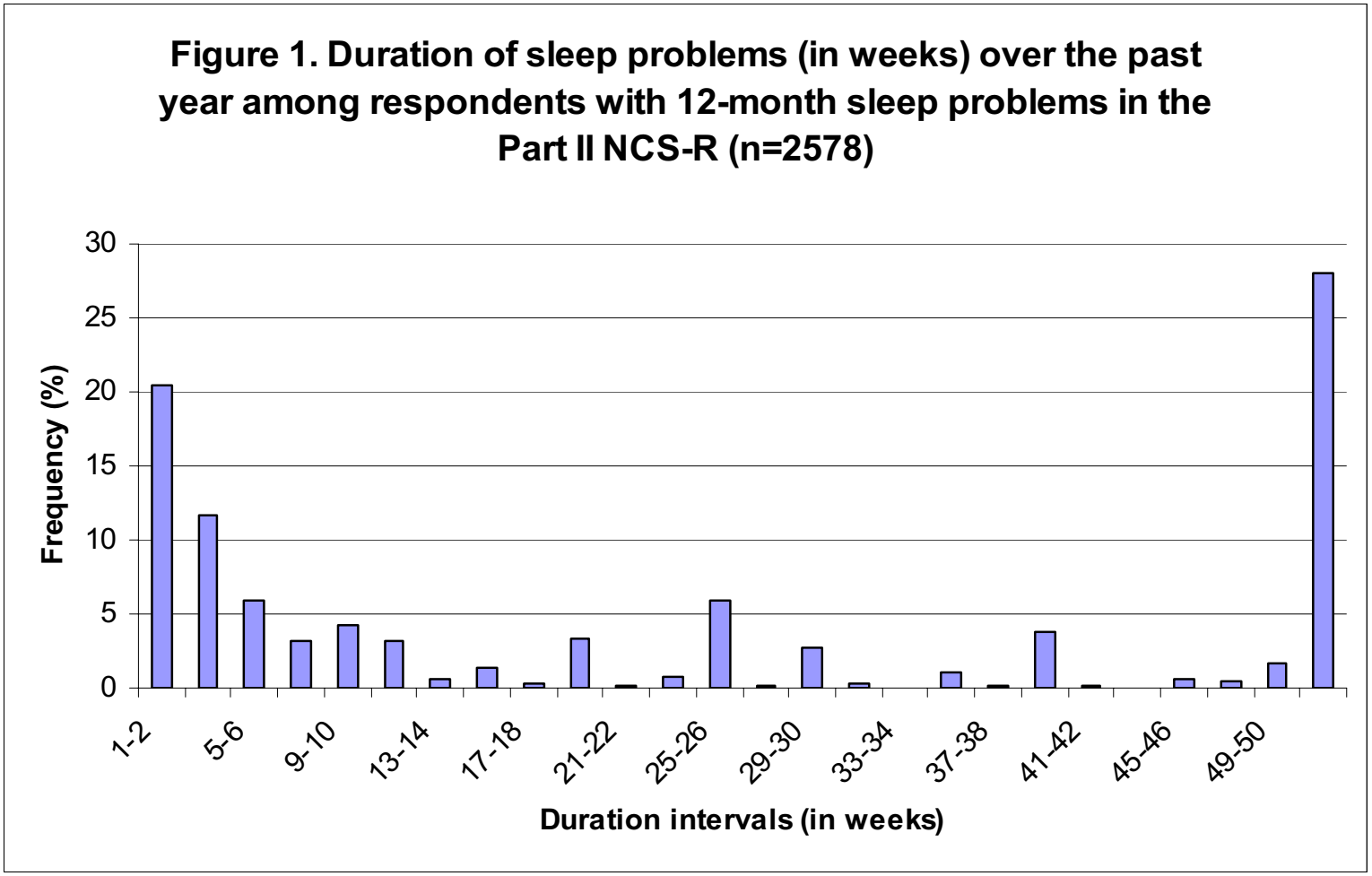


Figure 1. Duration of sleep problems (in weeks) over the past year among respondents with 12-month sleep problems in the Part II NCS-R (n=2578)
Frequency (%)
30
25
20
15
10
5
0
1-2
5-6
9-10
13-14
17-18
21-22
25-26
29-30
33-34
37-38
41-42
45-46
49-50
Duration intervals (in weeks)

Date completed: __ __ / __ __ / __ __ __ __

K6+

Provider: _____

Provider ID:

|___|___|___|___|___|___|___|___|

Please used gummed label if available

Patient or Client Identifier:

|__|__|__|__|__|__|__|__|__|__|__|__|

Surname:	
Other names:	
Date of Birth: ___ ___ / ___ ___ / ___ ___ ___ ___	Sex: Male ☐1 Female ☐2
Address:	

The following questions ask about how you have been feeling during the **past 30 days**. For each question, please circle the number that best describes how often you had this feeling.

Q1. During the past 30 days, about how often did you feel …	**All of the time**	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**
a. …nervous?	**1**	**2**	**3**	**4**	**5**
b. …hopeless?	**1**	**2**	**3**	**4**	**5**
c. …restless or fidgety?	**1**	**2**	**3**	**4**	**5**
d. …so depressed that nothing could cheer you up?	**1**	**2**	**3**	**4**	**5**
e. …that everything was an effort?	**1**	**2**	**3**	**4**	**5**
f. …worthless?	**1**	**2**	**3**	**4**	**5**

MENTAL HEALTH

Please turn over the page to continue

Q2. The last six questions asked about feelings that might have occurred during the past 30 days. Taking them altogether, did these feelings occur More often in the past 30 days than is usual for you, about the same as usual, or less often than usual? (If you never have any of these feelings, circle response option "4.")

More often than usual			About the same as usual	Less often than usual		
A lot	**Some**	**A little**		**A little**	**Some**	**A lot**
1	**2**	**3**	**4**	**5**	**6**	**7**

The next few questions are about how these feelings may have affected you in the past 30 days. You need not answer these questions if you answered "None of the time" to **all** of the six questions about your feelings.

Binding margin – do not write

Q3. During the past 30 days, how many days out of 30 were you totally unable to work or carry out your normal activities because of these feelings?

_____ **(Number of days)**

Q4. **Not counting the days you reported in response to Q3**, how many days in the past 30 were you able to do only half or less of what you would normally have been able to do, because of these feelings?

_____ **(Number of days)**

Q5. During the past 30 days, how many times did you see a doctor or other health professional about these feelings?

_____ **(Number of times)**

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
Q6. During the past 30 days, how often have physical health problems been the main cause of these feelings?	**1**	**2**	**3**	**4**	**5**

Thank you for completing this questionnaire.

Date completed: __ __ / __ __ / __ __ __ __

K10+

Provider: _____

Provider ID:

|___|___|___|___|___|___|___|___|

Please used gummed label if available

Patient or Client Identifier:

|__|__|__|__|__|__|__|__|__|__|__|__|

Surname:

Other names:

Date of Birth: ___ ___ / ___ ___ / ___ ___ ___ ___

Sex: Male ☐1 Female ☐2

Address:

The following questions ask about how you have been feeling during the **past 30 days**. For each question, please circle the number that best describes how often you had this feeling.

Q1. During that month, how often did you feel …	**All of the time**	**Most of the time**	**Some of the time**	**A little of the time**	**None of the time**
a. … tired out for no good reason?	**1**	**2**	**3**	**4**	**5**
b. …nervous?	**1**	**2**	**3**	**4**	**5**
c. …so nervous that nothing could calm you down?	**1**	**2**	**3**	**4**	**5**
d. …hopeless?	**1**	**2**	**3**	**4**	**5**
e. …restless or fidgety?	**1**	**2**	**3**	**4**	**5**
f. …so restless that you could not sit still?	**1**	**2**	**3**	**4**	**5**
g. …depressed?	**1**	**2**	**3**	**4**	**5**
h. …so depressed that nothing could cheer you up?	**1**	**2**	**3**	**4**	**5**
i. …that everything was an effort?	**1**	**2**	**3**	**4**	**5**
j. …worthless?	**1**	**2**	**3**	**4**	**5**

Please turn over the page to continue

Q2. The last ten questions asked about feelings that might have occurred during the past 30 days. Taking them altogether, did these feelings occur More often in the past 30 days than is usual for you, about the same as usual, or less often than usual? (If you never have any of these feelings, circle response option "4.")

More often than usual			About the same as usual	Less often than usual		
A lot	**Some**	**A little**		**A little**	**Some**	**A lot**
1	**2**	**3**	**4**	**5**	**6**	**7**

The next few questions are about how these feelings may have affected you in the past 30 days. You need not answer these questions if you answered "None of the time" to **all** of the ten questions about your feelings.

Binding margin – do not write

Q3. During the past 30 days, how many days out of 30 were you totally unable to work or carry out your normal activities because of these feelings?

_____ **(Number of days)**

Q4. **Not counting the days you reported in response to Q3**, how many days in the past 30 were you able to do only half or less of what you would normally have been able to do, because of these feelings?

_____ **(Number of days)**

Q5. During the past 30 days, how many times did you see a doctor or other health professional about these feelings?

_____ **(Number of times)**

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
Q6. During the past 30 days, how often have physical health problems been the main cause of these feelings?	**1**	**2**	**3**	**4**	**5**

Thank you for completing this questionnaire.

Psychological Medicine, 2002, **32**, 959–976. © 2002 Cambridge University Press
DOI: 10.1017/S0033291702006074 Printed in the United Kingdom

Short screening scales to monitor population prevalences and trends in non-specific psychological distress

R. C. KESSLER,[1] G. ANDREWS, L. J. COLPE, E. HIRIPI, D. K. MROCZEK, S.-L. T. NORMAND, E. E. WALTERS AND A. M. ZASLAVSKY

From the Department of Health Care Policy, Harvard Medical School, Division of Mental Disorders, Behavioral Research and AIDS, National Institute of Mental Health, Department of Psychology, Fordham University and Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; and Clinical Research Unit for Anxiety Disorders, School of Psychiatry, University of New South Wales, NSW, Australia

ABSTRACT

Background. A 10-question screening scale of psychological distress and a six-question short-form scale embedded within the 10-question scale were developed for the redesigned US National Health Interview Survey (NHIS).

Methods. Initial pilot questions were administered in a US national mail survey ($N = 1401$). A reduced set of questions was subsequently administered in a US national telephone survey ($N = 1574$). The 10-question and six-question scales, which we refer to as the K10 and K6, were constructed from the reduced set of questions based on Item Response Theory models. The scales were subsequently validated in a two-stage clinical reappraisal survey ($N = 1000$ telephone screening interviews in the first stage followed by $N = 153$ face-to-face clinical interviews in the second stage that oversampled first-stage respondents who screened positive for emotional problems) in a local convenience sample. The second-stage sample was administered the screening scales along with the Structured Clinical Interview for DSM-IV (SCID). The K6 was subsequently included in the 1997 ($N = 36116$) and 1998 ($N = 32440$) US National Health Interview Survey, while the K10 was included in the 1997 ($N = 10641$) Australian National Survey of Mental Health and Well-Being.

Results. Both the K10 and K6 have good precision in the 90th–99th percentile range of the population distribution (standard errors of standardized scores in the range 0·20–0·25) as well as consistent psychometric properties across major sociodemographic subsamples. The scales strongly discriminate between community cases and non-cases of DSM-IV/SCID disorders, with areas under the Receiver Operating Characteristic (ROC) curve of 0·87–0·88 for disorders having Global Assessment of Functioning (GAF) scores of 0–70 and 0·95–0·96 for disorders having GAF scores of 0–50.

Conclusions. The brevity, strong psychometric properties, and ability to discriminate DSM-IV cases from non-cases make the K10 and K6 attractive for use in general-purpose health surveys. The scales are already being used in annual government health surveys in the US and Canada as well as in the WHO World Mental Health Surveys. Routine inclusion of either the K10 or K6 in clinical studies would create an important, and heretofore missing, crosswalk between community and clinical epidemiology.

[1] Address for correspondence: Dr R. C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.

INTRODUCTION

Dimensional scales of non-specific psychological distress have been used in community epidemiological surveys since the end of World War II, beginning with the 20-item Health Opinion Survey in the Stirling County Study (MacMillan, 1957; Leighton, 1975) and the 22-item Langner Scale in the Midtown Manhattan Study (Langner, 1962; Srole *et al.* 1962). Although originally used as first-stage screens to target respondents with broadly defined emotional problems for more in-depth clinical assessment, these dimensional scales came to be used without clinical follow-up in later surveys (Myers *et al.* 1975). Controversy regarding the appropriate cut-point for case thresholds on these scales in community surveys (Seiler, 1973) led in later surveys to scale scores being reported primarily in dimensional terms (e.g. mean scores) rather than in terms of proportions of respondents screening positive (Pearlin *et al.* 1981).

Dimensional scales continue to be widely used to screen for mental illness in primary care (Coyne *et al.* 2001) and to assess symptom severity and treatment effectiveness in clinical studies (Rush *et al.* 2000). However, influenced by the widely published results of the Epidemiological Catchment Area Study, dimensional screening scales went out of vogue in community psychiatric epidemiology beginning in the early 1980s (Robins & Regier, 1991). Fully structured research diagnostic interviews administered by lay interviewers have become the standard measures of psychopathology in community epidemiological surveys since that time. A number of such structured diagnostic interviews now exist, including the Diagnostic Interview Schedule (DIS) (Robins *et al.* 1981), the Composite International Diagnostic Interview (CIDI) (Robins *et al.* 1988), the PRIME-MD (Spitzer *et al.* 1994); and the Mini-International Neuropsychiatric Interview (MINI; Sheehan *et al.* 1998).

We now know, based on the use of fully structured diagnostic interviews in a number of large community epidemiological surveys, that up to half the general population meet criteria for one or more lifetime ICD or DSM disorders and up to one-fifth carry a DSM or ICD diagnosis at any one point in time (Kessler *et al.* 1994; Kessler & Frank, 1997). Although the published reports of these high prevalence estimates were initially met with a good deal of scepticism, subsequent clinical reappraisal studies showed that they are accurate (Wittchen, 1994; Kessler *et al.* 1998), but that many community cases have considerably less severe disorders than those of cases in treatment (Kessler *et al.* 2001).

The finding that clinical severity is related to treatment is, of course, not surprising. However, given the high proportion of people in the population who meet criteria for a mental disorder in relation to the societal resources available for treatment, policy-orientated interpreters of the epidemiological evidence have called for (Regier *et al.* 2000), and in some cases created (National Advisory Mental Health Council, 1993), distinctions to be made between people with severe and less severe mental disorders in an effort to define medical necessity for policy planning purposes. For example, the US Substance Abuse and Mental Health Service Administration, which administers Block Grants to States to fund public mental health services for low-income people who are not otherwise insured, limits coverage to cases defined as having a serious mental illness (SMI). The criteria for SMI require not only a DSM diagnosis but also specified indicators of severity that characterize fewer than one-third of the people in the US population who meet criteria for a current DSM-III-R disorder (Kessler *et al.* 1996).

Dimensional measures of non-specific psychological distress have come to take on new importance in the context of this movement to distinguish community cases based on severity rather than purely on diagnosis. The architects of the redesigned US National Health Interview Survey (NHIS), being aware of this fact, opted to include a short dimensional measure of non-specific psychological distress rather than disorder-specific diagnostic measures in the core of the new NHIS interview. The 'core' of the NHIS is the part of the interview that is administered to all respondents every year. Given the severe time constraints in the NHIS core, this dimensional measure was required to have no more than six to eight questions. Although a number of non-specific distress scales

exist and have been used for many years in community surveys (Gurin *et al.* 1960; Dohrenwend *et al.* 1980; Derogatis, 1983), only a few of them are brief enough to meet this time requirement (Pearlin *et al.* 1981; Ware & Sherbourne, 1992) and none was developed using modern psychometric methods to maximize precision in the clinical range of the population distribution (van der Linden & Hambleton, 1997). Based on these considerations, the decision was made to develop a new screening scale for use in the redesigned NHIS.

The conceptualization of this task relied importantly on the work of Dohrenwend and his colleagues (Dohrenwend *et al.* 1980; Link & Dohrenwend, 1980). Their review of screening scales of nonspecific psychological distress showed that these scales typically include questions about a heterogeneous set of cognitive, behavioural, emotional and psychophysiological symptoms that are elevated among people with a wide range of different mental disorders. However, despite this heterogeneous content, the vast majority of the symptoms in these scales have high factor loadings on a first principal factor. People with a wide range of mental disorders typically have high scores on this core dimension of non-specific distress. Based on this result, we sought to measure this core dimension of non-specific psychological distress in our new scale. Due to the fact that the developers of the NHIS were unsure about how much space they had available for this scale, we decided to develop both 10-question and 6-question versions, optimizing the precision of the scale by using modern psychometric methods to select the questions with the maximum precision in the clinical range of the scale. Based on the fact that no more than 10%, and probably closer to 6%, of the US population are estimated to meet criteria for serious mental illness in a given year (Kessler *et al.* 1996), we decided at the onset to seek maximum precision in the 90th–99th percentile range of the general population distribution.

METHOD

Data

Data are reported from five community surveys. The first two were pilot surveys used sequentially to develop the screening scales from a larger pool of questions. The third was a calibration survey in which the screening scales were administered along with clinical interviews. The last two were large government health surveys carried out in the US and Australia in which we cross-validated the pilot results.

The mail pilot survey

The first pilot survey was carried out in a nationally representative mail sample of the continental US that included oversamples of people with Hispanic surnames and of people living in zip codes with high concentrations of Blacks. A total of 1403 respondents ages $\geqslant 18$ completed the questionnaire, for a response rate of 54·8%. The methods and procedures of the mail pilot survey were approved by the Human Subjects Committee of the Institute for Social Research at the University of Michigan.

The telephone pilot survey

A revised set of scale questions, based on analysis of the mail pilot survey, was administered to a nationally representative telephone sample of 1574 respondents ages $\geqslant 25$ in the continental US as part of the pilot work for the MacArthur Foundation Midlife Development in the US (MIDUS) survey (Ryff *et al.* 2002). The co-operation rate in households where we were able to make contact was 52·9%, while the estimated response rate (assuming that numbers never contacted were eligible) was 40·4%. Verbal informed consent was obtained before beginning the interviews. The methods and procedures of the telephone pilot survey were approved by the Human Subjects Committee of the Harvard Medical School.

The clinical reappraisal survey

Both a 10-question scale, which we refer to as the K10, and a 6-question short-form scale, which we refer to as the K6, were developed from the results of the pilot surveys using the methods of Item Response Theory (IRT) (Hambleton *et al.* 1991). A clinical reappraisal survey of these scales was carried out in a two-stage convenience sample. The first stage administered a brief telephone screening interview to a convenience sample of 1000 respondents ages

Table 1. *Sources of initial item pool for K10 and K6 scales*

Name of scale	Source	Questions *N*
Psychiatric Epidemiology Research Interview Demoralization Scale	Dohrenwend *et al.* (1980)	27
Carroll Depression Scale	Carroll *et al.* (1981)	52
Health Opinion Survey	MacMillan (1957)	20
Taylor Manifest Anxiety Scale	Taylor (1953)	50
Stimulus-Response Inventory	Endler *et al.* (1962)	62
Self-Rating Depression Scale	Zung (1963)	20
Anxiety Status Inventory	Zung (1971)	20
Self-Rating Anxiety Scale	Zung (1971)	20
Beck Depression Inventory	Beck *et al.* (1961)	21
Depression Adjective Checklist	Lubin *et al.* (1967)	32
General Well-Being Scale – HANES	Fazio (1977)	20
Symptom Checklist-90	Derogatis (1983)	90
Center for Epidemiologic Studies – Depression Scale	Radloff (1977)	20
General Well-Being Scale – Rand	Ware *et al.* (1979)	46
Gurin Scale	Gurin *et al.* (1960)	20
General Health Questionnaire	Goldberg (1972)	30
22-Item Screening Scale	Langner (1962)	22
State-Trait Anxiety Inventory	Hodges & Spielberger (1969)	40
Total		612

$\geqslant 18$ with listed phone numbers in the Boston Metropolitan Area. Verbal consent was obtained for this first-stage interview after informing respondents that they might also be invited to participate in a second stage in-person interview.

The second-stage interview was carried out face-to-face in the homes of a subsample of 155 first-stage respondents, oversampling those with perceived mental health problems. This interview administered the K10 followed by the 12-month version of the Structured Clinical Interview for DSM-IV (SCID) (First *et al.* 1997). The purpose of this study was to determine whether the K10 is a useful screen for the US Substance Abuse and Mental Health Service Administration definition of 12-month SMI. This definition requires a 12-month DSM-IV disorder, along with a GAF score < 60 for the worst month in that 12-month period. In order to map the K10 to this same time period, respondents in the second-stage interview were asked to answer the K10 questions for the 1 month during the past year when they had the most severe and persistent emotional distress rather than for the month before the interview. The second-stage interview data were weighted to match the joint distribution from the 1997 NHIS of age, sex, education, and a coarse four-category classification of scores on the 6-question scale. Written consent was obtained before beginning the interview. The methods and procedures of the clinical reappraisal survey were approved by the Human Subjects Committee of the Harvard Medical School.

The NHIS

The K6 was subsequently included in the Sample Adult (age $\geqslant 18$) questionnaire in the 1997 ($N = 36\,116$) and 1998 ($N = 32\,440$) NHIS. The NHIS is an annual nationally representative face-to-face household survey based on a multi-stage clustered area probability sample of the United States carried out by the US government. Blacks and Hispanics are oversampled. The response rate was 89·0% in 1997 and 83·8% in 1998 (National Center for Health Statistics, 2000*a*, 2000*b*). The NHIS Public Use data-tape was obtained and the data were weighted to approximate the Census distribution on the cross-classification of a range of sociodemographic variables. These data were analysed to cross-validate the pilot IRT results.

The NSMHWB

The K10 was included in the 1997 NSMHWB (Andrews *et al.* 2001). The NSMHWB was a nationally representative face-to-face survey of 10 641 households based on multi-stage clustered area probability sample of the Australian population. One member of each sample household aged $\geqslant 18$ was randomly selected for interview.

Table 2. *Final K10 and K6 item pool*

During the last 30 days, about how often did …*
1 Depressed mood
(a) … you feel unhappy?
(b) … you feel sad or blue?
(c) … you feel depressed? (10)
(d) … you feel so depressed that nothing could cheer you up? (6) (10)†
(e) … you feel hopeless? (6) (10)‡
2 Anhedonia
(a) … you feel that nothing was worthwhile anymore?
(b) … you lose interest in the people or things you usually care about?
3 Eating
(a) … you have a much bigger appetite than usual?§
(b) … you have a much smaller appetite than usual?§
4 Sleep
(a) … you have trouble falling asleep or staying asleep?§
(b) … you sleep much more than usual?§
5 Motor agitation
(a) … you feel restless or fidgety? (6) (10)
(b) … you feel so restless that you could not sit still? (10)
6 Motor retardation
(a) … your thoughts come more slowly than usual?
(b) … you feel like everything was happening in slow motion?
7 Fatigue
(a) … you feel tired out for no good reason? (10)
(b) … you feel that everything was an effort? (6) (10)
(c) … you feel full of energy?
8 Worthless guilt
(a) … you feel worthless? (6) (10)
(b) … you feel ashamed?‖
(c) … you feel guilty?‖
(d) … you feel inferior or not as good as other people?
9 Concentration
(a) … you have trouble making simple decisions?
(b) … you have trouble keeping your mind on what you were doing?
10 Death
(a) … you have thoughts of death or dying?
(b) … you have thoughts of killing yourself?
11 Anxiety
(a) … you feel anxious?
(b) … you feel nervous? (6) (10)
(c) … you feel so nervous that nothing could calm you down?(10)
(d) … you get upset by little things?§
(e) … you feel fearful?
12 Worry
(a) … you feel worried about things that were not really important?
(b) … you worry about things that were not likely to happen?§
13 Motor tension
(a) … you feel physically tense or shaky?
(b) … your muscles feel tense, sore, or aching?§
14 Hypersensitivity
(a) … your heart pound or race without exercising?§
(b) … your mouth feel dry?§
(c) … you feel short of breath without exercising?§
(d) … you have indigestion or an upset stomach?§
(e) … you have trouble swallowing?§
(f) … your hands feel sweaty or clammy?§
(g) … you feel dizzy?
(h) … your face feel hot and flushed?§
15 Vigilance
(a) … you feel keyed up or an edge?
(b) … you feel irritable?
(c) … you feel angry?‡
(d) … you feel resentful?‡
16 Positive affect
(a) … you feel in a really good mood?§
(b) … you feel happy?§

(6) Questions included in the final 6-item scale.
(10) Questions included in the final 10-item scale.
* The response options used in the mail pilot survey were most of the time, some of the time, a little of the time, and none of the time. All later surveys added the response option all of the time.
† The wording used in the studies reported here was '… so sad that …'. This wording was subsequently changed to '… so depressed that …' to ensure a nesting with responses to question 1(c).
‡ Questions added in the telephone pilot survey.
§ Deleted after mail pilot survey.
‖ A single question about being ashamed or guilty was used in the mail pilot survey. The questions were separated in the telephone pilot survey.

Written informed consent was obtained before beginning the interviews. The response rate was 78%. The sample data were weighted to match the joint distribution of age and sex in the Australian national census. These data were analysed to cross-validate the pilot IRT results.

Measures

An initial pool of 612 questions from existing screening scales (Table 1) was reduced to 235 by discarding redundant and obviously unclear questions. This reduced set was sorted into content domains and rewritten in a format that asked respondents how often they experienced each symptom over a 30-day recall period. Because of their dominance in the item pool, further scale development focused on the 15 domains represented in the DSM-III-R diagnoses of major depression (MD) and generalized anxiety disorder (GAD) plus the positive affect domain. Following Converse & Presser (1986), the questions in these domains were submitted to an expert advisory panel of survey researchers who were asked to rate each question for clarity of wording. Only questions consistently rated as being clear were retained. The

resulting test item pool consisted of 45 questions (Table 2). These questions were administered in the mail pilot survey using a four-category response scale (most of the time, some of the time, a little of the time, never). The mail pilot survey also assessed basic sociodemographic variables.

Based on the results of the mail pilot survey, the item pool was revised to retain 28 of the original 45 questions. In addition, one of the 28 questions (feeling ashamed or guilty) was split into two (feeling ashamed, feeling guilty) and three new questions were added to increase precision in the depressed mood (feeling hopeless) and vigilance (feeling angry, feeling resentful) domains. The resulting 32 questions were used in the telephone pilot survey. The response category 'all of the time' was added in the telephone pilot survey in an effort to increase precision at the upper end of the distribution. The final K10 and K6 scales were generated from analysis of the telephone pilot survey. Only the questions in these scales were included in later phases of data collection.

Analysis

Evaluating dimensionality

Analysis of the mail and telephone pilot surveys began by carrying out factor analyses of the four-category (mail survey) or five-category (telephone survey) variables based on matrices of Pearson correlations using the FACTOR procedure in SAS 8 (SAS Institute, 1999). Parallel analyses based on matrices of polychoric correlations, which allow for non-linear monotonic relationships between pairs of variables, were also carried out but are not reported here because they yielded the same substantive results. After documenting the unidimensionality of the variable set with these initial factor analyses, we replicated the factor analyses at the level of the response categories within variables. This was done by converting responses into a series of either three (mail survey) or four (telephone survey) ordered dichotomies (e.g. 1 *v.* 2–4, 1–2 *v.* 3–4 and 1–3 *v.* 4 on the four-category response scale). Each of the dichotomies created in this way is referred to below as an 'item'. The 45 questions in the mail pilot survey generated 135 items (45×3), while the 32 questions in the telephone pilot survey generated 128 items (32×4). Tetrachoric correlation matrices were created from these items and factor analysed using the TESTFACT program (Scientific Software International, Inc., 1998) to select items for the unidimensional scale with high loadings (at least 0·4 in the mail survey and at least 0·5 in the telephone survey) on the first unrotated principal factor and low relative loadings on the second principal factor (a ratio of factor 1 to factor 2 loading of at least 2·0). The rationale for analysing the data at the item-level rather than at the variable-level is described in the next subsection.

Item response models

The BILOG-MG program (Scientific Software International, Inc., 1996) was used to estimate item response theory (IRT) models for items that passed the unidimensionality test. Unlike classical psychometric test theory models, IRT models allow the researcher to evaluate the contribution of each item to the sensitivity of the total scale in the severity range of the distribution that is most relevant for purposes of scale development. As noted in the introduction, in the present case this is the 90th–99th percentile range of the population distribution.

The IRT analysis was based on the conventional one- and two-parameter IRT logistic regression models for binary scale items (van der Linden & Hambleton, 1997). The two-parameter model is given by:

$$P_{ij}(\mathrm{TPD}) = [1 + e^{-a_j(\mathrm{TPD}_i - b_j)}]^{-1}. \qquad (1)$$

The outcome variable in this model, $P_{ij}(\mathrm{TPD}_i)$, is the probability that respondent i will endorse binary item j as a function of his or her underlying true score on the dimension of nonspecific psychological distress (TPD). The slope, a_j, which we refer to as the sensitivity parameter, measures the steepness of the logistic curve at the point where the probability of endorsing item j is 0·5. A steep curve means that the item has strong discriminating ability at the point on the curve where it has maximum information value. The intercept, b_j, which we refer to as the severity parameter, is the point on the TPD distribution at which the probability of endorsing item j is 0·5. The one-parameter model differs from the two-parameter model in that the sensitivity parameter is constrained to be constant across items.

In both the one-parameter and two-parameter models, all selected items, including those based on different cutpoints of the same question, were analysed as if they were independent dichotomies. Treating the information obtained at different points on the response scale in this fashion is a sensible way to combine information across the scale, just as we might combine distinct dichotomous questions that are sensitive at different degrees of severity. However, items based on the same question form a perfect Guttman scale and consequently are not independent. As we do not model this dependence, our procedure is not based on a full probability model. The likelihoods for these data under the IRT models might consequently be more precisely called pseudo-likelihoods. As a result, while the parameter estimates are meaningful, the standard errors of these estimates are biased downwards due to the dependence among related items.

Items considered for inclusion in the final scale were required to have standardized severity parameters of 0·8 (i.e. eight-tenths of a standard deviation above the mean on the TPD distribution) or greater in the mail pilot survey and 1·2–2·3 (i.e. the 90th–99th percentile range of the standardized distribution) in the telephone pilot survey. This requirement was imposed in order to select items that had maximum sensitivity in the target severity range. A lower minimum inclusion value was imposed in the mail pilot survey for two reasons. First, because of the low response rate in the mail pilot survey, we wanted to be liberal in bringing forward items for further testing in the subsequent telephone pilot survey. Secondly, we became aware in the course of analysing the mail pilot data that the response categories should be increased from four to five. We were consequently aware, in analysing the mail pilot data, that the slope for the highest response category in the data obtained from that survey would be disaggregated in later data collections. We also recognized that more basic changes in responding might occur when response options were increased. Because of these uncertainties, we wanted to err on the side of being over-inclusive in selecting items for further investigation.

The items selected for further analysis in the mail pilot survey were also required to have standardized sensitivity parameters of $\geqslant 1{\cdot}0$. This requirement was imposed after carrying out preliminary analyses that showed the two-parameter model (i.e. a model that allow separate sensitivity parameters for each item) to provide a significantly better fit than the one-parameter model (i.e. a model that constrained the estimated sensitivity parameter to be constant across items). The selection of 1·0 as the minimum required sensitivity is conventional in the IRT literature.

Finally, all the items selected for further analysis in the mail pilot survey were required to have consistent (relative to other items) severity values across sociodemographic subsamples. This requirement was imposed to guarantee that scale scores have the same meaning in all major segments of society. This is a critical issue, based on previous research that has documented substantial subsample differences in severity for questions dealing with psychological distress. For example, Schaeffer (1988) showed that the IRT item severities for responses to the question 'How often did you cry?' are dramatically higher for men than women. This means that crying is an indicator of more severe distress among men than women. A distress scale that includes a question about frequency of crying will consequently overestimate the magnitude of sex differences in true distress.

The subset of items that fulfilled the above criteria was used to generate a maximum likelihood estimate of TPD (i.e. the estimate that has the highest likelihood of producing the respondent's observed pattern of item responses based on the parameter estimates in Equation (1)). Alternative scales to measure TPD made up of different subsets of items that fulfilled the above criteria were evaluated by using maximum likelihood to compute a graph known as the test information curve (TIC). Each point on the TIC is roughly equal to $1/[\text{S.E.}(\text{TPD})]^2$, where S.E. is the standard error of the estimate at that point on the scale distribution (Hambleton *et al.* 1991). A high test information value (low standard error) in a particular range of the distribution means that changes in observed scores in that range are strongly related to changes in true TPD scores. The height and shape of the TIC can be changed by adding or subtracting items that differ in severity (to change the skew of the curve) and sensitivity (to change the height of the curve).

This exercise was carried out with the telephone pilot data to construct scales that had maximum height in the top 90th–99th percentile range of the population distribution. Even though questions selected for inclusion in the various versions of the scales were chosen primarily on the basis of having eligible items in the target severity range, the TIC curves that were constructed to compare the performance of the alternative versions used information for all items in selected questions that passed the test of having consistent severity values across sociodemographic subsamples. This was done because items can contribute to precision in the target range even when their severities are outside this range (Thissen & Wainer, 2001).

It is important to note that IRT methods exist that would have allowed us to analyse the polychotomous responses to each of our scale questions as a single unit rather than as a set of ordered dichotomies (van der Linden & Hambleton, 1997). However, these methods require that the slopes of the implicit ordered dichotomies are constrained to be equal within each variable (i.e. the four slopes of the ordered dichotomies made up of responses to one five-category variable are estimated but constrained to have a single value). Preliminary analysis of the mail pilot survey results showed that this assumption was violated. This conclusion was confirmed in subsequent analyses of the telephone pilot data, the NHIS data, and the NSMHWB data. As a result, the ordered dichotomy approach was used in analysis. This is also why the factor analyses described above were carried out at the item-level as well as at the variable-level.

The clinical reappraisal survey

The associations of estimated TPD scores with DSM-IV/SCID diagnoses were evaluated in the clinical reappraisal sample using Receiver Operating Characteristic (ROC) curve analysis (Hanley & McNeil, 1982). ROC analysis displays the relationship between the sensitivity and (1-specificity) of each value of a dimensional screening scale in predicting a dichotomous clinical outcome. Two such outcomes were examined in our analysis: (*a*) a 12-month DSM-IV/SCID diagnosis of either an anxiety disorder, a mood disorder, or a nonaffective psychosis with a Global Assessment of Functioning (GAF) (Endicott *et al.* 1976) score in the range 0–70; and (*b*) a similar DSM-IV/SCID diagnosis with a GAF score in the range 0–50. The area under each ROC curve was calculated to evaluate the accuracy of the scales. This area can be interpreted as the probability that a randomly chosen case and a randomly chosen noncase would be correctly distinguished based on their screening scale scores.

Cross-validation in the NHIS and NSMHWB

The external validity of the pilot results was examined by re-estimating the IRT models in the NHIS and NSMHWB data. IRT parameters were examined to evaluate the severity, sensitivity and consistency of the severity of items. IRT test information curves for TPD were also compared to evaluate the relative sensitivities of the K10 and K6 in the target severity range. Finally, a score for each respondent was calculated by summing the item sensitivity parameters for each endorsed item. When the item parameters are fixed, as they would be when results in a benchmarking survey are used to define the metric of the scale in later surveys, this score is a sufficient statistic for the person parameter (TPD). This being the case, the summed sensitivity parameter score is a one-to-one monotonic transformation of the estimated TPD, but is much easier to calculate. A figure showing the relationship between TPD and the summed sensitivity parameter scores was calculated to show this transformation.

RESULTS

The mail pilot survey

Dimensionality

Principal axis factor analysis of a Pearson correlation matrix among the 45 questions included in the mail pilot survey found that the first unrotated principal factor had an eigenvalue of 15·7 compared to 2·1 for the second principal factor. All 45 questions had higher factor loadings on the first than second unrotated factor. In addition, tetrachoric factor analysis indicated that 106 of the 135 items made up from the 45 mail pilot survey questions had a factor loading of at least 0·4 on the first unrotated

Table 3. *Summary of mail pilot study IRT results*

Domain	Initial number of		Tests passed*				Final number of questions
	Questions	Items	Dimensionality	Severity	Sensitivity	Consistency	
1 Depressed mood	4	12	12	8	8	5	4
2 Anhedonia	2	6	5	4	4	4	2
3 Eating	2	6	2	1	0	0	0
4 Sleep	2	6	3	2	0	0	0
5 Motor agitation	2	6	6	4	4	2	2
6 Motor retardation	2	6	6	4	4	2	2
7 Fatigue	3	9	8	4	3	2	2
8 Worthless guilt	3	9	7	7	6	4	3
9 Concentration	2	6	6	4	4	4	2
10 Death	2	6	4	4	4	4	2
11 Anxiety	5	15	10	9	9	5	4
12 Worry	2	6	5	3	2	1	1
13 Motor tension	2	6	5	2	2	2	1
14 Hypersensitivity	8	24	21	18	1	1	1
15 Vigilance	2	6	4	4	2	2	2
16 Positive affect	2	6	2	0	0	0	0
Total	(45)	(135)	(106)	(78)	(53)	(38)	(28)

* Only items that passed the dimensionality test were evaluated for severity. Only items that passed the severity test were evaluated for sensitivity. All items were evaluated for consistency, but only those that also passed the sensitivity test are reported in this Table.

principal factor (F1) and a F1/F2 ratio of at least 2·0. These items were considered to pass the unidimensionality test. Items in the eat, sleep, and positive affect domains were least likely to pass this test (Table 3), with the positive affect domain completely eliminated because of failure to pass this test.

IRT analysis

One-parameter and two-parameter logistic models were estimated with the 106 items that passed the unidimensionality test. The two-parameter model provided a significantly better fit ($\chi^2_{106} = 973{\cdot}9$, $P < 0{\cdot}001$). Seventy-eight of the 106 items in the two-parameter model had severity values of $\geqslant 0{\cdot}8$ (Table 3), 53 of which had sensitivity parameters of $\geqslant 1{\cdot}0$. Thirty-eight of these 53 items had severity parameters that were consistent (i.e. did not vary significantly at the 0·05 level, two-sided tests) across sociodemographic subsamples defined in terms of age, sex, race-ethnicity, or education. These 38 items came from 28 of the original survey questions. Only these 28 questions were carried forward into the subsequent telephone second pilot survey. The positive affect, eating, and sleep domains were entirely eliminated because of failing at least one of these tests.

The telephone pilot survey

Dimensionality

Principal axis factor analysis of a Pearson correlation matrix among the 32 questions included in the mail pilot survey found that the first unrotated principal factor had an eigenvalue of 11·5 compared to 1·1 for the second principal factor. All 32 questions had higher factor loadings on the first than second unrotated factor. In addition, tetrachoric factor analysis indicated that 93 of the 128 items generated from the 32 questions passed the unidimensionality test using the criteria described in the Analysis section of this paper.

IRT analysis

One-parameter and two-parameter logistic models were estimated with the 93 items that passed the unidimensionality test. The two-parameter model provided a significantly better fit ($\chi^2_{93} = 607{\cdot}8$, $P < 0{\cdot}001$). Fifty-five of the 93 items in the two-parameter model had severity values in the 1·2–2·3 range (Table 4), 41 of which had sensitivity parameters of $\geqslant 1{\cdot}0$. Thirty-four of these 41 items had severity parameters that were consistent across sociodemographic sub-

Table 4. *Summary of telephone pilot study IRT results*

Domain	Initial number of		Tests passed*				Final number of questions
	Questions	Items	Dimensionality	Severity	Sensitivity	Consistency	
1 Depressed mood	5	20	17	10	10	8	4
2 Anhedonia	2	8	7	5	5	4	2
3 Eating	0	0	—	—	—	0	0
4 Sleep	0	0	—	—	—	0	0
5 Motor agitation	2	8	5	3	3	3	2
6 Motor retardation	2	8	5	5	0	0	0
7 Fatigue	2	8	7	3	3	3	3
8 Worthless guilt	4	16	14	10	8	6	4
9 Concentration	2	8	4	3	0	0	0
10 Death	2	8	4	1	0	0	0
11 Anxiety	4	16	12	6	5	3	2
12 Worry	1	4	3	2	1	1	1
13 Motor tension	1	4	3	2	1	1	1
14 Hypersensitivity	1	4	1	0	—	0	0
15 Vigilance	4	16	11	5	5	5	4
16 Positive affect	0	0	—	—	—	0	0
Total	(32)	(128)	(93)	(55)	(41)	(34)	(23)

* Only items that passed the dimensionality test were evaluated for severity. Only items that passed the severity test were evaluated for sensitivity. All items were evaluated for consistency, but only those that also passed the sensitivity test are reported in this Table.

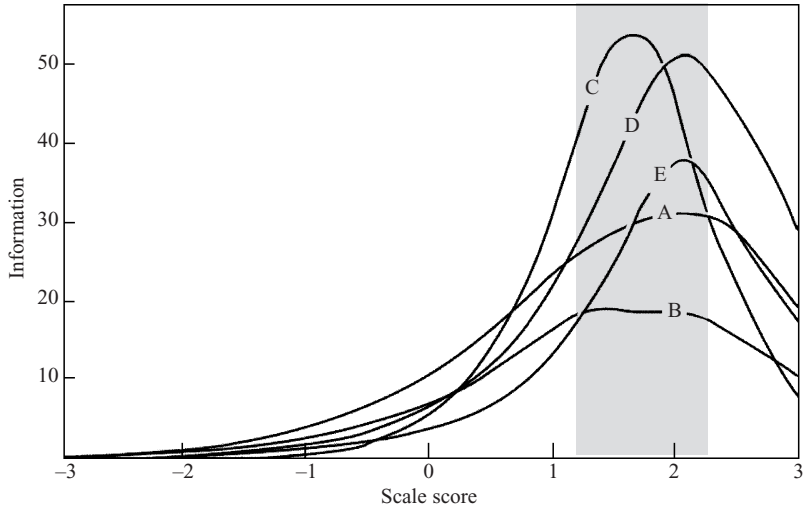


FIG. 1. Test information curves for the 10-question and 6-question scales in the telephone pilot survey, the NHIS and the NSMHWB. (A, Telephone pilot 10-question; B, telephone pilot 6-question; C, NHIS 6-question; D, NSMHWB 10-question; E, NSMHWB 6-question.)

samples. These 34 came from 23 different questions. These 23 questions were the focus of subsequent scale construction efforts.

A TIC was generated for the scale that included not only these 34 most informative items, but all other items that passed the dimensionality, sensitivity, and consistency tests in the questions that generated the 23 most informative questions. These additional items were added because, as noted above in the section on analysis methods, they can contribute to precision in the target range even though their severities are outside this range. Decomposition was then used to generate a separate TIC for each logically possible 10-question and 6-question subscale of this larger scale, again using all eligible items in the questions to construct the TIC.

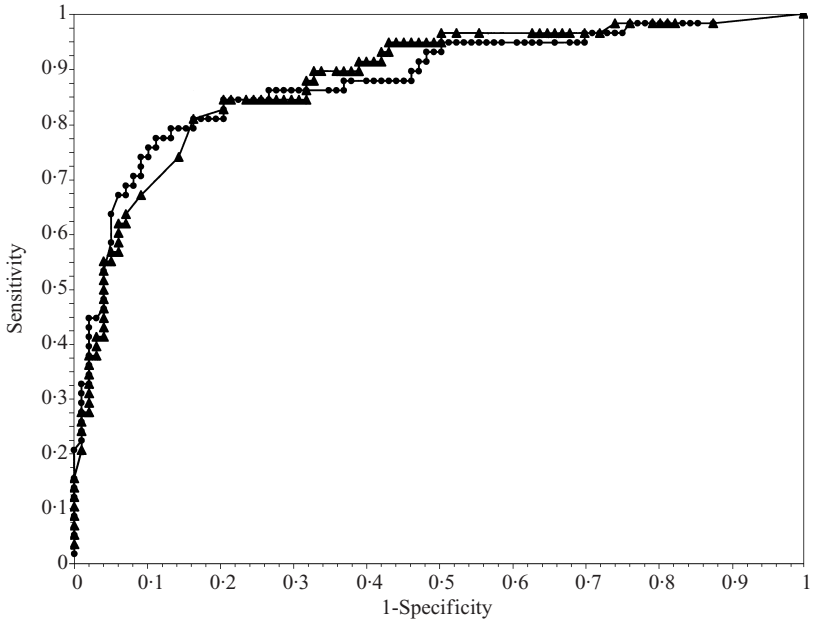


FIG. 2. Receiver operator characteristic (ROC) curves for the K10 and K6 predicting DSM-IV/SCID disorders with GAF scores in the range 0–70. (Area under the ROC curve: ●, K10 0·876; ▲, K6 0·879.)

The final K10 and K6 scales were selected based on inspection of these different TIC plots to maximize information in the top 90th–99th percentile range of the population distribution (Fig. 1). As we suspected that a 10-question scale was too long for use in the NHIS, the questions included in the final K10 were selected to contain three ordered pairs (questions 1c–d, 5a–b and 11b–c in Table 2), so that respondents who responded 'never' to the first question in the pair could be skipped over the second question. Given the distribution of responses in the pilot samples, we estimated that this use of ordered pairs would result in only eight of the 10 questions being administered to an average respondent in the general population. Both the K10 ($a = 0{\cdot}93$) and the K6 ($a = 0{\cdot}89$) had excellent internal consistency reliability (Cronbach's alpha) in the telephone pilot sample.

The clinical reappraisal survey

ROC analysis, using the OUTROC option of the LOGISTIC procedure in SAS 8, was used to evaluate the ability of the two scales to discriminate between community cases and non-cases of DSM-IV disorders (SAS Institute, 1999). Respondents were classified as cases if they met criteria for a 12-month DSM-IV/SCID diagnosis of either an anxiety disorder, mood disorder, or non-affective psychosis and had a GAF score in the range 0–70. Separate ROC curves were estimated for standardized K10 and K6 scales. These standardized scales were generated using the maximum-likelihood estimate of TPD based on the IRT parameters from the telephone pilot survey. Both the K10 and K6 were found to have very good discrimination (Fig. 2), with area under the curve of 0·876 for the K10 and 0·879 for the K6. A parallel ROC analysis to discriminate severe cases (GAF in the range 0–50) from all other community respondents found both scales to have excellent discrimination (Fig. 3), with area under the curve 0·955 for the K10 and 0·950 for the K6.

The NHIS and NSMHWB surveys

The IRT analysis was replicated in the NHIS and NSMHWB. Total-sample parameter estimates and ranges of subsample estimates are presented in Table 5. As shown there, all questions have at least one item with both severity and sensitivity parameters in the target range, while most of the questions have multiple

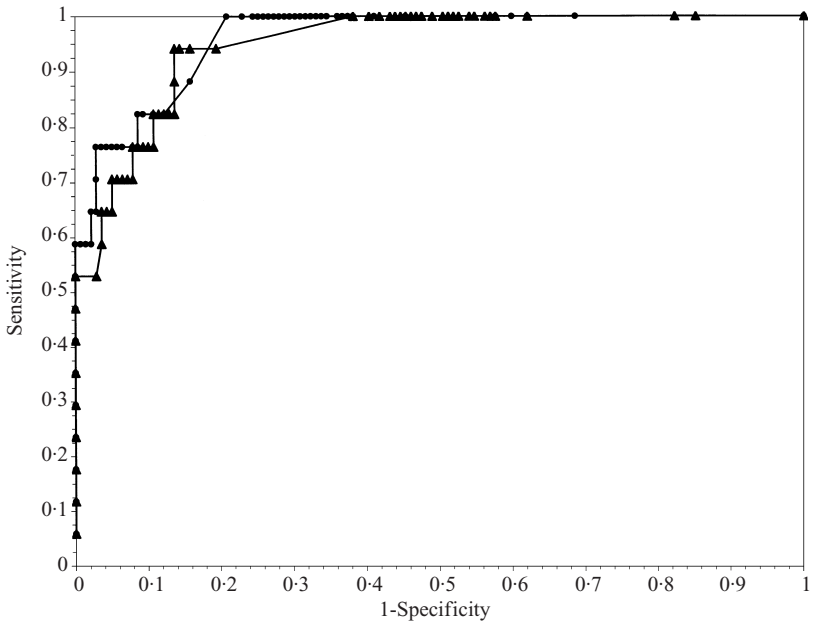


FIG. 3. Receiver operator characteristic (ROC) curves for the K10 and K6 predicting DSM-IV/SCID disorders with GAF scores in the range 0–50. (Area under the ROC curve: ●, K10 0·955; ▲, K6 0·950.)

items of this sort. It is noteworthy that significant variation can be seen in item sensitivities, justifying the use of a two-parameter IRT model. There is also significant within-question variation in item sensitivities for a number of questions, justifying the use of ordered dichotomies rather than polychotomies to estimate the IRT model.

As in the telephone pilot survey, both the K10 (a = 0·92 in the NSMHWB) and the K6 (a = 0·92 in the NHIS and a = 0·89 in the NSMHWB) had excellent internal consistency reliability. More importantly, the TIC values in the true score range 1·2–2·3 were 25–50 for the K10 in the NSMHWB, 34–54 for the K6 in the NHIS, and 17–37 for the K6 in the NSMHWB (Fig. 1). These values indicate that the precision of scale scores at the individual level is between 0·14 [i.e. $(1/54)^{1/2}$] and 0·24 [i.e. $(1/17)^{1/2}$] standard errors in the target severity range.

Finally, severity parameters were found to be very similar across sociodemographic subsamples defined on the basis of age, sex, and educational attainment, with Pearson correlations of the severity parameters across subsamples in the range 0·76–0·99 (with a mean of 0·91) for the K10 and 0·98–0·99 for the K6. Significance tests to evaluate subgroup variation in severity parameters were ignored because even extremely small differences are judged statistically significant in samples as large as the NSMHWB and especially the NHIS.

Scoring the scales

If the one-parameter model had fit, scoring the K10 and K6 would have required nothing more than summing the number of items endorsed, yielding scales with ranges of 0–40 (K10) and 0–24 (K6). However, scoring based on the two-parameter model is more complex, as optimal scaling requires the use of maximum-likelihood estimation. This is most easily implemented by using a standard IRT program, such as the BILOG-MG program used here, to estimate parameters and automatically generate maximum-likelihood scores for individuals. If a researcher wants to norm scores against a reference population, such as baseline values in the NHIS or NSMHWB that can be trended in subsequent surveys, the parameter estimates from the reference survey (e.g. the parameter estimates in Table 5 for the NHIS and NSMHWB) can be entered as start values in the IRT program and the number of iterations set to

Table 5. *Two-parameter IRT model results for the K10 and K6 scales in the NSMHWB and the NHIS*

		Sensitivity			Severity			Subsample severity range		
		NSMHWB		NHIS	NSMHWB		NHIS	NSMHWB		NHIS
		K10*	K6*	K6*	K10†	K6†	K6†	K10‡	K6‡	K6‡
Depressed (1c)§	All *v.* Others	2·0	—	—	2·7	—	—	2·3–2·8	—	—
	All-most *v.* Others	2·5	—	—	1·9	—	—	1·7–2·1	—	—
	All-some *v.* Others	1·9	—	—	1·2	—	—	1·1–1·5	—	—
	All-little *v.* Never	1·3	—	—	0·4	—	—	0·3–0·7	—	—
So depressed (1d)§	All *v.* Others	2·2	2·2	1·7	2·9	2·9	2·6	2·4–3·0	2·1–2·8	1·8–2·3
	All-most *v.* Others	2·5	2·5	2·3	2·2	2·2	1·8	1·9–2·5	1·7–2·2	1·3–1·7
	All-some *v.* Others	2·2	2·2	2·0	1·7	1·7	1·2	1·5–1·9	1·4–1·7	0·9–1·3
	All-little *v.* Never	1·5	1·4	1·6	1·2	1·2	0·7	1·1–1·5	1·0–1·4	0·5–1·0
Nervous (11b)§	All *v.* Others	1·5	1·5	1·6	3·0	3·0	2·4	2·6–3·6	2·4–3·3	1·7–2·2
	All-most *v.* Others	1·6	1·5	2·0	2·2	2·3	1·8	2·0–2·5	1·8–2·3	1·3–1·7
	All-some *v.* Others	1·2	1·1	1·7	1·6	1·7	1·1	1·4–1·8	1·4–1·7	0·8–1·2
	All-little *v.* Never	0·7	0·7	1·4	0·5	0·6	0·4	0·4–0·9	0·3–0·8	0·1–0·7
So nervous (11c)§	All *v.* Others	1·7	—	—	3·4	—	—	3·0–3·8‖	—	—
	All-most *v.* Others	2·3	—	—	2·6	—	—	2·3–2·8	—	—
	All-some *v.* Others	2·0	—	—	2·2	—	—	2·0–2·4	—	—
	All-little *v.* Never	1·4	—	—	1·8	—	—	1·5–2·0	—	—
Restless (5a)§	All *v.* Others	1·3	1·3	1·4	3·0	3·1	2·5	2·5–3·1	2·3–2·8	1·8–2·3
	All-most *v.* Others	1·5	1·4	1·8	2·0	2·1	1·8	1·8–2·2	1·7–2·0	1·3–1·7
	All-some *v.* Others	1·2	1·1	1·8	1·1	1·2	1·0	1·0–1·4	0·9–1·3	0·8–1·2
	All-little *v.* Never	0·9	0·8	1·5	0·1	0·1	0·5	−0·2–0·5	−0·2–0·4	0·2–0·7
So restless (5b)§	All *v.* Others	1·4	—	—	3·2	—	—	2·8–4·4	—	—
	All-most *v.* Others	1·5	—	—	2·4	—	—	2·1–2·6	—	—
	All-some *v.* Others	1·3	—	—	1·8	—	—	1·6–2·0	—	—
	All-little *v.* Never	1·2	—	—	1·2	—	—	1·0–1·5	—	—
Worthless (8a)§	All *v.* Others	2·0	2·4	2·7	2·8	2·7	2·5	2·2–3·1	1·9–2·8	1·7–2·1
	All-most *v.* Others	2·7	3·3	3·9	2·1	2·1	1·9	1·8–2·4	1·6–2·1	1·4–1·8
	All-some *v.* Others	2·2	2·6	3·3	1·8	1·8	1·6	1·6–2·0	1·4–1·8	1·2–1·5
	All-little *v.* Never	1·6	1·7	2·4	1·4	1·4	1·4	1·2–1·6	1·1–1·4	1·0–1·2
Tired out (7a)§	All *v.* Others	1·2	—	—	2·9	—	—	2·4–3·3	—	—
	All-most *v.* Others	1·3	—	—	2·0	—	—	1·6–2·3	—	—
	All-some *v.* Others	1·1	—	—	1·1	—	—	1·1–1·5	—	—
	All-little *v.* Never	0·8	—	—	0·3	—	—	0·2–0·5	—	—
Hopeless (1e)§	All *v.* Others	1·5	1·7	2·6	3·0	2·9	2·4	2·4–3·2	2·1–2·8	1·7–2·1
	All-most *v.* Others	2·0	2·4	3·9	2·3	2·2	1·9	2·0–2·4	1·7–2·1	1·4–1·8
	All-some *v.* Others	1·8	2·0	3·6	1·8	1·7	1·5	1·6–1·9	1·4–1·7	1·1–1·5
	All-little *v.* Never	1·3	1·4	2·5	1·2	1·2	1·2	1·0–1·5	1·0–1·3	0·9–1·3
Everything an effort (7b)§	All *v.* Others	1·5	1·6	1·4	2·6	2·5	2·4	2·0–2·6	2·1–2·3	1·7–2·2
	All-most *v.* Others	1·7	1·8	1·9	1·8	1·8	1·7	1·5–2·0	1·4–1·8	1·3–1·7
	All-some *v.* Others	1·5	1·6	2·2	1·2	1·2	1·1	1·0–1·4	0·9–1·3	0·9–1·2
	All-little *v.* Never	1·2	1·3	2·0	0·3	0·3	0·8	0·3–0·7	0·3–0·6	0·5–1·0

* Sensitivity standard errors ranged from 0·02–0·2 with a mean of 0·08 and a median of 0·06 in the NSMHWB K10, from 0·02–0·3 with a mean of 0·1 and a median of 0·09 in the NSMHWB K6, and from 0·01–0·1 with a mean of 0·04 and a median of 0·03 in the NHIS K6.

† Severity standard errors ranged from 0·01–0·2 with a mean of 0·04 and a median of 0·03 in the NSMHWB K10, from 0·01–0·1 with a mean of 0·04 and a median of 0·03 in the NSMHWB K6, and from 0·004–0·03 with a mean of 0·01 and a median of 0·01 in the NHIS K6.

‡ The two-parameter IRT model was estimated in nine subsamples (sex, male, female; years of education, <12, 12, 13–15, $\geqslant 16$; age groups: 18–34, 35–54, $\geqslant 55$). The presented range reflects the minimum and maximum values of the coefficients in those subsamples.

§ Number in parentheses refers to question numbers listed in Table 2.

‖ This range is based on eight subsamples. The severity parameter could not be estimated in the education 13–15 subsample due to non-endorsement of the item by all respondents in the other subsamples.

zero to estimate maximum-likelihood person parameters of TPD.

As noted in the section on analysis methods, a simple alternative to maximum-likelihood estimation, assuming that item parameters estimated from a reference sample are being used to score data in a new sample, is to sum the item sensitivities generated from analysis of the

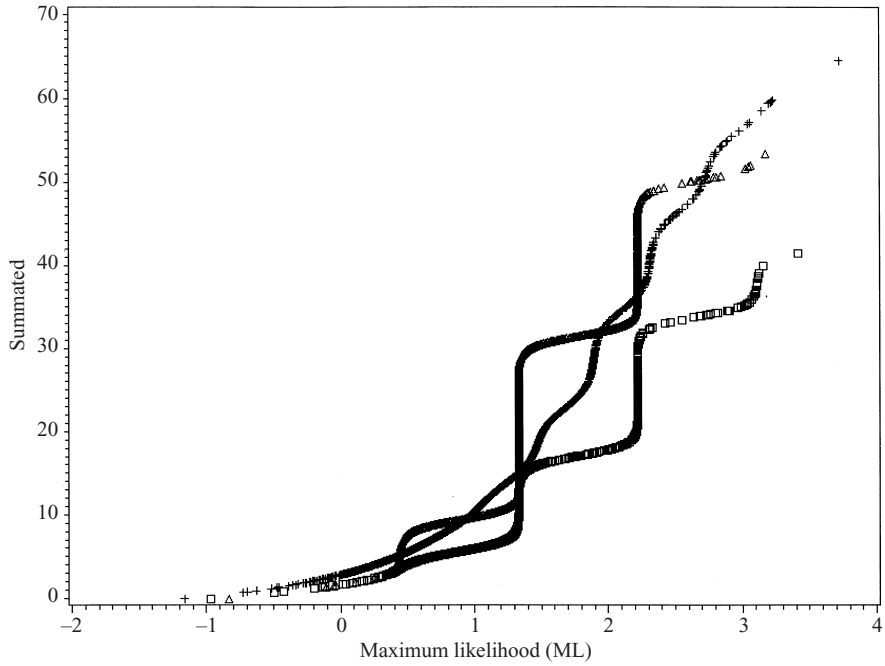


FIG. 4. Association between maximum likelihood and summated sensitivity K10 and K6 scores in the NSMHWB and NHIS. (+, NSMHWB K10; □, NSMHWB K6; △, NHIS K6.)

reference sample across all endorsed items in the new sample. As this sum is the sufficient statistic for TPD, rank orderings based on the maximum-likelihood (ML) scores and the summed scores will be identical, although the two scores will be non-linearly related.

This is illustrated in Fig. 4, where we plot the associations between ML scores and scores based on summing the sensitivities of endorsed items in the NHIS and NSMHWB surveys. The rank-order correlations of all three curves are 1·0. A 'stairstep' pattern can be seen in the curves. The reason for this, which can be derived mathematically from the likelihood equation, is that the summed score rises most sharply around points on the TPD distribution that correspond to the severity parameters for items with high sensitivities. This non-linearity is not a problem in working with the summed score rather than with the more computationally difficult ML score, as the metric of the ML score is not directly interpretable. Practical uses of the scores, furthermore, treat the scale non-linearity by classifying respondents in relation to population-normed cut-offs on clinical validators (as in the ROC analyses in Figs. 2 and 3). These uses are unaffected by the transformation. Based on these considerations, the summed scoring method is the one that should be used in practical applications. The relevant item 'weights' to use in creating these summed scales are the severity scores in Table 5.

DISCUSSION

We began with the goal of developing short screening scales of non-specific psychological distress that would be sensitive in the upper 90th–99th percentile range of the population distribution. This goal was achieved. The final K10 and K6 scales have excellent precision in the target range of the scale distribution as well as consistent levels of severity across socio-demographic subsamples. Furthermore, the scales discriminate with precision between community cases and non-cases of DSM-IV disorders. Principled methods exist to score the scales in such a way as to take account of differences in item sensitivities. Calibration methods also exist to translate scale scores into probabilities of various clinical outcomes based

on the clinical reappraisal study in order to facilitate the interpretation of population trends.

Precision

Regarding precision of the scales, is it noteworthy that analyses not reported above showed that the Test Information Curves of the final K10 and K6 scales have considerably larger ratios of information in the 90th–99th percentile range *versus* other parts of the distribution than the TIC of a scale that includes all the questions in the original question pool. This is an important result because it means that the targeted selection of items using IRT methods yielded a more useful distribution of information in the scales than we would have achieved by using classical test theory methods such as selecting the questions with the highest factor loadings or those that entered into stepwise regressions aimed at explaining variance in a scale made up of all questions.

One would expect, based on this result, that comparative analysis would show the K10 and K6 to outperform previously developed screening scales of comparable length in discriminating DSM cases and non-cases. No comparisons of this sort can be made with the data presented here because the clinical reappraisal survey did not include any other screening scales. However, a separate analysis carried out in the NSMHWB showed that the K10 and K6 both significantly outperform the GHQ-12 in discriminating CIDI/ICD-10 cases of anxiety and mood disorders (Furukawa *et al.* 2002).

The test information curves in the NSMHWB show that the K10 has between 20% and 50% more information than the K6 in the 1·2–2·3 severity range. Any researcher desiring optimal precision in this range should consequently prefer the K10 to the K6. However, given this greater precision, it is surprising that the ROC analysis failed to find the K10 to be more accurate than the K6 in discriminating between DSM-IV cases and non-cases. This presumably reflects the fact that the more subtle assessment of distress in the K10 than the K6 is not related to categorical diagnoses.

Scoring

As noted above in the section on scoring, K10 and K6 scoring can be carried out either by using optimal maximum-likelihood estimation from a target survey or by summing the sensitivities for endorsed items obtained from an IRT model estimated in a baseline survey. We recognize that re-estimating the IRT item parameters to generate sample-specific optimal scoring or using IRT parameter estimates from a reference sample to generate ML estimates in a new sample is much more complicated than using the informal unweighted summative scoring approach that is conventional in scoring screening scales (i.e. assigning one point on a scale for each item endorsed or, in the case of five-point scales, assigning between 1 and 5 points, and then summing). However, IRT-based scoring by summation of sensitivities of endorsed items estimated in a reference sample is an easy-to-use alternative that substantially increases the precision of the scale in comparison to unweighted summation.

In order to standardize the use of the sensitivity summation scoring method across many different populations and samples, the K10 and K6 were included in the WHO World Mental Health (WMH) surveys (Kessler & Üstün, 2000), a series of representative community surveys currently underway in 26 countries around the world that will have a combined sample of over 200000 respondents. The K10 and K6 severity parameter estimates obtained in the WMH surveys will be archived as soon as WMH data are available to serve as the official item weights for scoring the K10 and K6. These item weights will be archived at the URL http://www.hcp.med.harvard.edu/ncs/ under the heading 'Scoring the K10 and K6'.

Internationally valid K10 and K6 calibration rules will also be generated from the WMH survey data and made available for translating screening scale scores into predicted probabilities of various clinically significant outcomes (e.g. any DSM disorder, any ICD disorder, any serious disorder, etc.). Although preliminary rules of this sort can be derived from the ROC curves presented here, it will be possible to develop much more precise rules to predict a greater variety of clinical outcomes from the WMH data. WMH respondents are being administered the WHO Composite International Diagnostic Interview (CIDI) (WHO, 1997), a fully structured research diagnostic interview that generates diagnoses according to the definitions and criteria of both the ICD-10 and

DSM-IV diagnostic systems, and the WHO Disability Assessment Schedule (WHO-DAS) (Rehm *et al*. 1999), a fully structured instrument designed to assess the extent of impairment associated with physical and mental disorders. In addition, probability subsamples of respondents in a number of WMH surveys are participating in clinical reappraisal survey in which they are being administered a SCID by trained clinical interviewers. A wide range of clinical outcomes will be generated from these interviews for purposes of calibrating the K10 and K6. Calibration rules based on these data that take into consideration the impact of population prevalence on conditional probabilities of clinical outcomes (Furukawa *et al*. 2002) will be posted on the website mentioned above as soon as the WMH data are analysed.

Other uses of the K10 and K6

In addition to their use in trend surveys, the good results regarding precision of the K10 and K6 suggest that they would be very useful broad-gauged screening scales for mental illness in health risk appraisal surveys and primary care screening batteries. The fact that they can easily and quickly be either self-administered or interviewer-administered in 2–3 min is an important attraction here. In addition, the K10 and K6 might be useful secondary outcomes in clinical studies, as they sensitively measure the severity of non-specific distress in the range likely to be found in clinical samples. Such dimensional assessments of non-specific distress could be useful complements to the dimensional assessments of non-specific impairment, such as the GAF, that are often included in clinical studies. In addition, the inclusion of the K10 or K6 in clinical studies would provide a useful crosswalk between clinical research and community epidemiological research by allowing a comparison of the severity distribution of non-specific distress among community cases *versus* clinical cases. The absence of such comparative data has restricted our ability to interpret the clinical significance of categorical prevalence estimates in community epidemiological studies up to now. The inclusion of identical short dimensional assessments, like the K10 and K6, in both clinical and community studies would be an important step in the direction of addressing this important problem.

The research reported here was supported by US Public Health Service Grants R01 MH46376, R01 MH52861, R01 MH49098, and K05 MH00507, by the John D. and Catherine T. MacArthur Foundation Network on Successful Midlife Development (Gilbert Brim, Director) and by the Pfizer Foundation. The authors thank Jennifer Madans and her staff in the National Health Interview Survey Division of the National Center for Health Statistics for their encouragement and support of this work. We thank the expert panel members who rated the items included in the initial item pool for their consultation: Peggy Barker, David Brawn, Don Camburn, Charlie Cannell, Paul Cleary, John Eckenrode, Jack Fowler, Sue Gore, Neal Krause, Jersey Liang, Lou Magilavy, Betsy Martin, Nancy Mathiowetz, Woody Neighbors, Stanley Pressor, Pat Shrout, Eleanor Singer, Zulema Suarez, Peggy Thoits, Deborah Trunzo, Jay Turner, Aida Urtado, Elaine Wethington, Blair Wheaton, and Gordon Willis. We thank Peggy Barker, Joan Epstein and Joe Gfroefer from the Substance Abuse and Mental Health Service Administration for their support of the data analysis phase of the clinical calibration study. We thank Robert Gibbons, Norbert Swartz, and Pat Shrout for consultation regarding analysis and interpretation of results. Finally, we thank Alana Prills and T. J. Romano for manuscript preparation. A script for the interviewer-administered version of the scales and a layout for the self-administered version of the scales are archived in the NCS web page at the URL http://www.hcp.med.harvard.edu/ncs/.

REFERENCES

Andrews, G., Henderson, S. & Hall, W. (2001). Prevalence, comorbidity, disability and service utilization: an overview of the Australian National Mental Health Survey. *British Journal of Psychiatry* **178**, 145–153.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychology* **4**, 561–571.

Carroll, B. J., Feinberg, M., Smouse, S. E., Rawson, S. G. & Greden, J. F. (1981). The Carroll rating scale for depression: I. Development, reliability and validation. *British Journal of Psychiatry* **138**, 194–200.

Converse, J. M. & Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Sage: Beverley Hills, CA.

Coyne, J. C., Thompson, R. & Racioppo, M. W. (2001). Validity and efficiency of screening for history of depression by self-report. *Psychological Assessment* **13**, 163–170.

Derogatis, L. R. (1983). *SCL-90-R Revised Manual*. Johns Hopkins School of Medicine: Baltimore, MD.

Dohrenwend, B. P., Shrout, P. E., Ergi, G. E. & Mendelsohn, F. S. (1980). Measures of non-specific psychological distress and other dimensions of psychopathology in the general population. *Archives of General Psychiatry* **37**, 1229–1236.

Endicott, J., Spitzer, R. L., Fleiss, J. L. & Cohen, J. (1976). The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disorders. *Archives of General Psychiatry* **33**, 766–771.

Endler, N. S., Hunt, J. M. & Rosenstein, A. J. (1962). An S-R inventory on anxiousness. *Psychological Monographs: General and Applied* **76**, 1–33.

Fazio, A. (1977). A concurrent validational study of the NCHS General Well-Being Schedule. In *Vital and Health Statistics Publications* (Series 2, No. 73). US Government Printing Office: Washington, DC.

First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-patient Edition* (*SCID-I/NP*). Biometrics Research, New York State Psychiatric Institute: New York.

Furukawa, T. A., Andrews, G., Slade, T. & Kessler, R. C. (2002). The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychological Medicine* (in the press).

Goldberg, D. P. (1972). *The Detection of Psychiatric Illness by Questionnaire: A Technique for the Identification and Assessment of Non-Psychotic Psychiatric Illness*. Oxford University Press: London.

Gurin, G., Veroff, J. & Feld, S. C. (1960). *Americans View Their Mental Health*. Basic Books Inc.: New York, NY.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage: Newbury Park, CA.

Hanley, J. & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hodges, W. F. & Spielberger, C. D. (1969). Digit Span: an indicant of state or trait anxiety? *Journal of Consulting and Clinical Psychology* **33**, 430–434.

Kessler, R. C. & Frank, R. G. (1997). The impact of psychiatric disorders on work loss days. *Psychological Medicine* **27**, 861–873.

Kessler, R. C. & Üstün, B. T. (2000). The World Health Organization World Mental Health 2000 Initiative. *Hospital Management International* 195–196.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H.-U. & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Archives of General Psychiatry* **51**, 8–19.

Kessler, R. C., Berglund, P. A., Zhao, S., Leaf, P. J., Kouzis, A. C., Bruce, M. L., Friedman, R. M., Grosser, R. C., Kennedy, C., Kuehnel, T. G., Laska, E. M., Manderscheid, R. W., Narrow, W. E., Rosenheck, R. A., Santoni, T. W. & Schneier, M. (1996). The 12-month prevalence and correlates of Serious Mental Illness (SMI). In *Mental Health, United States 1996* (ed. R. W. Manderscheid and M. A. Sonnenschein), pp. 59–70. US Government Printing Office: Washington, DC.

Kessler, R. C., Wittchen, H.-U., Abelson, J. M., McGonagle, K., Schwarz, N., Kendler, K. S., Knäuper, B. & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *International Journal of Methods in Psychiatric Research* **7**, 33–55.

Kessler, R. C., Berglund, P. A., Bruce, M. L., Koch, J. R., Laska, E. M., Leaf, P. J., Manderscheid, R. W., Rosenheck, R. A., Walters, E. E. & Wang, P. S. (2001). The prevalence and correlates of untreated serious mental illness. *Health Services Research* **36**, 987–1007.

Langner, T. S. (1962). A twenty-two item screening score of psychiatric symptoms indicating impairment. *Journal of Health and Human Behavior* **3**, 269–276.

Leighton, A. H. (1975). *My Name is Legion. Vol. 1 of the Stirling County Study*. Basic Books: New York.

Link, B. G. & Dohrenwend, B. P. (1980). Formation of hypotheses about the true relevance of demoralization in the United States. In *Mental Illness in the United States: Epidemiological Estimates* (ed. B. P. Dohrenwend, B. S. Dohrenwend, M. S. Gould, B. Link, R. Neugebauer and R. Wunsch-Hitzig), pp. 114–132. Praeger: New York.

Lubin, B., Dupre, V. A. & Lubin, A. W. (1967). Comparability and sensitivity of Set 2 (Lists E, F and G) of the Depression Adjective Checklists. *Psychological Reports* **20**, 756–758.

MacMillan, A. M. (1957). The Health Opinion Survey: techniques for estimating prevalence of psychoneurotic and related types of disorder in communities. *Psychological Reports* **3**, 325–339.

Myers, J. K., Lindenthal, J. J. & Pepper, M. P. (1975). Life events, social integration and psychiatric symptomatology. *Journal of Health and Social Behavior* **16**, 421–427.

National Advisory Mental Health Council. (1993). Health care reform for Americans with severe mental illnesses. *American Journal of Psychiatry* **150**, 1447–1465.

National Center for Health Statistics. (2000*a*). *Data File Documentation, National Health Interview Survey, 1997* (machine-readable data file and documentation). National Center for Health Statistics: Hyattsville, MD.

National Center for Health Statistics. (2000*b*). *Data File Documentation, National Health Interview Survey, 1998* (machine-readable data file and documentation). National Center for Health Statistics: Hyattsville, MD.

Pearlin, L. I., Lieberman, M. A., Menaghan, E. S. & Mullen, J. T. (1981). The stress process. *Journal of Health and Social Behavior* **22**, 337–356.

Radloff, L. S. (1997). The CES-D Scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385–401.

Regier, D., Narrow, W., Rupp, A., Rae, D. & Kaelber, C. (2000). The epidemiology of mental disorder treatment need: community estimates of medical necessity. In *Unmet Need in Psychiatry* (ed. G. Andrews and S. Henderson), pp. 41–58. Cambridge University Press: Cambridge.

Rehm, J., Üstün, T. B., Saxena, S., Nelson, C. B., Chatterji, S., Ivis, F. & Adlaf, E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *International Journal of Methods in Psychiatric Research* **8**, 110–123.

Robins, L. N. & Regier, D. A. (1991). *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. The Free Press: New York.

Robins, L. N., Helzer, J. E., Croughan, J. L. & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics and validity. *Archives of General Psychiatry* **38**, 381–389.

Robins, L. N., Wing, J., Wittchen, H.-U., Helzer, J. E., Babor, T. F., Burke, J. D., Farmer, A., Jablenski, A., Pickens, R., Regier, D. A., Sartorius, N. & Towle, L. H. (1988). The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry* **45**, 1069–1077.

Rush, J., Carmody, T. & Reimitz, P.-E. (2000). The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) ratings of depressive symptoms. *International Journal of Methods in Psychiatric Research* **9**, 45–59.

Ryff, C., Brim, O. & Kessler, R. C. (2002). *A Portrait of Midlife in the US*. University of Chicago Press: Chicago, IL.

SAS Institute, Inc. (1999). *SAS User's Guide, Release 8*. SAS Institute, Inc.: Carey, NC.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. In *Sociological Methodology* (ed. C. Clogg), pp. 271–307. American Sociological Association: Washington, DC.

Scientific Software International, Inc. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Scientific Software International, Inc.: Chicago, IL.

Scientific Software International, Inc. (1998). *TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis*. Scientific International, Inc.: Chicago, IL.

Seiler, L. H. (1973). The 22-item scale used in field studies of mental illness: a question of method, a question of substance, and a question of theory. *Journal of Health and Social Behavior* **14**, 252–264.

Sheehan, D., Lecrubier, Y., Sheehan, K., Amorim, P. & Janavs, J. (1998). The Mini-International Neuropsychiatric Interview

(MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry* **59**, 22–33.

Spitzer, R., Williams, J., Kroenke, K., Linzer, M., deGruy, F. III., Hahn, S., Brody, D. & Johnson, J. (1994). Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA: Journal of the American Medical Association* **272**, 1749–1756.

Srole, L., Langner, T. S., Michael, S. T., Opler, M. K. & Rennie, T. A. (1962). *Mental Health in the Metropolis: The Midtown Manhattan Study*. McGraw-Hill: New York.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology* **48**, 285–290.

Thissen, D. & Wainer, H. (2001). *Test Scoring*. Lawrence Erlbaum Associates, Publishers: Mahwah, NJ.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. Springer-Verlag: New York.

Ware, J. & Sherbourne, C. (1992). The SF-36 short-form health status survey: 1. Conceptual framework and item selection. *Medical Care* **3**, 473–481.

Ware, J. E., Johnston, S. A. & Davies-Avery, A. (1979). *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Mental Health* (pub. no. R-1987/3-HEW). Rand Corporation: Santa Monica, CA.

Wittchen, H.-U. (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research* **28**, 57–84.

World Health Organization (1997). *Composite International Diagnostic Interview (CIDI): Version 2.1*. World Health Organization: Geneva.

Zung, W. W. (1963). A self-report depression scale. *Archives of General Psychiatry* **12**, 63–70.

Zung, W. W. (1971). A rating instrument for anxiety disorders. *Psychosomatics* **12**, 371–379.

33348

POST-DEPLOYMENT Health Assessment

Authority: 10 U.S.C. 136 Chapter 55. 1074f, 3013, 5013, 8013 and E.O. 9397

Principal Purpose: To assess your state of health after deployment outside the United States in support of military operations and to assist military healthcare providers in identifying and providing present and future medical care to you.

Routine Use: To other Federal and State agencies and civilian healthcare providers, as necessary, in order to provide necessary medical care and treatment.

Disclosure: (Military personnal and DoD civilian Employees Only) Voluntary. If not provided, healthcare WILL BE furnished, but comprehensive care may not be possible.

INSTRUCTIONS: Please read each question completely and carefully before marking your selections. Provide a response for each question. If you do not understand a question, ask the administrator.

Demographics

Last Name

First Name **MI**

Name of Your Unit or Ship during this Deployment

Today's Date (dd/mm/yyyy) / /

Social Security Number — —

DOB (dd/mm/yyyy) / /

Date of arrival in theater (dd/mm/yyyy) / /

Date of departure from theater (dd/mm/yyyy) / /

Gender
- ○ Male
- ○ Female

Service Branch
- ○ Air Force
- ○ Army
- ○ Coast Guard
- ○ Marine Corps
- ○ Navy
- ○ Other

Component
- ○ Active Duty
- ○ National Guard
- ○ Reserves
- ○ Civilian Government Employee

Pay Grade

○ E1	○ O01	○ W1
○ E2	○ O02	○ W2
○ E3	○ O03	○ W3
○ E4	○ O04	○ W4
○ E5	○ O05	○ W5
○ E6	○ O06	
○ E7	○ O07	○ Other
○ E8	○ O08	
○ E9	○ O09	
	○ O10	

Location of Operation
- ○ Europe
- ○ SW Asia
- ○ SE Asia
- ○ Asia (Other)
- ○ Australia
- ○ Africa
- ○ Central America
- ○ Unknown
- ○ South America
- ○ North America
- ○ Other

To what areas were you mainly deployed:
(mark all that apply - list where/date arrived)
- ○ Kuwait
- ○ Qatar
- ○ Afghanistan
- ○ Bosnia
- ○ On a ship
- ○ Iraq
- ○ Turkey
- ○ Uzbekistan
- ○ Kosovo
- ○ CONUS
- ○ Other

Name of Operation:

Occupational specialty during this deployment (MOS, NEC or AFSC)

Combat specialty:

Administrator Use Only

Indicate the status of each of the following:

Yes	No	N/A	
○	○	○	Medical threat debriefing completed
○	○	○	Medical information sheet distributed
○	○	○	Post Deployment serum specimen collected

Reset

33348

Please answer all questions in relation to THIS deployment

1. Did your health change during this deployment?

- ○ Health stayed about the same or got better
- ○ Health got worse

2. How many times were you seen in sick call during this deployment?

No. of times

3. Did you have to spend one or more nights in a hospital as a patient during this deployment?

- ○ No
- ○ Yes, reason/dates: _____

4. Did you receive any vaccinations just before or during this deployment?

- ○ Smallpox (leaves a scar on the arm)
- ○ Anthrax
- ○ Botulism
- ○ Typhoid
- ○ Meningococcal
- ○ Other, list: _____
- ○ Don't know
- ○ None

5. Did you take any of the following medications during this deployment?
(mark all that apply)

- ○ PB (pyridostigmine bromide) nerve agent pill
- ○ Mark-1 antidote kit
- ○ Anti-malaria pills
- ○ Pills to stay awake, such as dexedrine
- ○ Other, please list _____
- ○ Don't know

6. Do you have any of these symptoms now or did you develop them anytime during this deployment?

No	Yes During	Yes Now	
○	○	○	Chronic cough
○	○	○	Runny nose
○	○	○	Fever
○	○	○	Weakness
○	○	○	Headaches
○	○	○	Swollen, stiff or painful joints
○	○	○	Back pain
○	○	○	Muscle aches
○	○	○	Numbness or tingling in hands or feet
○	○	○	Skin diseases or rashes
○	○	○	Redness of eyes with tearing
○	○	○	Dimming of vision, like the lights were going out

No	Yes During	Yes Now	
○	○	○	Chest pain or pressure
○	○	○	Dizziness, fainting, light headedness
○	○	○	Difficulty breathing
○	○	○	Still feeling tired after sleeping
○	○	○	Difficulty remembering
○	○	○	Diarrhea
○	○	○	Frequent indigestion
○	○	○	Vomiting
○	○	○	Ringing of the ears

7. Did you see anyone wounded, killed or dead during this deployment?
(mark all that apply)

○ No ○ Yes - coalition ○ Yes - enemy ○ Yes - civilian

8. Were you engaged in direct combat where you discharged your weapon?

○ No ○ Yes (○ land ○ sea ○ air)

9. During this deployment, did you ever feel that you were in great danger of being killed?

○ No ○ Yes

10. Are you currently interested in receiving help for a stress, emotional, alcohol or family problem?

○ No ○ Yes

11. Over the LAST 2 WEEKS, how often have you been bothered by any of the following problems?

None	Some	A Lot	
○	○	○	Little interest or pleasure in doing things
○	○	○	Feeling down, depressed, or hopeless
○	○	○	Thoughts that you would be better off dead or hurting yourself in some way

33348

Reset

12. Have you ever had any experience that was so frightening, horrible, or upsetting that, IN THE PAST MONTH, you ....

No	Yes	
○	○	Have had any nightmares about it or thought about it when you did not want to?
○	○	Tried hard not to think about it or went out of your way to avoid situations that remind you of it?
○	○	Were constantly on guard, watchful, or easily startled?
○	○	Felt numb or detached from others, activities, or your surroundings?

13. Are you having thoughts or concerns that ...

No	Yes	Unsure	
○	○	○	You may have serious conflicts with your spouse, family members, or close friends?
○	○	○	You might hurt or lose control with someone?

14. While you were deployed, were you exposed to:
(mark all that apply)

No	Sometimes	Often	
○	○	○	DEET insect repellent applied to skin
○	○	○	Pesticide-treated uniforms
○	○	○	Environmental pesticides (like area fogging)
○	○	○	Flea or tick collars
○	○	○	Pesticide strips
○	○	○	Smoke from oil fire
○	○	○	Smoke from burning trash or feces
○	○	○	Vehicle or truck exhaust fumes
○	○	○	Tent heater smoke
○	○	○	JP8 or other fuels
○	○	○	Fog oils (smoke screen)
○	○	○	Solvents
○	○	○	Paints
○	○	○	Ionizing radiation
○	○	○	Radar/microwaves
○	○	○	Lasers
○	○	○	Loud noises
○	○	○	Excessive vibration
○	○	○	Industrial pollution
○	○	○	Sand/dust
○	○	○	Depleted Uranium (If yes, explain) _____
○	○	○	Other exposures _____

15. On how many days did you wear your MOPP over garments?

No. of days

16. How many times did you put on your gas mask because of alerts and NOT because of exercises?

No. of times

17. Were you in or did you enter or closely inspect any destroyed military vehicles?

○ No ○ Yes

18. Do you think you were exposed to any chemical, biological, or radiological warfare agents during this deployment?

○ No ○ Don't know
○ Yes, explain with date and location

_____

_____

_____

33348

Reset

Health Care Provider Only

SERVICE MEMBER'S SOCIAL SECURITY # ___ — __ — ____

Post-Deployment Health Care Provider Review, Interview, and Assessment

Interview

1. Would you say your health in general is: ○ Excellent ○ Very Good ○ Good ○ Fair ○ Poor
2. Do you have any medical or dental problems that developed during this deployment? ○ Yes ○ No
3. Are you currently on a profile or light duty? ○ Yes ○ No
4. During this deployment have you sought, or do you now intend to seek, counseling or care for your mental health? ○ Yes ○ No
5. Do you have concerns about possible exposures or events during this deployment that you feel may affect your health? ○ Yes ○ No
   Please list concerns: _____
6. Do you currently have any questions or concerns about your health? ○ Yes ○ No
   Please list concerns: _____

Health Assessment

After my interview/exam of the service member and review of this form, there is a need for further evaluation as indicated below. (More than one may be noted for patients with multiple problems. Further documentation of the problem evaluation to be placed in the service member's medical record.)

REFERRAL INDICATED FOR:

- ○ None
- ○ Cardiac
- ○ Combat/Operational Stress Reaction
- ○ Dental
- ○ Dermatologic
- ○ ENT
- ○ Eye
- ○ Family Problems
- ○ Fatigue, Malaise, Multisystem complaint
- ○ Audiology
- ○ GI
- ○ GU
- ○ GYN
- ○ Mental Health
- ○ Neurologic
- ○ Orthopedic
- ○ Pregnancy
- ○ Pulmonary
- ○ Other _____

EXPOSURE CONCERNS (During deployment):

- ○ Environmental
- ○ Occupational
- ○ Combat or mission related
- ○ None

Comments: _____

I certify that this review process has been completed.
Provider's signature and stamp:

This visit is coded by V70.5 _ _ 6

Date (dd/mm/yyyy) __ / __ / ____

End of Health Review

Journal of Personality and Social Psychology
1993, Vol. 65, No. 4, 757–768

Copyright 1993 by the American Psychological Association, Inc.
0022-3514/93/$3.00

A Comparison of Three Structural Models for Personality: The Big Three, the Big Five, and the Alternative Five

Marvin Zuckerman, D. Michael Kuhlman, Jeffrey Joireman, Paul Teta, and Michael Kraft

The major factors from 3 models of personality are compared: Eysenck's Three Factor model, Costa and McCrae's version of the Big Five, and Zuckerman and Kuhlman's Alternative Five. The 1st study describes the development of a questionnaire measure for the Alternative Five and the reliability assessments of the scales. The 2nd study used factor analysis to compare the factors among the scales from the 3 models. Extraversion and Neuroticism were quite similar across all 3 models. Eysenck's Psychoticism scale marked a factor that included Conscientiousness and Impulsive Sensation Seeking factors from the other 2 models. Agreeableness and Aggression-Hostility formed a 4th factor. Openness could be identified as a factor using facet scales, but it showed no convergence with other factors. Four of the 5 factors showed convergence across at least 2 of the models.

For many years, the field of personality structure, or the taxonomy of traits, was dominated by two models, H. J. Eysenck's (1947, 1967) three-factor and Cattell's (1957) 16-factor models. Almost unnoticed at first, the Big Five model emerged and evolved (see Digman, 1990) from its origins in the Fiske (1949) study on the selection of clinical psychologists and Norman's (1963) research (developed from work by Tupes & Christal, 1961), to its present claim to be the best paradigm for personality structure (Costa & McCrae, 1992a, 1992c; Goldberg, 1992). This claim has not gone unchallenged (Eysenck, 1992a, 1992b; Zuckerman, 1992), and the issues have been debated at recent conferences (Angleitner, 1992; Halverson & Kohnstamm, 1991; Zuckerman, 1991a).

Five- and three-factor theorists generally agree on two points. First, 16 basic factors are at least 11 too many, although they could represent narrow factors comprising the broader ones. The problem is that Cattell (1957) and Guilford and Zimmerman's (1956) factors are generally not replicable across gender, age, or methods, and many investigators have simply failed to find them (Digman, 1990; H. J. Eysenck & Eysenck, 1969). The second point is that in all five- and three-factor models, two of the major factors consist of Extraversion–Introversion (E, or Surgency) and Neuroticism (N, or Emotional Stability–Instability, or Negative Emotionality [Tellegen, 1985]). Although all models concur on the existence of these two basic factors, there is less agreement on the narrower traits comprising them, and the conformance between measures of similar traits derived from different models is open to investigation. The more perplexing and contentious question, however, is "what lies beyond E and N?" (Zuckerman, Kuhlman, & Camac, 1988). H. J. Eysenck (1967; H. J. Eysenck & Eysenck, 1985) saw the necessity for only one additional factor, which he called *Psychoticism* (P). The Big Five includes three more factors labeled *Agreeableness, Conscientiousness,* and (variously) *Intelligence, Intellect, Intellectance, Culture,* or *Openness to Experience.*

An alternative five-factor model (Zuckerman, Kuhlman, Thornquist, & Kiers, 1991) adds Impulsive Unsocialized Sensation Seeking (ImpUSS), Aggression–Hostility (Agg-Host), and Activity (Act), to Sociability (Sy) and Neuroticism–Anxiety (N-Anx). Eysenck suggested that factors other than his Big Three, such as Agreeableness, Aggression, Conscientiousness, Impulsivity, Activity, and Sensation Seeking, are either components of E, N, or P, or combinations of two of them. Conscientiousness and Agreeableness factors of the Big Five, for instance, are regarded by Eysenck (1992b) as components of P, whereas Big Five proponents (Costa & McCrae, 1992c; Goldberg & Rosolack, in press) regard P as a combination of the primary traits of Agreeableness and Conscientiousness. Although Goldberg and Rosolack have said that "Psychoticism is a hybrid of factors II [disagreeableness] and III [unconscientiousness]," they assert that P is not a higher order or super factor because it does not emerge in three-factor rotations along with the clusters of trait adjectives. However, the lexical criterion is not necessarily the final word on which are the basic personality factors. The encoding of personality traits in the language may not be proportional to the behavioral importance of the traits or their biological relevance (Zuckerman, 1992). Agreeableness and Conscientiousness are important and salient traits for human observers, but the capacity to be socialized by societal conditioning may be more basic and fundamental than these two of its components.

The definition of a basic trait of personality cannot be settled by factor analyses of trait measures alone (Eysenck, 1992a, 1992b; Zuckerman, 1992). The outcomes of factor analyses obviously depend on the sampling of variables to be factored. Eysenck and Zuckerman have insisted that basic traits should be

Marvin Zuckerman, D. Michael Kuhlman, Jeffrey Joireman, Paul Teta, and Michael Kraft, Department of Psychology, University of Delaware.

Copies of the Zuckerman–Kuhlman Personality Questionnaire Form III (ZKPQ III) and scoring keys may be obtained from Marvin Zuckerman.

Correspondence concerning this article should be addressed to Marvin Zuckerman, Department of Psychology, University of Delaware, Newark, Delaware, 19716–2577.

part of a nomological network or theory and have some demonstrable biological basis. Costa and McCrae (1992a) responded that "it is poor science to explain the known on the basis of the unknown" (p. 659). Although the psychobiology of basic personality (Zuckerman, 1991b) is in its infancy, much is known about certain traits such as E (H. J. Eysenck & Eysenck, 1985), Anxiety (Gray, 1982, 1991), P (H. J. Eysenck & Eysenck, 1976), and Sensation Seeking (Zuckerman, 1979, 1983, 1984), but less is known about Conscientiousness, Agreeableness, and Openness to Experience. It might be argued that the former kinds of traits are dimensions of temperament, whereas the latter are a model for personality. Strelau (1983) made the distinction between temperament and personality traits in terms of the stronger role of biological factors, early appearance in development, analogues in other species, and expressive, energetic expressions in temperament traits as opposed to directional, motivational expressions in personality traits. However, Zuckerman (1991b) has argued that the distinction between temperament and basic personality traits is not at all clear or meaningful. What are called *traits of temperament*, such as Sociability, Emotionality, Aggression, Impulsivity, and Sensation Seeking, are not very different in content from E, N, Agreeableness, and Conscientiousness. One of the purposes of this article is to assess the degree of correspondence between such traits as operationally defined in questionnaires.

The alternative five-factor model emerged from factor analyses (Zuckerman et al., 1988, Zuckerman et al., 1991) that also included S. B. G. Eysenck, Eysenck, and Barrett's (1985) revision of the Eysenck Personality Questionnaire (EPQ) scales for the three-factor model. The revision of the EPQ was done primarily to correct psychometric deficiencies in the P scale, including low internal reliability, a curtailed scoring range, and a highly skewed distribution of scores. The new scale achieved some improvement, but reliabilities are still lower than those for other EPQ scales, and the distribution is still far from normal.

The sampling of scales for the original factor analyses was based on hypothesized factors of basic personality from the research literature on temperament and personality, and many of the scales had been used in psychobiological research (Zuckerman et al., 1988). The second study used the findings from the first study to more sharply define the factors and test their replicability across four samples (Zuckerman et al., 1991). This rational selection of scales for the initial factor analyses differs from the lexical sampling determining the Big Five and therefore might have yielded different kinds of factors. There was no extensive sampling of scales for an Intellect–Openness factor, for instance, but more extensive sampling of Sensation Seeking scales because of their demonstrated biological relevance (Zuckerman, Buchsbaum, & Murphy, 1980). Social desirability scales were also included to assess the importance of this trait in the content factors.

In the three-factor solution in both studies, the E scale from the EPQ loaded on a Sociability factor, along with other scales of affiliative tendencies, the N scale was the strongest marker for a neuroticism–emotionality factor, and the P scale was the best marker for ImpUSS. ImpUSS was the same at the five-factor level in the Zuckerman et al. (1991) study, and P was still a strong marker for the factor. At this level, N remained a strong marker for a dimension containing other N and anxiety scales, but aggression, anger, hostility, and inhibition of aggression scales formed a factor distinct from that in N-Anx. Scales of activity, energy level, and persistence formed a fifth factor, distinct from their location in the E factor in the three-factor analysis. E continued to group with the scales in the Sy factor but, unlike other sociability scales, showed a secondary loading on P-ImpUSS. The results demonstrated a convergence of the Eysenck model with three of the factors in the alternative five-factor model.

Studies comparing the Eysenck three-factor and Costa and McCrae five-factor models (Angleitner & Ostendorf, in press; McCrae & Costa, 1985) showed that E and N were highly convergent with the corresponding trait measures in both systems. McCrae and Costa found that Eysenck's measure of P was related to Agreeableness and Conscientiousness factors in their Big Five model. Angleitner and Ostendorf also found relationships between P and Agreeableness, Conscientiousness, and Openness in a five-factor analysis. However, using a six-factor model, P correlated most highly with Openness.

Angleitner and Ostendorf (in press) also compared other putative measures of adult "temperament" with Costa and McCrae's (NEO) measures of the five-factor model. They included the four subscales of the Sensation Seeking Scale (SSS Form V, Zuckerman, 1979; Zuckerman, Eysenck, & Eysenck, 1978) and the subscales of the Buss and Plomin (1975) EASI temperament measures. All of these temperament scales were used in the studies first defining the alternative five-factor model (Zuckerman et al., 1988; Zuckerman et al., 1991), allowing a comparison of the Big Five and the Alternative Five models. Three of the four sensation seeking subscales of the SSS and a sensation seeking subscale of impulsivity on the EASI loaded most highly on the openness factor of the NEO. The Experience Seeking subscale of the SSS had the strongest relationship with the Openness factor, and the other two subscales of the SSS and the EASI had equal or near-equal loadings (negative) on the Agreeableness factor. The disinhibition subscale of the SSS had its primary and substantial loading (negative) on the Agreeableness factor.

Costa and McCrae (1992a) reanalyzed the factor analysis data of Zuckerman et al. (1991), rerotating the five factors to align with putative markers for their model. They used the SSS Experience Seeking subscale to identify the Openness factor. The new rotations resulted in Eysenck's P scale loading moderately on both the Openness (positively) and Conscientiousness (negatively) factors and somewhat lower on the Agreeableness (negatively) factor. The loading of P on the Openness factor is somewhat inconsistent with prior notions of its identity as the opposite of Conscientiousness and Agreeableness, but Eysenck (1992c) has maintained that creativity is associated with the P dimension and provides one of the adaptive advantages of the trait. Supporting this notion is the fact that creative and successful artists and writers are among the highest scoring groups on the P scale (Zuckerman, 1989).

There has been no study directly comparing all three models because until recently there were no direct measures for the alternative five-factor model. The first study in this article describes the development of a questionnaire for this model. The second study was designed to compare the three models using

the operational definitions provided by the primary scales for each. In view of the previous studies, a high degree of convergence was expected for the Extraversion–Sociabililty and N factors in all three models. ImpUSS in the alternative model was expected to be negatively related to Conscientiousness, and Agg-Host in our model was predicted to measure the obverse of Agreeableness in the Big Five model. The major questions concern the equivalencies for P from the Eysenck model in the other two. The results from previous studies have not been consistent, some suggesting Agreeableness and others Openness or Conscientiousness as primary equivalents of P.

A second type of comparison of the components of the three models is the extent to which they correlate with the same scales from other questionnaire measures. Things that correlate with the same things are not necessarily the same, but when the total profile of comparisons is similar, they obviously share some important characteristics. Other models, including the original four-factor temperament scheme of Buss and Plomin (1975) and the two-factor system of Block and Block (1980), were compared with the three- and five-factor models central to this study. Measures that are of a narrower or more specific focus such as cognitive trait and reinforcement sensitivity scales were also of interest. Concepts such as "generalized reward expectancy" are relevant to the theoretical mechanisms underlying some of the traits.

The hypotheses generated from the three models are as follows:

1. According to Eysenck, the scales from the other two tests (NEO and Zuckerman–Kuhlman Personality Questionnaire [ZKPQ]) should load on three factors identifiable by his own markers as E, N, and P, respectively, with one exception. Because he and Zuckerman deliberately excluded markers for the Culture–Intellect–Openness factor, the NEO Openness scale would not be expected to be encompassed within his three-factor model. The P factor should consist of scales for Agg-Host (ZKPQ), Agreeableness and Conscientiousness (NEO), and ImpUSS (ZKPQ), the latter because of the impulsivity rather than the Sensation Seeking component. It is important that P be the highest, or among the highest loading scales on this factor in a three-factor analysis on the assumption that it is the best comprehensive measure of the broad factor.

2. The Costa and McCrae model would predict at least four major factors using the total scores on the NEO scales (N, E, Agreeableness, and Conscientiousness), with their own scales loading highly on these supertraits. The fifth factor of Openness would not be expected when using only the single NEO total score for Openness because there would be no corresponding scale within the EPQ or ZKPQ. The fifth factor should be defined by the six NEO facet scales for Openness, but might also include some secondary loadings from the ZKPQ ImpUSS scale or the EPQ P scales according to Angleitner and Ostendorf's (in press) results.

3. The Zuckerman and Kuhlman model would also predict at least four factors using the broad scales from all of the tests with high loadings from the ZKPQ Sy, N-Anx, ImpUSS, and Agg-Host scales for the four factors. A fifth factor of Act would not be expected because of the absence of a corresponding major scale in the other tests. Act might emerge as a fifth factor in an analysis using the NEO facet scales because there is an Activity Facet scale within the major E scale that might combine with ZKPQ Act. ZKPQ ImpUSS could be most strongly related to Agreeableness, Conscientiousness, or Openness scales of the NEO because previous findings have found one or more of these correlating with ImpSS. According to previous factor analyses (Zuckerman et al., 1988; Zuckerman et al., 1991), ImpSS will be more strongly related to a dimension marked by EPQ P than one identified with E.

Study 1: Development of Scales for the Alternative Five-Factor Model

Method

Factor scores were computed for individual subjects on each of the basic five factors in the the Zuckerman et al. (1991) study. All of the items contained in the scales used in the Zuckerman et al. (1991) study, except those from the EPQ, were correlated with the five factor scores in the combined samples (n = 522). Twenty items were selected for each factor scale on the basis of having the highest correlation with that factor, little correlation with other factors, and minimal correlation with the Crowne–Marlowe (1960) Social Desirability scale. Although there is evidence that social desirability is not an important confounding influence in most scales (McCrae & Costa, 1983), it can affect the endorsement frequency of certain types of items, particularly those dealing with socially unacceptable traits such as aggression, and may account for the skewed distribution of scales such as Eysenck and Eysenck's (1976) P.

Many of the items were rewritten for the second form of the ZKPQ (ZKPQ II) because some were from published tests and others needed improvement in wording. The ZKPQ II, consisting of 100 items, was given to 589 subjects, and their item responses were factor analyzed using principal components and a varimax rotation. Scree tests indicated the appropriateness of the five-factor solution. Ninety of the 100 items loaded at least .30 on their predicted factor and higher on that factor than any of the other four. One item met this criterion, but the loading on a second factor was so close to that on the primary factor that it was eliminated. The remaining 89 items were selected for the revised scale (ZKPQ III). An examination of the distributions of scores on the five scales showed that all but the Sy scale were fairly normally distributed, but the distribution of scores on Sy was skewed toward the high end, so new items were written for this scale (primarily introversion-type items) in an attempt to normalize the distribution. An examination of the distribution in a subsequent sample showed that the distribution was now much closer to normal. Ten new items were written for an infrequency scale to detect careless responding.

The five scales can be described in terms of their typical content:

1. *Impulsive Sensation Seeking (ImpSS, 19 items).* The impulsivity items involve a lack of planning and the tendency to act impulsively without thinking. The ImpSS items are general in content and do not describe specific activities such as drinking or sex. Most can be described as experience seeking, or the willingness to take risks for the sake of excitement or novel experience.

2. *N-Anx (19 items).* These items describe emotional upset, tension, worry, fearfulness, obsessive indecision, lack of self-confidence, and sensitivity to criticism.

3. *Agg-Host (17 items).* About half of the items describe a readiness to express verbal aggression. Others concern rude, thoughtless, or antisocial behavior, vengefulness, and spitefulness. The persons scoring high admit a quick temper and impatience with others.

4. *Act (17 items).* About half of the items describe the need for activity and an inability to relax and do nothing when the opportunity presents itself (Type A personality?). The other items portray a prefer-

ence for hard or challenging work, an active busy life, and a high energy level.

5. *Sy (17 items).* These items concern the number of friends one has and the amount of time spent with them, outgoingness at parties, and a preference for being with others as opposed to being alone and pursuing solitary activities.

Infrequency (Inf, 10 items). These items are mostly exaggerated, true scored, socially desirable but unlikely to be completely true statements about anyone. The scale is highly skewed, with most scores around 0 or 1, and this may account for its low internal reliability. Rather than being regarded as a scale, it should only be used to detect inattention to the task or an excessive concern with appearing socially desirable. Only 4% of the subjects scored higher than 3 on the scale. A higher score may indicate the need to investigate the validity of the test for the particular person achieving that score.

The revised 99-item version (ZKPQ III-R) was given to classes in introductory psychology ($n = 730$) and personality psychology ($n = 187$) to assess reliabilities, sex differences, and intercorrelations of the five scales.

Results

Table 1 presents the alpha coefficient reliabilities in men and women and the combined samples in both courses. These range from .70 to .86. N-Anx had the highest reliabilities and Act had the lowest reliabilities of the five scales, but all reliabilities were quite satisfactory.

Table 2 presents the means and standard deviations of the five scales in men and women in the two courses. Men scored significantly higher on the ImpSS scale in the general psychology course, and the difference was close to significant in the personality course. Women were significantly higher than men on N-Anx in both courses. Men in the introductory psychology course were significantly higher than women on Agg-Host, Act, and infrequent response, but of these, only the difference on infrequent response was significant in the personality course.

Table 3 shows the correlations between the ZKPQ III-R subscales in the two courses. Most of these were minimal ($r < .20$), except for the correlations between ImpSS and Agg-Host and Sy in both groups and that between Agg-Host and Sy in the personality course.

Table 1
Reliabilities (Alpha Coefficients) for the ZKPQ III

	Experiment 1			Experiment 2		
Scale	Men	Women	All	Men	Women	All
ImpSS	.74	.78	.77	.82	.82	.82
N-Anx	.81	.82	.83	.82	.86	.86
Agg-Ho	.72	.76	.75	.80	.74	.76
Sy	.76	.75	.75	.79	.83	.81
Act	.72	.74	.73	.70	.74	.72
n	256	474	730	61	126	187

Note. Imp-SS = Impulsive Sensation Seeking; ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; N-Anx = Neuroticism–Anxiety; Agg-Ho = Aggression–Hostility; Sy = Sociability; Act = Activity.

Table 2
Means, Standard Deviations, and t Tests Between Men and Women on the ZKPQ III Scales

	Experiment 1			Experiment 2		
Scale	Men	Women	*t*	Men	Women	*t*
ImpSS						
M	11.12	9.81	4.47**	10.39	9.10	1.91
SD	3.68	3.96		4.35	4.30	
N-Anx						
M	7.82	11.02	9.85**	7.41	10.37	4.41**
SD	4.17	4.23		4.11	4.68	
Agg-Ho						
M	9.18	8.08	4.05**	8.34	7.30	1.80
SD	3.47	3.57		3.83	3.42	
Sy						
M	10.51	10.51	0	9.31	10.06	1.22
SD	3.58	3.43		3.91	4.00	
Act						
M	8.14	7.44	2.40*	7.02	7.56	1.04
SD	3.80	3.70		3.59	3.81	
Infreq						
M	1.40	.78	6.11**	1.08	.73	2.04*
SD	1.46	.97		1.14	1.01	

Note. ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; ImpSS = Impulsive Sensation Seeking; N-Anx = Neuroticism-Anxiety; Agg-Ho = Aggression-Hostility; Sy = Sociability; Act = Activity; Infreq = Infrequency.
* $p < .05$. ** $p < .0001$.

Study 2: Relationships Among the Three Models and With Other Scales

Procedure and Method

Students in a course in personality psychology were given the personality tests on a series of weekly psychometric laboratory days. Their own data furnished part of the basis for a term paper on the analysis of their own personality. The first personality test given was the EPQ-R (S. B. G. Eysenck et al., 1985). A week later they were administered the ZKPQ III, and one week after that they took the Costa and McCrae (1992b) Revised NEO Personality Inventory (NEO-PI-R).

The factor analyses of the EPQ, ZKPQ, and NEO scales were based on a total of 157 subjects who took all three of these scales. Correlations between the subscales of these three questionnaires and other scales were based on *n*s varying from 135 to 177.

The ZKPQ has already been described, and the three scales of the EPQ (E, N, and P) are probably familiar to most readers. The revised NEO contains scores for five major factors, including E, N, Agreeableness, Conscientiousness, and Openness (to Experience), each of which is a sum of six "facet," or narrow trait scales. For instance, N is made up of subscales for anxiety, anger, hostility, depression, self-consciousness, impulsiveness, and vulnerability.

Other Scales Given and Rationale for Their Selection

On other weeks, subjects took response set tests (Crowne–Marlowe [1960] Social Desirability, Jackson [1974] Social Desirability, and Bass [1956] Social Acquiescence scales); the Sensation Seeking Scale, Form V (Zuckerman, 1979; Zuckerman, et al., 1978); the Buss and Plomin (1975) Emotionality, Activity, Sociability, Impulsivity (EASI-III) temperament scales; Block

Table 3
Correlations Among ZKPQ III Scales

	Study 1, all subjects, n = 589				Study 2, all subjects, n = 187			
Scale	N-Anx	Agg-Ho	Act	Sy	N-Anx	Agg-Ho	Act	Sy
ImpSS	−.09*	.25**	.15**	.28**	−.17*	.29**	.14	.37**
N-Anx	—	.11**	−.02	−.14**	—	.08	.03	−.14
Agg-Ho		—	.05	.11**		—	−.12	.22**
Act			—	.13**			—	.18*

Note. ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; N-Anx = Neuroticism–Anxiety; Agg-Ho = Aggression–Hostility; Act = Activity; Sy = Sociability; ImpSS = Impulsive Sensation Seeking.
* $p < .05$. ** $p < .01$.

and Block's (1980) Ego-Control and Ego-Resiliency scales; Ball and Zuckerman's (1990) Generalized Reward and Punishment Expectancy Scales (GRAPES); Fenigstein, Scheier, and Buss's (1975) Public and Private Self-Consciousness scales; and Rotter's (1966) Locus of Control scale.

The EASI III is a putative temperament questionnaire assessing four temperaments: emotionality, activity, sociability, and impulsivity. The distinction between temperament and basic personality is not well defined when we get down to actual measures in tests like this. There is an obvious resemblance between the EASI factors and the scales in the three models compared in this study, and therefore the EASI was included to assess the actual similarity and degree of correspondence with the three models.

The original Sensation Seeking Scales (SSS V) were included to clarify their position within the three models. Of course, they should correlate highly with the Zuckerman–Kuhlman ImpSS score, because it explicitly includes Sensation Seeking items but eliminates the specific interests and activities in the content of the original SSS items. Whereas ImpSS would be expected to correlate highly with the Total SSS score, it could be more highly related to one or more of the subscales than to the remaining ones.

The correlation of the SS scales with the scales in the other models is more problematic. Eysenck and Eysenck (1985) and Costa and McCrae (1992b) included sensation seeking (or "excitement seeking" in Costa and McCrae) as a component of E, but our analyses (Zuckerman et al., 1991) suggest that it is more closely identified with P and is also likely to be correlated with Conscientiousness. In their reanalysis of the data from that study within the framework of the Big Five model, Costa and McCrae (1992a) concluded that three of the SSS scales (experience seeking, thrill and adventure seeking, and boredom susceptibility, but especially experience seeking) were primarily related to Openness, but disinhibition definitely belonged within a Conscientiousness factor. Angleitner and Ostendorf (in press) had similar findings with regard to Openness, but in their study, disinhibition loaded primarily on agreeableness. The inclusion of the SSS in the present comprehensive study of all three models offers the chance to resolve the discrepancies in previous analyses.

Block has been a persistent critic of the Big Five and Eysenck models and defended a two-factor model consisting of Ego Control and Ego Resiliency (Block & Block, 1980). The scales designed to measure these two factors were included to clarify their position in relation to the other three models compared in this study. Ego Control (or undercontrol) seems to involve the traits of impulsivity and sensation seeking and is therefore a likely correlate of ImpSS and Conscientiousness. Ego Resiliency is a measure of adjustment and therefore has N as its most likely correlate.

Several cognitive-type scales were included. Users of cognitive measures often show little concern about the position of these types of scales within the broader factors of personality, even though they may have some obvious theoretical relevance to these factors. Rotter's Locus of Control (external) has been correlated with Sensation Seeking (Zuckerman, 1979) and a Social Deviancy factor (Daitzman & Zuckerman, 1980), suggesting that it might be related to P, ImpSS, and Conscientiousness.

Ball and Zuckerman's (1990) Generalized Reward and Punishment Expectancy Scales (GRAPES) contains scales for generalized expectancies of reward and punishment that are of particular theoretical significance. Gray (1982, 1991) maintained that extraverts have a relatively greater sensitivity to stimuli associated with reward, whereas introverts have a relatively greater sensitivity to stimuli associated with punishment. Neurotics are said to be more sensitive to stimuli conditioned to both reward and punishment, compared with emotionally stable persons, and equally sensitive to both types of stimuli. It is true that the construct of "sensitivity to stimulation," derived from the comparative literature, is not the same as "generalized expectancy," but the latter is more applicable to humans and probably relevant to the more behavioral construct. A human who is sensitive to certain types of stimuli is likely to anticipate or expect such stimuli in the future. Ball and Zuckerman (1990), using the GRAPES, found E correlated positively with generalized expectancy of reward, but not with generalized expectancy of punishment. N was related negatively to reward expectancy and positively to punishment expectancy. Sensation Seeking was positively related to reward and negatively related to punishment expectancies, but the latter was stronger than the former. The present study offered the chance to replicate these findings and extend them from the Eysenckian context to the Costa–McCrae and Zuckerman–Kuhlman models.

Fenigstein et al. (1975) related the concept of private self-consciousness to the Jungian conception of introversion [presumably thinking-introversion]. Public self-consciousness sug-

gests a sensitivity to the reactions of others and a neurotic kind of social dependence. However, Fenigstein et al. did not interpret this aspect of self-consciousness as a neurotic tendency, but as a necessary antecedent of such anxiety. Their third scale of social anxiety is clearly relevant to the N dimension in all three models.

Despite the question of the existence of social desirability (SD) as a style of personality confounding other dimensions (McCrae & Costa, 1983), its relationships to these dimensions is a matter of interest. A substantial relationship between SD and a scale does not indicate that the scale is confounded by conscious or unconscious dissimulation in self-report, but it does indicate that the items of that scale are heavily weighted by a general evaluation factor that does have reliability as a separately measureable individual trait. Crowne and Marlowe (1964) conceived of this trait as the "need for approval."

Results

Factor analyses of 13 scales from the three major questionnaires were done with principal-components analysis followed by varimax rotations. Figure 1 shows the plot of the 13 eigenvalues. Eigenvalues began to level off after five factors, and factors beyond the sixth showed little change. But beyond the first four factors, eigenvalues dropped below 1.0, usually considered a minimal value for factor extraction. Furthermore, factors beyond the first four consisted of only one scale, and a minimum of two markers is necessary to define a factor. Three factors accounted for 63% and four factors for 74% of the variance.

The results of the three-factor analysis are shown in Table 4. The first factor is composed of the E or Sy scales from all three models plus the ZKPQ Act scale, and the second factor consists of the three N scales with high loadings from all three. The third factor has its highest loading from the EPQ-R P scale, with additional loadings from the NEO Agreeableness and Conscientiousness scales and the ZKPQ ImpSS and Agg-Host scales. The three-factor results are a good fit to the one proposed by Eysenck. The third factor contains scales for aggression (and agreeableness), and impulsivity, sensation seeking (and conscientiousness). That the P scale is the best marker for the factor suggests that the scale is a good measure of the broad trait regardless of how it is conceptualized.

Table 5 shows the results of the four-factor rotation. The first and second factors represent E and N, with the appropriate scales from each of the models, as in the three-factor analysis. The broad third factor from that analysis is broken into two in the four-factor analysis. The third factor is defined by NEO

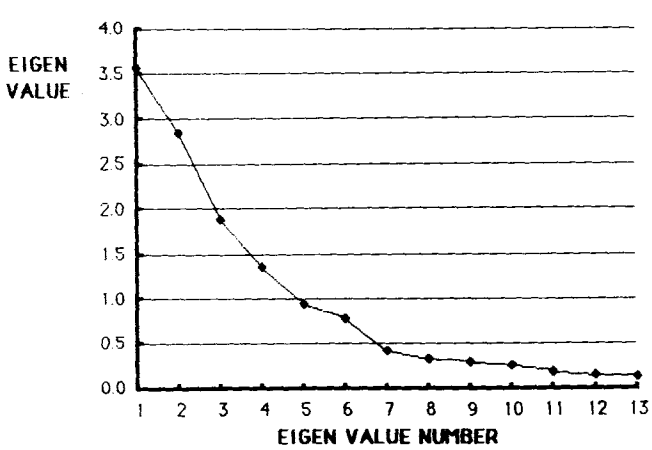


Figure 1. Scree plot for the factor analysis of the five NEO Personality Inventory, three Eysenck Personality Questionnaire, and five Zuckerman–Kuhlman Personality Questionnaire scales.

Table 4
Three-Factor Analysis

Scale	Factor loadings		
	Factor 1	Factor 2	Factor 3
NEO Extraversion	**.89**	−.17	−.07
EPQ Extraversion	**.76**	−.35	.23
ZKPQ Sociability	**.75**	−.19	.17
ZKPQ Activity	**.60**	.00	−.11
NEO Openness	.35	.11	−.21
ZKPQ N-Anx	−.10	**.93**	−.03
NEO Neuroticism	−.13	**.90**	.17
EPQ Neuroticism	−.12	**.92**	.03
EPQ Psychoticism	−.12	−.11	**.80**
NEO Agreeableness	.06	−.06	**−.72**
NEO Conscientiousness	.14	−.03	**−.68**
ZKPQ ImpSS	.46	.02	**.65**
ZKPQ Agg-Ho	.27	.32	**.63**

Note. Three factors accounted for 63% of the variance. Defining factor loadings are in boldface. NEO = NEO Personality Inventory; EPQ = Eysenck Personality Questionnaire; ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; N-Anx = Neuroticism–Anxiety; ImpSS = Impulsive Sensation Seeking; Agg-Ho = Aggression–Hostility.

Table 5
Four-Factor Analysis

Scale	Factor loadings			
	Factor 1	Factor 2	Factor 3	Factor 4
NEO Extraversion	**.88**	−.14	−.05	.17
EPQ Extraversion	**.79**	−.32	.17	−.08
ZKPQ Sociability	**.76**	−.16	.10	−.07
ZKPQ Activity	**.60**	.01	−.18	.02
ZKPQ N-Anx	−.13	**.92**	−.01	.08
NEO Neuroticism	−.15	**.90**	.10	−.11
EPQ Neuroticism	−.16	**.91**	−.04	−.08
NEO Conscientiousness	.15	−.07	**−.86**	−.02
EPQ Psychoticism	−.09	−.08	**.80**	−.28
ZKPQ ImpSS	.48	.08	**.74**	−.02
NEO Agreeableness	−.04	−.07	−.31	**.81**
ZKPQ Agg-Ho	.35	.34	.24	**−.72**
NEO Openness	.27	.14	.18	**.67**

Note. Four factors account for 74% of the variance. Defining factor loadings are in boldface. NEO = NEO Personality Inventory; EPQ = Eysenck Personality Questionnaire; ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; N-Anx = Neuroticism–Anxiety; ImpSS = Impulsive Sensation Seeking; Agg-Ho = Aggression–Hostility.

Conscientiousness, EPQ P, and ZKPQ ImpSS. The fourth factor is composed of NEO Agreeableness and Openness and of ZKPQ Agg-Host.

Although the eigenvalues do not warrant extraction and rotation of more than four factors, we decided to look at the results of five- and six-factor analyses, accounting for 81% and 87% of the variance. The results of the five-factor rotation were the same as those for the four-factor rotation, except that Act was separated from the E factor to form the single-variable fifth factor. Similarly, in the six-factor analysis, Openness was separated from the Agreeableness versus Agg-Host factor to form its own factor.

The lack of confirmation for Openness and Act can be accounted for by the fact that they were each only represented by a single scale within one of the three models. The revised NEO, however, contains six facet scales for Openness and one facet scale for Act, subsumed under E, which together with EPQ and ZKPQ scales might allow identification of these factors as independent ones or as components of other factors. The ZKPQ and EPQ scales and the 30 facet scales of the NEO were factor analyzed and rotated to a five-factor solution accounting for 62% of the variance (Table 6).

Factor 1 is N, defined by N-Anx from the ZKPQ, EPQ N, and four of the six N facet scales from the NEO. Factor 2 was composed of all six Conscientiousness facet scales from the NEO at one pole, and ZKPQ ImpSS, EPQ P, and the NEO Impulsiveness facet scale. This latter scale is assigned to the N factor in the NEO but clusters with other impulsivity and conscientiousness scales in this study.

Factor 3 is clearly E with ZKPQ Sy, EPQ E, and all six E facet scales from the NEO. Act from the ZKPQ also has a somewhat weaker loading on the factor. Factor 4 may be defined as Agreeableness versus Agg-Host, because it contains all six agreeableness facet scales from the NEO, Agg-Host from the ZKPQ, and the Anger–Hostility facet scale from the NEO. The latter scale is normally scored as a component of the N factor in the NEO, and it does load almost as highly on this factor as on Agg-Host in our analysis. Factor 5 consists entirely of the six NEO facet scales defining the Openness factor. Eysenck's P scale had no loading of any magnitude on this factor. A few other NEO facet scales had secondary loadings on Openness including N5, Impulsivity; E6, Positive Emotions; and A3, Altruism.

The NEO, EPQ, and ZKPQ scales were correlated with the other test scores taken by the class. The results for the three factors that have representative scales in all three questionnaires are shown in Table 7. The three N scales all correlated highly with the Emotionality score from the EASI; the three E scales all correlated with EASI Sociability; and the factor defined by P, ImpSS, and Conscientiousness (P-ImpUSS) correlated with the EASI Impulsivity factor.

The P-ImpUSS factor scales (P, ImpSS, and Conscientiousness) were also the ones correlating most highly with all of the Sensation Seeking subscales and the total score. The ImpSS scale itself correlated highly with the SSS total, as expected, but it also correlated nearly equally with all of the subscales. E and N scales correlated significantly and consistently only with the TAS subscale of the SSS.

The P-ImpUSS factor scales, particularly ImpSS, correlated very highly with Block's Ego Control scale; EPQ E and P correlated equally with undercontrol. Block's Ego Resiliency correlated primarily with the N and E factors.

The Locus of Control scale showed few correlations of any magnitude with the factors, but all correlations with N were significant. GRAPES Generalized Reward Expectancy correlated negatively with all N and positively with all E factor scales, but GRAPES Generalized Punishment Expectancy correlated to any degree with N only. Fenigstein's Private and Public Self-Consciousness scales showed little consistent correlation with the scales comprising the primary factors, but the Social Anxiety scale had a clear pattern of positive correlation with the Anxiey and negative correlation with the Sociability factors.

The Bass Social Acquiesence Scale showed little correlation with any of the factors despite the scales from the EPQ and ZKPQ being strongly imbalanced between true and false items. The correlations of SD, particularly as measured by the Jackson scale, were strongly negative with N scales and moderately positive with E scales. The Crowne–Marlowe SD scale also correlated negatively with N scales, although to a lower degree than the Jackson scale. Conscientiousness correlated positively and P correlated negatively with the Crowne–Marlowe, Jackson SD, and EPQ Lie scales, but the ZKPQ ImpSS scale was significantly correlated only with the Lie scale. Agreeableness correlated positively, Agg-Host correlated negatively, and Openness correlated positively with all three SD scales.

Discussion

The Big Three and Alternative Five emerged from psychobiological approaches to personality (H. J. Eysenck, 1967; H. J. Eysenck & Eysenck, 1985; Zuckerman, 1983; Zuckerman, 1991b). Eysenck's Big Three paradigm has been the framework for a half century of experimental and correlational research in personality, although until recently most of this dealt with the E and N dimensions. P was added as a third dimension to the standard questionnaire in the mid-1970s (S. B. G. Eysenck & Eysenck, 1975), and a theoretical model for the dimension soon followed (H. J. Eysenck & Eysenck, 1976). The existence or significance of the P dimension and its psychometric characteristics were immediately challenged (Bishop, 1977; Block, 1977a, 1977b), and the Eysencks' theory of the nature of P and its relationship to actual psychoses were contested somewhat later (Claridge, 1981; Zuckerman, 1989).

The Alternative Five emerged from factor analytic studies of personality and temperament scales, most of which had already been used in psychobiological studies (Zuckerman, 1991b). Surprisingly, Eysenck's P scale proved to be the best marker for a basic dimension of personality, which also included scales for Impulsivity, Sensation Seeking, and Socialization (Zuckerman et al., 1988). Beyond the Big Three, a highly replicable five-factor analysis separated Agg-Host from its placement with anxiety in a general Emotionality factor, and Act from its location with Sy in an E factor, at the three-factor level.

The distinction between Agg-Host and anxiety is important because aggression (Valzelli, 1981) and anxiety (Gray, 1982) have distinctive psychobiological bases and should not be confounded within a single factor as they are in general Emotionality at the three-factor level. In the Costa and McCrae (1992a)

Table 6
Five Factor Analysis of EPQ, ZKPQ, and NEO Facet Scales

Scale	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
ZKPQ N-Anx	.89	−.03	−.08	.05	.06
EPQ Neuroticism	.86	.03	−.12	−.15	.04
N3-Depression	.84	−.18	−.20	−.07	.04
N1-Anxiety	.82	.18	−.06	.00	.11
N6-Vulnerability	.77	−.17	−.11	−.05	−.09
N4-Self-Conscious	.77	−.01	−.21	.12	−.03
C4-Achievement	−.02	.84	.20	−.13	.10
C5-Self-Discipline	−.18	.83	.02	.05	−.04
C6-Deliberation	.08	.76	−.25	.23	−.02
C3-Dutifulness	.00	.74	.14	.10	.09
C1-Competence	−.36	.73	.16	−.01	.09
C2-Order	.07	.66	−.01	.00	−.04
ZKPQ ImpUSS	−.06	−.59	.41	−.24	.12
EPQ Psychoticism	−.13	−.58	−.14	−.43	.07
N5-Impulsivity	.30	−.44	.29	−.27	.36
ZKPQ-Sociability	.16	−.11	.81	−.03	−.14
E2-Gregariousness	−.09	−.05	.80	.12	−.07
E1-Warmth	−.05	.14	.76	.36	.19
EPQ Extraversion	−.37	−.09	.73	−.21	.04
E4-Activity	−.07	.32	.65	−.18	.18
E5-Excitement Seeking	.01	−.32	.64	−.17	.27
E6-Positive Emotions	−.24	.13	.59	.25	.36
E3-Assertiveness	−.29	.20	.58	−.41	.10
ZKPQ-Activity	−.10	.29	.41	−.07	.09
A4-Compliance	−.08	−.02	−.18	.81	.13
ZKPQ-Agg-Ho	.25	−.08	.25	−.74	−.13
A2-Straightforwardness	.10	.26	−.11	.67	.02
A1-Trust	−.39	.16	.23	.61	.04
A3-Altruism	.08	.29	.35	.61	.35
N2-Anger–Hostility	.57	−.01	−.04	−.60	−.06
A5-Modesty	.34	−.10	−.17	.57	−.07
A6-Tendermindedness	.01	.07	.32	.50	.21
O2-Aesthetics	.08	.08	.06	.14	.80
O5-Ideas	−.29	.15	−.09	.01	.69
O3-Feelings	.31	.23	.21	.07	.68
O1-Fantasy	.14	−.33	.00	−.01	.62
O6-Values	.02	−.04	.13	.28	.50
O4-Actions	−.28	−.21	.25	.04	.43

Note. Five factors accounted for 62% of the variance. Numbered scales are from the revised NEO Personality Inventory and designated by their placement among the five major trait scores: N = Neuroticism, C = Conscientiousness, E = Extraversion, A = Agreeableness, O = Openness to Experience. EPQ = Eysenck Personality Questionnaire; ZKPQ = Zuckerman–Kuhlman Personality Questionnaire; N-Anx = Neuroticism–Anxiety; ImpUSS = Impulsive Unsocialized Sensation Seeking; Agg-Ho = Aggression–Hostility.

version of the Big Five, "angry-hostility" is subsumed in the N factor along with anxiety and depression, although logically one would expect it to be the opposite pole of Agreeableness.

Nearly all classification of temperament in children includes activity as a basic factor (Kohnstamm, Bates, & Rothbart, 1989). In both the Big Three and Big Five, activity is subsumed as a component of E (although Costa and McCrae's [1992a] factor analysis shows a higher loading on the Conscientiousness factor). Our factor analyses (Zuckerman et al., 1988, Zuckerman et al., 1991) also show activity encompassed within a broad E trait at the three-factor level, but with a negative relationship to N-Anx as well. Despite this dual influence of E and Anxiety on activity, it did emerge as a distinct factor at the five-factor level in earlier analyses (Zuckerman et al., 1991). Because of its developmental and psychobiological significance, activity level merits a distinctive assessment as a major trait of temperament or personality. Buss and Plomin (1984) included it as one of their Big Three (Emotionality, Activity, and Sociability) in their revised theory of temperament.

The conventional Big Five developed from lexical analyses of trait terms in language systems and their translation into simple rating scales (Digman, 1990; Goldberg, 1990). The factors probably reflect the importance and observability of trait concepts in social interactions. However, the sampling of traits in the language may not reflect their importance as basic traits of personality. Surgency (E) and Agreeableness always emerge as the two strongest factors in language analyses, but Emotional Stability (N) comes out fourth in size, indicating that there are

Table 7
Correlations Between NEO, ZKPQ, and EPQ Scales and Other Tests

Scale	NEO-N	ZKPQ N-Anx	EPQ-N	NEO-E	ZKPQ-Sy	EPQ-E	NEO-Consc.	ZKPQ-ImpSS	EPQ-P	NEO Agree	ZKPQ-Agg-Ho	NEO Open	ZKPQ-Act
EASI													
Emotionality	**.74****	**.68****	**.72****	−.16	−.13	−.18*	−.11	−.11	−.02	−.29**	.46**	−.11	−.14
Sociability	−.16*	−.28**	−.25**	**.70****	**.67****	**.69****	.10	.25**	−.12	−.09	.24**	.08	.27**
Impulsivity	.14	−.06	.02	.09	.23**	.31**	**−.63****	**.70****	**.56****	−.41**	.42**	.07	−.03
Activity	−.12	−.18	−.10	.52**	.19*	.38**	.38**	.08	−.17*	.04	.11	.16	**.59****
SSS													
Total	.05	−.10	−.06	.21*	.20*	.31**	**−.47****	**.66****	**.55****	−.37**	.31**	.13	.01
TAS	−.24**	−.24**	−.19*	.30**	.21*	.33**	−.26**	.49**	.21**	−.09	.11	.02	.17*
ES	.09	.00	−.05	.07	−.05	.17*	−.36**	.46**	.45**	−.04	.07	**.43****	−.12
Dis	.21*	.05	.10	.13	.23**	.13	−.41**	.48**	.43**	−.40**	.36**	.02	−.09
BS	.07	−.08	−.01	.07	.15	.23**	−.23**	.37**	.43**	−.48**	.33**	−.14	.08
Block													
Ego UC	.06	−.15	−.07	.39**	.24**	.48**	**−.53****	**.63****	**.46****	−.36**	.30**	.32**	.11
Ego Res	−.34**	−.30**	−.30**	.51**	.17*	.39**	.18*	.14	−.09	.34**	−.24**	.44**	.23**
Rotter LocusC	−.35**	−.31**	−.28*	.23**	.09	.09	.29*	−.12	−.08	−.02	.04	.07	.10
GRAPES													
Rew Exp	**−.45****	**−.49****	**−.43****	**.51****	**.39****	**.54****	.23**	.23**	−.08	−.11	.09	.21*	.34**
Pun Exp	**.40****	**.46****	**.54****	−.08	−.13	−.18*	.14	−.10	−.19*	−.13	.11	−.05	−.06
Self Cs													
Private	.15	.22*	.30**	.08	−.08	−.03	.13	.00	−.07	.04	−.11	.34**	.01
Public	.18*	.16	.26**	.05	.04	−.06	.22*	−.11	−.13	−.15	.10	−.01	−.01
Soc Anx	**.47****	**.44****	**.45****	**−.53****	**−.39****	**−.56****	−.07	−.22*	.00	−.05	.02	−.23**	−.27*
Res. Set													
C-M SD	−.34**	−.27**	−.28**	.21**	.01	.05	.29**	−.05	−.15*	.40**	−.33**	.21**	.25**
Jac. SD	−.58**	−.54**	−.53**	.41**	.28**	.42**	.43**	−.09	−.35**	.30**	−.21**	.10	.24**
Bass SA	−.08	−.09	−.20**	−.13	.01	−.11	−.06	−.01	−.01	−.01	−.09	.00	−.08
EPQ Lie	−.20*	−.07	−.10	−.09	−.08	−.18*	.30**	−.30**	−.27*	.34**	−.29**	.00	.18*

Note. N = Neuroticism; N-Anx = Neuroticism–Anxiety; E = Extraversion; Sy = Sociability, Consc. = Conscientiousness, ImpSS = Impulsive Sensation Seeking. P = psychoticism; Agree = Agreeableness, Agg-Ho = Aggression–Hostility, Open. = Openness, Act = Activity. EASI = Buss–Plomin Temperament scales; SSS = Sensation Seeking Scale (Form V): Total = total score (sum of four subscales), TAS = Thrill and Adventure Seeking, ES = Experience Seeking, Dis = Disinhibition, BS = Boredom Susceptibility; Block = Block Ego Scales: Ego UC = Ego (Under) Control, Ego Res = Ego Resiliency; LocusC = Rotter Locus of Control Scale; GRAPES = Ball and Zuckerman (1990) Generalized Reward and Punishment Expectancy Scales: Rew Exp = Reward Expectancy, Pun Exp = Punishment Expectancy, Self Cs = Fenigstein, Scheier, and Buss (1975) Self-Consciousness Scales: Private = Private Self-Consciousness, Public = Public Self-Consciousness, Soc Anx = Social Anxiety; Res. Set = Response Set Scales: C-M SD = Crowne–Marlowe Social Desirability scale, Jac. SD = Jackson Social Desirability Scales, Bass SA = Bass Social Acquiescence Scale, EPQ Lie = Eysenck Personality Questionnaire Lie scale.
* $p < .05$, two-tailed test. ** $p < .01$, two tailed test.

fewer words to represent it. In analyses of questionnaire scales, however, N almost always emerges, along with E, as one of the two strongest factors.

Traits such as Impulsivity and Sensation Seeking, which have important psychobiological bases (Zuckerman, 1983, 1984, 1991b), are only represented as single scales within the larger factors of the Big Five and are listed under different major factors: "impulsiveness" is regarded as a facet of N and "excitement seeking" as a facet of E (Costa & McCrae, 1992b). Similarly, H. J. Eysenck and Eysenck (1985) placed sensation seeking under E and classified impulsivity under P. However, sensation seeking and impulsivity are closely related and with socialization form a distinctive factor in five-factor analyses of scales (Zuckerman et al., 1991) and items.

Despite these differences between the three models, the results of this empirical analysis of the major factors in all models suggest a great deal of convergence between them, particularly the two five-factor models. E and N are quite similar across all three models, despite some differences in their components. Restricting the analysis to three factors produces a model very much like that proposed by H. J. Eysenck and M. W. Eysenck (1985). The first factor is E, including sociability and activity, and the second is N, with very close identity of the measures from all three models. The third factor looks very much like P as described by the Eysencks, with Agreeableness, Conscientiousness, Aggression, and ImpSS included and the P scale itself as the strongest marker. ImpSS has its major loading on the P factor, as predicted from the Zuckerman–Kuhlman model, but ImpSS does have a moderate secondary loading on the E factor as well.

In the four-factor analysis, P is identified with Conscientiousness and ImpSS, but Agreeableness from the NEO and Agg-Host from the ZKPQ form a separate dimension. Thus, the two five-factor models show a high degree of convergence on four of the five basic factors.

Activity in the alternative five-factor model is not represented by a major dimension in the other two models so could not emerge as a separate factor defined by at least two scales. The Act factor in the ZKPQ correlated .52 with the Activity facet scale in the NEO and .59 with the EASI Activity scale. How-

ever, in the five-factor analysis using the NEO facet scales along with the EPQ and ZKPQ scales, NEO and ZKPQ Act scales both loaded most strongly on the E dimension. Act seems to be the weakest factor of our alternative five and from a statistical viewpoint alone, one could argue for reducing the major factors to four and including activity in the E dimension. However, the viability of a basic dimension goes beyond its strength in factor analyses and to the question of its theoretical and biological significance (H. J. Eysenck, 1992a, 1992b; Zuckerman, 1992); consequently, judgment should be postponed until there are enough data to demonstrate the construct validity of Act. The same conclusion can be applied to all of the dimensions beyond E and N from any of the models.

Like Activity, Openness could not be identified as a factor in factor analyses of the three scales because it had no direct corresponding measure in the EPQ and ZKPQ. In the four-factor analysis, it loaded positively on the Agreeableness versus Agg-Host factor. However, when we used the facet scales instead of the total factor scales for the NEO, all six of the Openness facet scores formed the fifth factor. However, despite previous data from the Angleitner and Ostendorf (in press) study suggesting that Openness might have some relationship with P, Sensation Seeking, or Agreeableness, little relationship was found between this factor and most of the scales from the other four factors. The strongest correlations between the Openness total score and other scales were with Block's Ego Resiliency scale ($r = .44$) and the Experience Seeking subscale of the SSS ($r = .43$).

The patterns of correlations between corresponding scales from the three models and other tests were quite similar. Three of the four scales in the EASI scales of temperament correlate with E, N, and P factors in the three models, and EASI activity correlates highly with ZKPQ Act. P, ImpSS, and Conscientiousness all correlate moderately with all of the Sensation Seeking subscales, but Disinhibition and Boredom Susceptibility also correlate with Agg-Host and Agreeableness, as we might expect from their relationships with antisocial types of Sensation Seeking (Zuckerman, 1979).

The relationships of generalized reward and punishment expectancy scales from the GRAPES with E and N factors are similar to what was found in the Ball and Zuckerman (1990) study: E is related only to reward expectancy, whereas N is related to both generalized punishment and reward expectancies (positively to the former and negatively to the latter). The results provide only partial support for Gray's theory if selective "stimulus sensitivity" and "generalized expectancy" can be equated. Whereas his theory suggests that E is related positively to reward and negatively to punishment stimulus sensitivity, the findings from this and the previous study show that E is only related to reward expectancy. His theory also predicts positive relationships between N and punishment and reward sensitivities, but the present data show a positive correlation with punishment and a negative correlation with reward expectancy.

Self-consciousness scales show little relationship to these personality factors. Private self-consciousness is not related to Introversion as suggested by the test authors' theory, but seems to be a weaker form of social anxiety somewhat related to N. But social anxiety seems to represent a combination of N and Introversion.

Acquiescence does not correlate significantly overall with the scales involved in these models, which is reassuring considering the imbalanced true–false keying of some of them. Acquiescence has been discounted as a response set influence on most scales for some time (Block, 1965). SD, particularly of the type measured by the Jackson scale, does correlate substantially with measures of N and E in all models and measures of Conscientiousness and P from the NEO and EPQ. This does not indicate that these scales are somehow confounded or less valid because of a response set influence. McCrae and Costa (1983) have shown that correcting the correlations between NEO scales and spouse ratings for scores on the Marlowe–Crowne (1960) SD and Eysenck Lie scales had no effect on the spouse–self validity coefficients. Even in the current study, these two scales did not correlate very highly with most of the three model scales. The Jackson SD scale, like the now seldom used Edwards (1957) SD scale, seems to be more loaded with the type of psychopathology contained in N-Anx scales. The ImpSS measure was unrelated to SD except for a low negative correlation with the EPQ Lie scale. This finding has been generally true for Sensation Seeking measures (Zuckerman, 1979).

Correlations between personality questionnaires primarily reflect similarities in content domains. The crucial tests of convergence or specificity will depend on use of the measures from the three models in studies of behavioral and biological phenomena. These tests will also determine whether the absence of an Openness factor in the Big Three and Alternative Five models, or a separate activity factor in the Big Three and Big Five models, or a separate Agg-Host (vs. Agreeableness) factor in the Big Three are important deficits in the respective models.

References

Angleitner, A. (1992, June). *Symposium on Temperament and the Five Factor Model.* Presented at the Sixth Conference of the European Association for Personality Psychology, Groningen, The Netherlands.

Angleitner, A., & Ostendorf, F. (in press). Temperament and the Big Five factors of personality. In C. F. Halverson, G. A. Kohnstamm, & R. P. Martin (Eds.), *The developing structure of temperament and pesonality from infancy to adulthood.* Hillsdale, NJ: Erlbaum.

Ball, S. A., & Zuckerman, M. (1990). Sensation seeking, Eysenck's personality dimensions and reinforcement sensitivity in concept formation. *Personality and Individual Differences, 11,* 343–353.

Bass, B. M. (1956). Development and evaluation of a scale for measuring social acquiescence. *Journal of Abnormal and Social Psychology, 53,* 296–299.

Bishop, D. V. M. (1977). The P scale and psychosis. *Journal of Abnormal Psychology, 86,* 127–134.

Block, J. (1965). *The challenge of response sets.* New York: Appleton-Century Crofts.

Block, J. (1977a). The Eysencks and psychoticism. *Journal of Abnormal Psychology, 86,* 653–654.

Block, J. (1977b). P scale and psychosis: Continued concerns. *Journal of Abnormal Psychology, 86,* 431–434.

Block, J. H., & Block, J. (1980). The role of ego-control and ego-resiliency in the organization of behavior. In W. A. Collins (Ed.), *The Minnesota Symposia on Child Psychology: Vol. 13. Development of cognition, affect, and social relations* (pp. 39–101). Hillsdale, NJ: Erlbaum.

Buss, A. H., & Plomin, R. (1975). *A temperament theory of personality development.* New York: Wiley.

Buss, A. H., & Plomin, R. (1984). *Temperament: Early developing personality traits.* Hillsdale, NJ: Erlbaum.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* New York: Harcourt, Brace, & World.

Claridge, G. (1981). Psychoticism. In R. Lynn (Ed.), *Dimensions of personality* (pp. 79–109). Oxford, England: Pergamon Press.

Costa, P. T., Jr., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653–665.

Costa, P. T., Jr., & McCrae, R. R. (1992b). *NEO-PI-R: Revised NEO Personality Inventory (NEO-PI-R).* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992c). Reply to Eysenck. *Personality and Individual Differences, 13,* 861–865.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive.* New York: Wiley

Daitzman, R., & Zuckerman, M. (1980). Disinhibitory sensation seeking, personality and gonadal hormones. *Personality and Individual Differences, 1,* 103–110.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Edwards, A. L., (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

Eysenck, H. J. (1947). *Dimensions of personality.* New York: Praeger.

Eysenck, H. J. (1967). *The biological basis of personality.* Springfield, IL: Charles C Thomas.

Eysenck, H. J. (1992a). A reply to Costa and McCrae. P or A and C—the role of theory. *Personality and Individual Differences, 13,* 867–868.

Eysenck, H. J. (1992b). Four ways five factors are not basic. *Personality and Individual Differences, 13,* 667–673.

Eysenck, H. J. (1992c). The definition and measurement of psychoticism. *Personality and Individual Differences, 7,* 757–785.

Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum Press.

Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement.* San Diego, CA: Robert R. Knapp.

Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality.* New York: Crane, Russak.

Eysenck, S. B. G., & Eysenck, H. J. (1975). *Manual of the EPQ (Personality Questionnaire).* San Diego, CA: Educational and Industrial Testing Service.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences, 6,* 21–29.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43,* 522–527.

Fiske, D. W. (1949). Consistency of factorial structures for personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44,* 329–344.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Goldberg, L. R. (1992, June). *The structure of phenotypic personality traits (or the magical number five, plus or minus zero).* Paper presented at the Sixth Conference of the European Association for Personality Psychology, Groningen, The Netherlands.

Goldberg, L. R., & Rosolack, T. K. (in press). Putting the Big Five factor structure against its competitors: An empirical comparison with Eysenck's P. E. N. model. In C. F. Halverson, G. A. Kohnstamm, & R. P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood.* Hillsdale, NJ: Erlbaum.

Gray, J. A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system.* New York: Oxford University Press.

Gray, J. A. (1991). *The neuropsychology of temperament. Explorations in temperament: International perspectives on theory and measurement* (pp. 105–128). New York: Plenum Press.

Guilford, J. P., & Zimmerman, W. S. (1956). Fourteen dimensional temperament factors. *Psychological Monographs, 70*(10), 1–26.

Halverson, C. F., & Kohnstamm, G. A. (1991, June). *Conference on the development of the structure of temperament and personality from infancy to adulthood.* Netherlands Institute for Advanced Study, Wassenaar, The Netherlands.

Jackson, D. N. (1974). *Personality Research Form manual.* Goshen, NY: Consulting Psychologists Press.

Kohnstamm, G. A., Bates, J. E., & Rothbart, M. K. (1989). *Temperament in childhood.* New York: Wiley.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51,* 882–888.

McCrae, R. R., & Costa, P. T., Jr. (1985). Comparison of EPI and Psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences, 6,* 587–597.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure. *Journal of Abnormal and Social Psychology, 66,* 574–583.

Rotter, J. B. (1966). Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs, 80*(Whole No. 609).

Strelau, J. (1983). *Temperament, personality, activity.* New York: Academic Press.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681–706). Hillsdale, NJ: Erlbaum.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings.* U.S. Air Force, ASD Technical Reports (No. 61–79).

Valzelli, L. (1981). *Psychobiology of aggression and violence.* New York: Raven Press.

Zuckerman, M. (1979). *Sensation seeking: Beyond the optimal level of arousal.* Hillsdale, NJ: Erlbaum.

Zuckerman, M. (1983). *Biological bases of sensation seeking, impulsivity and anxiety.* Hillsdale, NJ: Erlbaum.

Zuckerman, M. (1984). Sensation seeking: A comparative approach to a human trait. *Behavioral and Brain Sciences, 7,* 413–471.

Zuckerman, M. (1989). Personality in the third dimension: A psychobiological approach. *Personality and Individual Differences, 10,* 391–418.

Zuckerman, M. (1991a). *Symposium: Basic dimensions of personality.* Presented at the meeting of the International Society for the Study of Individual Differences, Oxford, England, July, 1991.

Zuckerman, M. (1991b). *Psychobiology of personality.* New York: Cambridge University Press.

Zuckerman, M. (1992). What is a basic factor and which factors are basic? Turtles all the way down. *Personality and Individual Differences, 13,* 675–682.

Zuckerman, M., Buchsbaum, M. S., & Murphy, D. L. (1980). Sensation seeking and its biological correlates. *Psychological Bulletin, 88,* 187–214.

Zuckerman, M., Eysenck, S. B. G., & Eysenck, H. J. (1978). Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology, 46,* 139–149.

Zuckerman, M., Kuhlman, D. M., & Camac, C. (1988). What lies beyond E and N? Factor analyses of scales believed to measure basic dimensions of personality. *Journal of Personality and Social Psychology, 54,* 96–107.

Zuckerman, M., Kuhlman, D. M., Thornquist, M., & Kiers, H. (1991). Five (or three): Robust questionnaire scale factors of personality without culture. *Personality and Individual Differences, 12,* 929–941.

Received November 18, 1992
Revision received March 12, 1993
Accepted March 16, 1993 ■

Call for Nominations

The Publications and Communications Board has opened nominations for the editorships of *Behavioral Neuroscience*, the *Journal of Experimental Psychology: General*, and the *Journal of Experimental Psychology: Learning, Memory, and Cognition* for the years 1996–2001. Larry R. Squire, PhD, Earl Hunt, PhD, and Keith Rayner, PhD, respectively, are the incumbent editors. Candidates must be members of APA and should be available to start receiving manuscripts in early 1995 to prepare for issues published in 1996. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. To nominate candidates, prepare a statement of one page or less in support of each candidate.

- For *Behavioral Neuroscience*, submit nominations to J. Bruce Overmier, PhD, Elliott Hall—Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455 or to psyjbo@vx.cis.umn.edu. Other members of the search committee are Norman Adler, PhD, Evelyn Satinoff, PhD, and Richard F. Thompson, PhD.

- For the *Journal of Experimental Psychology: General*, submit nominations to Howard E. Egeth, PhD, Chair, *JEP: General* Search, Department of Psychology, Johns Hopkins University, Charles & 34th Streets, Baltimore, MD 21218, to egeth@jhuvm.bitnet, or to fax number 410-516-4478. Other members of the search committee are Donald S. Blough, PhD, Martha Farah, PhD, and Edward E. Smith, PhD.

- For the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, submit nominations to Donna M. Gelfand, PhD, Dean, Social and Behavioral Science, 205 Osh, University of Utah, Salt Lake City, UT 84112-1102 or to fax number 801-585-5081. Other members of the search committee are Marcia Johnson, PhD, Michael Posner, PhD, Henry L. Roediger III, PhD, and Richard M. Shiffrin, PhD.

First review of nominations will begin December 15, 1993.

Development and Validation of a Screening Instrument for Bipolar Spectrum Disorder: The Mood Disorder Questionnaire

Robert M.A. Hirschfeld, M.D.
Janet B.W. Williams, D.S.W.
Robert L. Spitzer, M.D.
Joseph R. Calabrese, M.D.
Laurie Flynn, B.S.
Paul E. Keck, Jr., M.D.
Lydia Lewis, B.A.
Susan L. McElroy, M.D.
Robert M. Post, M.D.
Daniel J. Rapport, M.D.
James M. Russell, M.D
Gary S. Sachs, M.D.
John Zajecka, M.D.

Objective: Bipolar spectrum disorders, which include bipolar I, bipolar II, and bipolar disorder not otherwise specified, frequently go unrecognized, undiagnosed, and untreated. This report describes the validation of a new brief self-report screening instrument for bipolar spectrum disorders called the Mood Disorder Questionnaire.

Method: A total of 198 patients attending five outpatient clinics that primarily treat patients with mood disorders completed the Mood Disorder Questionnaire. A research professional, blind to the Mood Disorder Questionnaire results, conducted a telephone research diagnostic interview by means of the bipolar module of the Structured Clinical Interview for DSM-IV.

Results: A Mood Disorder Questionnaire screening score of 7 or more items yielded good sensitivity (0.73) and very good specificity (0.90).

Conclusions: The Mood Disorder Questionnaire is a useful screening instrument for bipolar spectrum disorder in a psychiatric outpatient population.

(Am J Psychiatry 2000; 157:1873–1875)

The lifetime prevalence of bipolar I disorder is approximately 1% (1, 2). However, the prevalence of bipolar spectrum disorder is substantially higher. Bipolar spectrum disorder has been described and defined in several ways (3, 4), but it usually includes bipolar I, bipolar II, cyclothymia, and bipolar disorder not otherwise specified. The lifetime prevalence of bipolar spectrum disorder has been found to be between 2.6% and 6.5% (5), which is similar to that of drug abuse (4.4%) and many anxiety disorders (1).

Unfortunately, bipolar spectrum disorders often go unrecognized and undiagnosed (6), largely because of the wide range of symptoms seen in patients with bipolar spectrum disorder, including impulsive behavior, alcohol and substance abuse, fluctuations in energy level, and legal problems. These symptoms are often attributed to problems other than bipolar disorder. The consequences of delayed diagnoses or misdiagnoses can be devastating.

One method of increasing recognition of an illness is to screen for it. Although several screening instruments exist for a variety of psychiatric disorders, none exist to screen for bipolar spectrum disorder. This article describes the development and validation of a brief and easy-to-use screening instrument for bipolar spectrum disorder called the Mood Disorder Questionnaire.

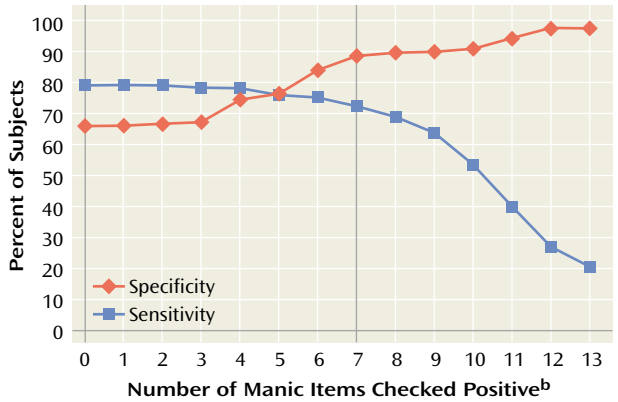
The Mood Disorder Questionnaire is a self-report, single-page, paper-and-pencil inventory that can be quickly and easily scored by a physician, nurse, or any trained medical staff assistant. The Mood Disorder Questionnaire screens for a lifetime history of a manic or hypomanic syndrome by including 13 yes/no items derived from both the DSM-IV criteria and clinical experience (Appendix 1). A yes/no question also asks whether several of any reported manic or hypomanic symptoms or behaviors were experienced during the same period of time. Finally, the level of functional impairment due to these symptoms ("no problem" to "serious problem") is queried on a 4-point scale. The original version of the Mood Disorder Questionnaire was administered to a convenience group of bipolar patients to assess feasibility and face validity. The items were then revised on the basis of this experience. The present study was designed to determine the optimal symptom threshold for identifying bipolar spectrum disorder and to assess the sensitivity and specificity of this threshold by using a professional mental health diagnosis of bipolar spectrum disorder as the criterion standard.

Method

The study was conducted at five outpatient psychiatric clinics that primarily treat patients with mood disorders, especially bipolar disorder. The protocol was approved by the institutional review board at each site. Signed informed consent was obtained from each subject. All subjects were English-speaking and at least 18 years old.

Outpatients being seen for treatment were asked to complete the Mood Disorder Questionnaire. Patients were contacted to receive a telephone research diagnostic interview within 2 weeks. An experienced psychiatric research social worker, who was blind to the Mood Disorder Questionnaire results, used the Structured

FIGURE 1. Operating Characteristics of the Mood Disorder Questionnaire for Various Threshold Scores Among 198 Patients From Outpatient Mood Disorder Clinics[a]



[a] A score of 7 or higher (gray vertical line) was chosen as the optimal cutoff.
[b] In addition to achieving the threshold number of symptom items, the subject *must* also have indicated that the symptoms clustered in the same time period ("yes" on question 2) and caused moderate or serious problems ("moderate" or "serious" on question 3).

Clinical Interview for DSM-IV (SCID) (7) to obtain a diagnosis of bipolar spectrum disorder (including bipolar I, bipolar II, and bipolar disorder not otherwise specified).

Data for the telephone-diagnosed subjects were analyzed with SPSS 8.0 for Windows (SPSS, Inc., Chicago). A scoring algorithm calculated the number of symptom items scored "yes" (range=0–13). In order to screen positively for bipolar spectrum disorder, in addition to a threshold number of symptom items, the respondent had to check "yes" for the item asking if the symptoms clustered in the same time period and had to indicate that the symptoms caused either "moderate" or "serious" problems. Sensitivity and specificity for each possible Mood Disorder Questionnaire score were plotted by using results from the SCID telephone interview as the standard. Sensitivity (percent of criterion standard diagnoses correctly diagnosed by the Mood Disorder Questionnaire) and specificity (percent of criterion standard noncases correctly identified as noncases by the Mood Disorder Questionnaire) for various symptom threshold cutoff scores were calculated in order to determine the optimal screen threshold.

Results

A group of 198 subjects received the telephone SCID interview. A total of 63% of the subjects were female. The mean age was 44 years (SD=13, range=18–80). A total of 86% had an education of high school level or higher. A total of 90% of the subjects were Caucasian, and 9% were African American.

A SCID diagnosis of bipolar spectrum disorder (bipolar I: N=70, bipolar II: N=26, and bipolar disorder not otherwise specified: N=13) was given to 109 (55%) of the 198 patients. The frequency of endorsement of Mood Disorder Questionnaire items ranged from 34.2% to 77.2% (the highest item endorsements were "easily distracted," "racing thoughts," and "irritability"). A Cronbach's alpha coefficient of 0.90 was achieved for the Mood Disorder Questionnaire. Individual item correlations with total score on the Mood Disorder Questionnaire ranged from 0.50 to 0.75.

Figure 1 presents the sensitivity and the specificity for various threshold cutoffs of the total score. A Mood Disorder Questionnaire screening score of 7 or more was chosen as the optimal cutoff, as it provided good sensitivity (0.73, 95% confidence interval [CI]=0.65–0.81) and very good specificity (0.90, 95% CI=0.84–0.96). Higher threshold cutoffs resulted in a loss of sensitivity without an appreciable increase in specificity; lower threshold cutoffs resulted in considerable loss of specificity. By using this 7-or-more-item threshold, seven out of 10 people with a bipolar spectrum disorder would be correctly identified by the Mood Disorder Questionnaire, whereas nine out of 10 of those who did not have a bipolar spectrum disorder would be successfully screened out.

Discussion

This study assessed the sensitivity and specificity of a brief, self-rated screening instrument for bipolar spectrum disorder by using a research diagnostic interview as the standard for diagnosis in a psychiatric outpatient population. The operating characteristics of the Mood Disorder Questionnaire are quite good and are comparable to those of other instruments that are used to screen for other psychiatric disorders. Mulrow et al. (8) reviewed 18 studies using nine different screening instruments for depression in primary care settings. The sensitivities and specificities of the instruments ranged from 0.67 to 0.99 (mean=0.84) and from 0.40 to 0.95 (mean=0.72), respectively. The Mood Disorder Questionnaire's sensitivity of 0.73 and specificity of 0.90 compare well with the accuracy of these other instruments.

Further research is needed to assess whether the Mood Disorder Questionnaire would be useful in primary care, community agencies, and other psychiatric settings to identify individuals who might benefit from a comprehensive diagnostic evaluation.

Received March 1, 2000; revision received June 1, 2000; accepted June 7, 2000. From the Department of Psychiatry and Behavioral Sciences, Medical Branch, the University of Texas at Galveston; the Department of Psychiatry, Columbia University, and the New York State Psychiatric Institute, New York; the School of Medicine, Case Western Reserve University, Cleveland; the National Association for the Mentally Ill, Washington, D.C.; the Biological Psychiatry Program, Department of Psychiatry, College of Medicine, University of Cincinnati, Cincinnati; the National Depressive and Manic-Depressive Association, Chicago; the NIMH, Bethesda, Md.; Massachusetts General Hospital, Harvard Medical School, Boston; and Rush-Presbyterian–St. Luke's Medical Center, Chicago.

Address reprint requests to Dr. Hirschfeld, Department of Psychiatry and Behavioral Sciences, Medical Branch, University of Texas at Galveston, 1.302 Rebecca Sealy, 301 University Blvd., Galveston, TX 77555-0188; rohirsch@utmb.edu (e-mail).

Supported by an unrestricted educational grant from Abbott Laboratories.

The authors thank David M. Medearis, B.S., for assistance in the preparation of this article.

APPENDIX 1. The Mood Disorder Questionnaire

1.	Has there ever been a period of time when you were not your usual self and...	YES	NO
	...you felt so good or so hyper that other people thought you were not your normal self or you were so hyper that you got into trouble?	☐	☐
	...you were so irritable that you shouted at people or started fights or arguments?	☐	☐
	...you felt much more self-confident than usual?	☐	☐
	...you got much less sleep than usual and found you didn't really miss it?	☐	☐
	...you were much more talkative or spoke faster than usual?	☐	☐
	...thoughts raced through your head or you couldn't slow your mind down?	☐	☐
	...you were so easily distracted by things around you that you had trouble concentrating or staying on track?	☐	☐
	...you had much more energy than usual?	☐	☐
	...you were much more active or did many more things than usual?	☐	☐
	...you were much more social or outgoing than usual, for example, you telephoned friends in the middle of the night?	☐	☐
	...you were much more interested in sex than usual?	☐	☐
	...you did things that were unusual for you or that other people might have thought were excessive, foolish, or risky?	☐	☐
	...spending money got you or your family into trouble?	☐	☐
2.	If you checked YES to more than one of the above, have several of these ever happened during the same period of time? *Please circle one response only.* **YES** **NO**		
3.	How much of a problem did any of these cause you — like being unable to work; having family, money, or legal troubles; getting into arguments or fights? *Please circle one response only.* **No problem** **Minor problem** **Moderate problem** **Serious problem**		

References

1. Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, Wittchen H-U, Kendler KS: Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. Arch Gen Psychiatry 1994; 51:8–19
2. Weissman MM, Bruce LM, Leaf PJ: Affective disorders, in Psychiatric Disorders in America: The Epidemiologic Catchment Area Study. Edited by Robins LN, Regier DA. New York, Free Press, 1991, pp 53–80
3. Klerman GL: The spectrum of mania. Compr Psychiatry 1981; 22:11–20
4. Akiskal HS, Pinto O: The evolving bipolar spectrum: prototypes I, II, III, and IV. Psychiatr Clin North Am 1999; 22:517–534
5. Angst J: The emerging epidemiology of hypomania and bipolar II disorder. J Affect Disord 1998; 50:143–151
6. Lish JD, Dime-Meenan S, Whybrow PC, Price RA, Hirschfeld RM: The National Depressive and Manic-Depressive Association (National DMDA) survey of bipolar members. J Affect Disord 1994; 31:281–294
7. Spitzer RL, Williams JBW, Gibbon M, First MB: The Structured Clinical Interview for DSM-III-R (SCID), I: history, rationale, and description. Arch Gen Psychiatry 1992; 49:624–629
8. Mulrow CD, Williams JW Jr, Gerety MB, Ramirez G, Montiel OM, Kerber C: Case-finding instruments for depression in primary care settings. Ann Intern Med 1995; 122:913–921

Prevalence and Effects of Mood Disorders on Work Performance in a Nationally Representative Sample of U.S. Workers

Ronald C. Kessler, Ph.D.

Hagop S. Akiskal, M.D.

Minnie Ames, Ph.D.

Howard Birnbaum, Ph.D.

Paul Greenberg, M.A.

Robert M.A. Hirschfeld, M.D.

Robert Jin, M.S.

Kathleen R. Merikangas, Ph.D.

Gregory E. Simon, M.D.

Philip S. Wang, M.D., Dr.P.H.

Objective: Research on the workplace costs of mood disorders has focused largely on major depressive episodes. Bipolar disorder has been overlooked both because of the failure to distinguish between major depressive disorder and bipolar disorder and by the failure to evaluate the workplace costs of mania/hypomania.

Method: The National Comorbidity Survey Replication assessed major depressive disorder and bipolar disorder with the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI) and work impairment with the WHO Health and Work Performance Questionnaire. A regression analysis of major depressive disorder and bipolar disorder predicting Health and Work Performance Questionnaire scores among 3,378 workers was used to estimate the workplace costs of mood disorders.

Results: A total of 1.1% of the workers met CIDI criteria for 12-month bipolar disorder (I or II), and 6.4% meet criteria for 12-month major depressive disorder. Bipolar disorder was associated with 65.5 and major depressive disorder with 27.2 lost workdays per ill worker per year. Subgroup analysis showed that the higher work loss associated with bipolar disorder than with major depressive disorder was due to more severe and persistent depressive episodes in those with bipolar disorder than in those with major depressive disorder rather than to stronger effects of mania/hypomania than depression.

Conclusions: Employer interest in workplace costs of mood disorders should be broadened beyond major depressive disorder to include bipolar disorder. Effectiveness trials are needed to study the return on employer investment of coordinated programs for workplace screening and treatment of bipolar disorder and major depressive disorder.

(Am J Psychiatry 2006; 163:1561–1568)

Although bipolar disorder has traditionally been thought to have a lifetime prevalence of about 1% of the population (1, 2), a substantial upward revision of this estimate is occurring based on mounting evidence for a broad bipolar spectrum that includes people with a history of hypomania, subthreshold manic symptoms, and medication-induced manic symptoms (3–5). The bipolar spectrum is now thought to characterize as much as 5% of the general population (6). People with bipolar spectrum disorder spend a considerably higher proportion of time with depressive than manic symptoms (7, 8), resulting in frequent confusion between major depressive episodes due to major depressive disorder or to bipolar disorder (9). Failure to make this distinction can have dire clinical implications.

The same distinction between major depressive disorder and bipolar disorder would be useful to make in depression cost-of-illness studies. However, with rare exceptions, these studies failed to distinguish between major depressive episodes associated with major depressive disorder and those associated with bipolar disorder (10–13). Furthermore, although several recent cost-of-illness studies (14–16) and reviews (17, 18) have focused on the costs of bipolar disorder, none has presented comparative information on the workplace costs of major depressive disorder and bipolar disorder. The current report does this using data from the recently completed National Comorbidity Survey Replication (NCS-R) (19), a nationally representative survey of the prevalence and costs of mental disorders in the U.S. household population.

Method

Sample

The NCS-R is a nationally representative survey of mental disorders among English-speaking household residents ages 18 and older in the continental United States. Interviews were carried out with 9,282 respondents between February 2001 and April 2003. Verbal informed consent was obtained before data collection. Consent was verbal rather than written to maintain consistency with the baseline NCS. The rate of response was 70.9%. Respondents were given a $50 incentive for participation. In addition, a probability subsample of hard-to-recruit predesignated respon-

This article is featured in this month's AJP Audio and is the subject of an editorial by Dr. Goldman on p. 1490.

TABLE 1. Demographic Distributions and Correlates of 12-Month DSM-IV Bipolar Disorder and Major Depressive Disorder Among 3,378 Employed Respondents to Part 2 of the National Comorbidity Survey Replication

	Demographic Distribution		12-Month Prevalence				Analysis				
			Bipolar Disorder		Major Depressive Disorder		Bipolar Disorder				
Variable	%	SE	%	SE	%	SE	Odds Ratio	95% CI	χ^2	df	p
Sex											
Men	53.4	1.3	0.8	0.2	4.0	0.4	1.0	—			
Women	46.6	1.3	1.5	0.3	9.0	0.9	1.4	0.6–3.2			
χ^2									0.8	1	0.37
Age											
18–29	25.3	1.4	1.6	0.3	7.3	1.0	4.3	0.6–34.1			
30–44	36.6	1.3	1.3	0.3	7.3	0.6	3.8	0.6–22.7			
45–59	32.1	1.2	0.7	0.3	5.3	0.7	2.2	0.2–20.0			
60+	6.0	0.5	0.5	0.5	2.1	0.4	1.0	—			
χ^2									6.3	3	0.10
Race-ethnicity											
Non-Hispanic white	73.5	1.6	1.2	0.2	6.6	0.5	1.0	—			
Non-Hispanic black	11.5	1.0	0.8	0.3	4.8	1.0	0.5	0.2–1.2			
Hispanic	11.3	1.1	1.1	0.5	6.1	1.0	0.6	0.2–1.6			
Other	3.7	0.4	1.5	0.9	7.9	2.6	1.3	0.3–4.9			
χ^2									5.3	3	<0.16
Education											
Less than high school	10.4	0.8	1.6	0.6	6.2	1.4	7.0*	1.8–27.9			
Completed high school	31.0	1.5	1.6	0.4	6.1	0.6	5.6*	2.1–15.1			
Some college	30.4	1.0	1.2	0.3	6.5	0.7	3.8*	1.3–11.2			
Completed college	28.2	1.2	0.4	0.2	6.6	0.9	1.0	—			
χ^2									14.0	3	0.003*
Occupation											
Professional	34.2	0.9	0.9	0.3	7.0	0.8	1.0	—			
Technical	3.0	0.5	0.3	0.4	5.0	2.1	0.2	0.0–2.0			
Service and clerical	21.1	0.8	1.8	0.4	9.3	1.0	0.8	0.3–2.0			
Labor	41.7	1.3	1.0	0.3	4.5	0.5	0.6	0.3–1.3			
χ^2									3.4	3	0.33
Average work hours											
20–34	13.0	0.9	1.8	0.5	7.6	1.2	1.3	0.6–2.7			
35–44	55.5	1.5	1.1	0.2	7.0	0.6	1.0	0.5–2.1			
45+	31.5	1.4	0.9	0.3	4.7	0.7	1.0	—			
χ^2									0.7	2	0.71

*p=0.05, two-tailed.

TABLE 2. Relation of 12-Month DSM-IV Bipolar Disorder and Major Depressive Disorder With Annualized Work Loss Days Due to Absenteeism and Presenteeism Among 3,378 Employed Respondents to Part 2 of the National Comorbidity Survey Replication

	Individual Level				Aggregate Level (total U.S. labor force)[a]			
	Days per Year		U.S. Dollars per Year		Million Days per Year		Million U.S. Dollars per Year	
Disorder	Days	SE	Dollars	SE	Days	SE	Dollars	SE
Bipolar disorder								
Absenteeism	27.7*	7.0	4,067*	1,034	40.7*	10.3	5,973*	1,518
Presenteeism[b]	35.3*	7.7	5,184*	1,137	51.8*	11.4	7,613*	1,670
Total[c]	65.5*	10.4	9,619*	1,527	96.2*	15.3	14,128*	2,242
Major depressive disorder								
Absenteeism	8.7*	2.6	1,420*	418	72.2*	21.2	11,742*	3,456
Presenteeism	18.2*	3.6	2,961*	591	150.5*	30.1	24,482*	4,890
Total[c]	27.2*	4.8	4,426*	784	225.0*	39.9	36,602*	6,485

[a] These results are based on a projection to the total civilian U.S. labor force based on data from the 2002 Current Population Survey.
[b] Presenteeism is defined in lost day equivalents.
[c] Entries do not sum to the parallel entries for absenteeism and presenteeism because the totals were based on a separate regression equation in which the dependent variable was a measure of total lost days of work rather than the simple summation of the results in the earlier rows.
*p=0.05 level, two-tailed.

dents was selected for a brief telephone nonrespondent survey, the results of which were used to weight the main sample for nonresponse bias. Nonrespondent survey participants were given a $100 incentive. The Human Subjects Committees of Harvard Medical School and the University of Michigan both approved these recruitment and consent procedures.

The NCS-R interview was administered in two parts. Part 1 included a core diagnostic assessment of all 9,282 respondents. Part 2 included questions about correlates and additional disorders administered to all part 1 respondents who met lifetime criteria for any core disorder plus a roughly 1-in-3 probability subsample of 5,692 other respondents. The Health and Work Performance

Analysis				
Major Depressive Disorder				
Odds Ratio	95% CI	χ^2	df	p
1.0	—			
2.0	1.4–2.8			
		18.6	1	<0.001*
4.0*	2.5–6.3			
4.1*	2.6–6.6			
2.9*	1.7–4.8			
1.0	—			
		44.9	3	<0.001*
1.0	—			
0.6	0.4–1.0			
0.8	0.6–1.2			
0.8	0.6–2.1			
1.1	0.6–2.1			
		5.3	3	<0.16
1.3	0.6–2.7			
1.1	0.8–1.6			
1.1	0.7–1.8			
1.0	—			
		0.7	3	0.87
1.0	—			
0.8	0.3–2.1			
1.0	0.7–1.5			
0.7	0.4–1.3			
		2.8	2	0.43
1.3	0.8–2.1			
1.3	1.0–1.8			
1.0	—			
		3.3	2	<0.19

Questionnaire assessment of work performance was included in part 2. A subsample of 3,378 part 2 respondents was either employed or self-employed 20 hours or more per week in the month before the interview and had valid data on all measures used in the following analyses. This is the sample used here. The records for these respondents were weighted to adjust for differential probability of selection into part 2 of the interview and for differential nonresponse. A more detailed discussion of NCS-R sampling and weighting is presented elsewhere (20).

Mood Disorders

NCS-R diagnoses were based on version 3.0 of the World Health Organization's Composite International Diagnostic Interview (CIDI) (21), a fully structured lay-administered measure. DSM-IV criteria were used to define major depressive episodes, dysthymic disorder, bipolar I disorder, and bipolar II disorder. Because of the small sample size, bipolar I disorder and bipolar II disorder were combined into a single category of bipolar disorder for the current analysis. All diagnoses excluded patients with plausible organic causes for their illness. Blind clinical reappraisal interviews with the lifetime nonpatient version of the Structured Clinical Interview for DSM-IV (SCID) (22) were administered to a probability subsample of 325 NCS-R respondents to assess concordance with CIDI hierarchy-free diagnoses. CIDI-SCID concordance was excellent for bipolar disorder, with an area under the receiver-operator-characteristic curve of 0.93, an odds ratio of 582.6, and a nonsignificant McNemar test (χ^2=0.6, df=1, p=0.45). The McNemar test evaluated whether the CIDI prevalence estimate differed significantly from the SCID prevalence estimate. Concordance between the CIDI and the SCID was also good for major depression, with an area under the curve of 0.75, an odds ratio of 18.4, and a McNemar test (χ^2=7.2, df=1, p=0.006). The McNemar test was significant because the CIDI prevalence estimate was conservative relative to the SCID estimate. Concordance between the CIDI and the SCID was not assessed for dysthymia because the number of respondents with dysthymia in the clinical reappraisal sample was too small for reliable analysis.

Once the mood disorders were operationalized, the respondents who were classified as having lifetime bipolar disorder were defined as 12-month patients with 12 months of illness if they experienced a major depressive, a manic, or a hypomanic episode at any time in the 12 months before the interview. The respondents classified as having lifetime major depressive disorder were defined as having 12-month cases if they had experienced a major depressive episode at any time in the 12 months before the interview. The vast majority of respondents with a hierarchy-free diagnosis of 12-month dysthymia also met criteria for 12-month major depressive disorder. These "double depressives" (23) were subsequently compared with other patients with major depressive disorder in the ability to predict their work performance. The handful of respondents with 12-month dysthymia who failed to meet criteria for major depressive disorder was excluded from the analysis because of the group's low statistical power.

The persistence and severity of 12-month major depressive episodes were compared for respondents with bipolar disorder who had 12-month major depressive episodes and for respondents with 12-month major depressive disorder to determine whether more severe or persistent depression could account for observed differences in work performance between the two subsamples. Persistence was assessed by asking respondents with a 12-month major depressive episode to estimate how many days out of 365 in the past year they had experienced a depressive episode. Severity was assessed with the self-report version of the Quick Inventory of Depressive Symptomatology (24), referring to the 1 month in the past year when the respondents reported their depression as most severe.

Work Performance

Work performance was assessed with the WHO Health and Work Performance Questionnaire (25, 26). This measure uses self-reports about absenteeism (missed days of work) and "presenteeism" (low performance while at work transformed to lost workday equivalents) to generate a summary measure of overall lost workdays in the month before the interview. Absenteeism was defined on a 0–100 scale for the percentage of work days the respondent missed in the past 30 days, while presenteeism was defined on a separate 0–100 scale in which 0 meant doing no work at all on days spent at work and 100 meant performing at the level of a top worker. Absenteeism and presenteeism were combined into a measure of total lost work performance by adding absenteeism to the value ([100–absenteeism] × [100–presenteeism]). Information about salary was used to transform the measures of lost work performance from a time metric to a salary metric for the purposes of estimating human capital loss associated with mood disorders. Salary was incremented by 25% to estimate fringe benefits.

Control Variables

All analyses included control for sex, age (18–29, 30–44, 45–59, and 60 and over), race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic, other), education (less than high school, completed high school, some college, completed college), and occupation (professional, technical, service-clerical, laborer), as well as for average expected hours of work per week (20–34, 35–44, 45 or more).

TABLE 3. Persistence and Severity of 12-Month Major Depressive Episodes With 12-Month DSM-IV Bipolar Disorder or Major Depressive Disorder Among Employed Respondents to Part 2 of the National Comorbidity Survey Replication

	Bipolar Disorder							
	Major Depressive Episode Only (N=7)				Major Depressive Episode and Mania/Hypomania (N=37)			
Variable	Mean	SE	Median	Range	Mean	SE	Median	Range
Persistence (number of days in major depressive episodes in the past 365 days)	134.0	53.5	90.0	30.0–183.0	164.0	19.8	150.0	52.0–250.00
Severity (scores on the Quick Inventory of Depressive Symptomatology Self-Report)	14.1	1.5	15.7	11.5–17.5	17.3	0.8	16.5	15.7–19.6

[a] All respondents with a 12-month major depressive episode who had either bipolar disorder or major depressive disorder were compared.

Analysis

Subgroup comparison of prevalence estimates was used to study the sociodemographic correlates of mood disorders, while linear regression analysis was used to estimate associations of mood disorders with work performance. Mood disorders were coded as "yes/no" dummy predictor variables in linear regression equations that included sociodemographic variables (age, sex, race/ethnicity, education, occupation) as controls. The dependent variables in these equations were measures of lost work performance in the metrics of day equivalents and salary equivalents that distinguished absenteeism and presenteeism as well as combined absenteeism and presenteeism into a summary measure of overall lost work performance.

These basic equations were elaborated in three ways. The first distinguished among patients with bipolar disorder who reported 12-month episodes of major depressive episodes only, mania or hypomania only, or both. The second controlled for the severity duration of major depressive episodes. The third evaluated interactions between mood disorders and sociodemographic variables.

The key predictors in the regression equations were measures of the prevalence of 12-month disorders, whereas the outcomes were measures of 1-month (not 12-month) decrements in work performance. The coefficients were multiplied by 12 to estimate decrements in work performance over the past 12 months because of 12-month mood disorders. These individual-level estimates were then projected to the total U.S. civilian labor force by adjusting for 12-month disorder prevalence and for the fact that the seasonally adjusted number of workers in the U.S. civilian labor force ages 18 and over at the time of the NCS-R was 130 million.

A question might be raised as to why the time frame of the measures was not made consistent by using either 12-month decrements in work performance as outcomes or 1-month prevalence of mood disorders as predictors. The former was not possible because methodological research has shown that retrospective self-reports about health-related decrements in work performance are inaccurate beyond a 1-month recall period (27). The latter (i.e., using measures of 1-month mood disorders as predictors) would have been possible but would have left unresolved the possibility that remitted mood disorders continue to have residual adverse effects on work performance after episode resolution. The use of 12-month disorders to predict 1-month work performance resolves this problem by generating an averaged estimate of the effects on 1-month work performance of both active episodes and remitted episodes that were active in the past 12 months. The multiplication of this estimate by 12 then produces an unbiased estimate of the effect of mood disorders active in the past 12 months on decrements in work performance in the same time period.

Because the NCS-R data are weighted and clustered, the Taylor series linearization method (28) implemented in the SUDAAN software system (29) was used to obtain design-based estimates of statistical significance. Significance tests of sets of coefficients in the logistic regression equations were made using Wald χ^2 tests based on design-corrected coefficient variance-covariance matrices. Statistical significance was consistently evaluated as p=0.05, two-tailed.

Results

Prevalence and Sociodemographic Correlates

Twelve-month prevalence estimates of DSM-IV bipolar disorder and major depressive disorder (standard errors in parentheses) among employed NCS-R respondents were 1.1% (SE=0.2) and 6.4% (SE=0.5), respectively. The estimated prevalence of bipolar disorder did not differ significantly by the respondents' sex, age, race/ethnicity, occupation, or expected work hours but was inversely related to education (Table 1). The estimated prevalence of major depressive disorder did not differ significantly by respondent race/ethnicity, education, occupation, or expected work hours but was significantly higher among women than men and inversely related to age. Neither bipolar disorder nor major depressive disorder was related to average hours worked per week.

Associations of Mood Disorders With Work Performance

Bipolar disorder and major depressive disorder both significantly predicted overall lost work performance in the regression analysis, with annualized regression slopes equivalent to 65.5 lost workdays per worker with bipolar disorder and 27.2 lost work days per worker with major depressive disorder (Table 2). Disaggregation showed that absenteeism, while significantly elevated for both people with bipolar disorder (27.7 days) and people with major depressive disorder (8.7 days), was less important than presenteeism (35.3 days for those with bipolar disorder and 18.2 days for those with major depressive disorder). Projections of individual-level associations to the total U.S. civilian labor force yielded estimates of 96.2 million lost workdays and $14.1 billion salary-equivalent lost productivity per year associated with bipolar disorder and 225.0 million workdays and $36.6 billion salary-equivalent lost productivity per year associated with major depressive disorder.

Bipolar Disorder					
Major Depressive Disorder (N=342)				Analysis	
Mean	SE	Median	Range	z[a]	p
98.1	5.1	60.0	28.0–150.00	2.7	0.01
14.5	0.3	14.7	11.5–17.5	2.9	0.007

Variation in Associations Based on the Persistence and Severity of Depressive Episodes

Roughly three-fourths of the respondents with 12-month bipolar disorder had depressive episodes in the 12 months before the interview (63.1% who also had manic/hypomanic episodes and 11.1% who had only depressive episodes). Persistence (days in depressive episodes in the 365 days before the interview) was consistently higher in individuals with bipolar disorder (mean=134.0–164.0, median=90–150) than in those with major depressive disorder (mean=98.1, median=60.0; z=2.7, p=0.01) (Table 3). Severity (scores on the Quick Inventory of Depressive Symptoms) was also consistently higher in individuals with bipolar disorder (mean=14.1–17.3, median=15.7–16.5) than in those with major depressive disorder (mean=14.5, median=14.7; z=2.9, p=0.007).

The individual-level elevations of absenteeism, presenteeism, and total lost work performance in individuals with bipolar disorder were consistently higher among respondents with 12-month major depressive episodes than only manic/hypomanic episodes (Table 4). Furthermore, bipolar disorder with major depressive episodes was consistently associated with significantly more lost work performance than major depressive disorder. Statistical control for major depressive episode persistence and severity reduces these discrepancies somewhat but does not make them disappear. Bipolar disorder with only manic/hypomanic episodes, in comparison, is associated with levels of lost work performance roughly equal to those with major depressive disorder.

Variation in Associations Based on Sociodemographic Variables

No significant differences in the associations of bipolar disorder and major depressive disorder with work performance were found by sex (χ^2=0.1–1.0, df=1, p=0.31–0.76) or, in the case of bipolar disorder, by age (χ^2=0.5–0.8, df=3, p=0.66–0.79), but the major depressive disorder coefficients varied with age (χ^2=8.0–29.0, df=3, p=0.001–<0.02) because of larger coefficients among workers in the age range of 30 to 44 years than either younger or older workers. (Detailed results are available upon request from the first author.) We also found variation in associations by occupation among those with bipolar disorder (χ^2=36.8–212.9, df=3, p<0.0001) but not major depressive disorder (χ^2=1.5–5.9, df=3, p=0.12–0.67). The work loss associated with bipolar disorder, although consistently significant in each occupational group, was significantly greater among technical and professional workers in the case of absenteeism and among laborers and professional workers in the case of presenteeism. (Detailed results are available upon request from the first author.)

Discussion

Two potential limitations of this study are the possible existence of inaccuracy in the key measures and the possible existence of unmeasured common causes of the disorders and outcomes. With regard to the first of these two, the accuracy of diagnostic assessment was documented in the SCID reappraisal interviews mentioned in the section on measures. However, fully structured instruments, such as the CIDI, are less able to distinguish mixed episodes than are semistructured clinical interviews, leading to the imposition of a more rigid distinction between major depressive episodes and manic/hypomanic episodes in individuals with major depressive disorder than would have been ideal (30). The accuracy of the Health and Work Performance Questionnaire work performance assessment was evaluated in a series of workplace validity studies (25, 26) that documented strong relationships of questionnaire measures with independent payroll records and supervisor evaluations of job performance.

The possibility of unmeasured common causes is much more difficult to evaluate. To the extent that common causes exist, the estimated effects of bipolar disorder and major depressive disorder on lost work performance will be biased. No definitive way exists to evaluate this possibility other than by experimentally changing the prevalence of these disorders, presumably in a treatment effectiveness trial, and evaluating the effects on work performance. The results of such experiments in representative workplace samples have not been reported either for bipolar disorder or major depressive disorder, although such an experiment is currently underway to evaluate the workplace effects of treating major depressive disorder (31). Despite the absence of experimental evidence, simulations of likely ef-

TABLE 4. Individual-Level Associations of 12-Month DSM-IV Bipolar Disorder Disaggregated by Type of 12-Month Episode and Major Depressive Disorder With Annualized Work Loss Days Due to Absenteeism and Presenteeism With and Without Control for Persistence and Severity of Major Depressive Episodes Among 3,378 Employed Respondents to Part 2 of the National Comorbidity Survey Replication

	Bipolar Disorder							
	Mania/ Hypomania Only		Major Depressive Episodes Only		Both		Major Depressive Disorder	
Variable	Days	SE	Days	SE	Days	SE	Days	SE
Without control for persistence and severity of major depressive episodes								
Absenteeism	12.5	8.6	32.2*	13.9	33.1*	11.9	8.7*	2.6
Presenteeism[a]	27.8*	12.4	62.0*	29.9	33.3*	11.4	18.2*	3.6
Total[b]	39.6*	18.0	105.4*	29.7	69.0*	16.6	27.2*	4.8
With control for persistence and severity of major depressive episodes								
Absenteeism	12.6	8.6	25.6*	13.1	25.4	19.6	2.9	5.5
Presenteeism[a]	28.2*	12.4	42.6	30.2	3.4	15.8	–2.0	9.6
Total[b]	40.1*	17.9	79.4*	24.7	32.5	31.3	3.8	15.3

[a] Presenteeism is defined in lost day equivalents.
[b] Entries do not sum to the parallel entries for absenteeism and presenteeism because the totals were based on a separate regression equation in which the dependent variable was a measure of total lost days of work rather than the simple summation of the results in the earlier rows.
*p=0.05, two-tailed.

fects have been carried out using parameter estimates gleaned from clinical trials (32, 33). The estimated decrements in work performance associated with major depressive disorder in these simulations are broadly consistent with the NCS-R estimates. In addition, the results of a recently reported experimental effectiveness trial aimed at increasing work performance by improving the quality of major depressive disorder treatment yielded estimates of effects on work performance broadly consistent with the NCS-R estimates (34).

Within the context of these limitations, the results reported here show that bipolar disorder and major depressive disorder are both common disorders in the U.S. civilian labor force associated with substantial lost work performance. Our prevalence estimates of bipolar disorder and major depressive disorder are consistent with those in other national surveys (35, 36). As noted in the introduction though, bipolar spectrum disorders could be defined more broadly than in the current report (3–5). The same is true for subthreshold depression (37, 38). Future research should investigate the effects of these subthreshold disorders on work performance (5, 39).

Our finding that both bipolar disorder and major depressive disorder are associated with substantial losses in work performance is consistent with other estimates of workplace costs (12, 17, 18, 40, 41). The estimated annual population-level workplace cost of major depressive disorder, $36.6 billion, is similar in magnitude to the $31.0 billion estimate reported in another recent study (31). In addition, the workplace cost of major depressive disorder plus bipolar disorder, $50.7 billion, is very similar to the $51.5 billion estimate reported elsewhere (12), although the distribution of workplace cost components is quite different across studies. Whereas presenteeism is estimated here to account for about two-thirds of the total workplace costs of illness, the earlier findings were skewed in the opposite direction, with more than two-thirds of workplace costs estimated to arise from absenteeism (12). The current results are likely to be more accurate than the earlier ones because the Health and Work Performance Questionnaire produces a better measure of the on-the-job component of work performance than the measure used in the previous study. Finally, the only previous estimate of the population-level workplace cost of bipolar disorder, $2.3 billion in 1990 dollars or in the range of $3 billion to $4 billion today (11), is much lower than our $14.1 billion estimate, presumably reflecting the fact that this earlier report, which was based on synthetic estimation rather than primary data collection, assumed a much lower prevalence than we found to be the case in our nationally representative survey.

By considering bipolar disorder and major depressive disorder simultaneously, we documented that bipolar disorder is associated with substantially more lost work performance than major depressive disorder at the individual level, although aggregate impairment is greater for major depressive disorder than for bipolar disorder because of the higher prevalence of the former than the latter disorder. Decomposition showed that the higher individual-level impairment of bipolar disorder than major depressive disorder was due largely to major depressive episodes being more impairing in the context of bipolar disorder than in major depressive disorder rather than to mania/hypomania being more impairing than major depressive episodes. The finding that mania/hypomania in the absence of major depressive episodes is associated with significantly less work impairment than bipolar disorder with major depressive episodes is consistent with the observation in a prospective patient study that functional impairment was associated with variation in depressive symptoms but not manic symptoms (42). More detailed analysis of the NCS-R data showed that the higher individual-level work impairment of major depressive episodes in bipolar disorder than in major depressive disorder is

due partly to the greater persistence and severity of major depressive episodes in bipolar disorder than in major depressive disorder. However, the persistence/severity of major depressive episodes explained only part of the association between bipolar disorder and work impairment. The remaining part of this association could be due to either imprecision in our measures or the effects of unmeasured correlates of bipolar disorder and work impairment.

An important practical problem related to the finding that most workers with bipolar disorder had major depressive episodes is that major depressive episodes due to bipolar disorder are sometimes incorrectly treated as if they were due to major depressive disorder (43, 44). This problem is exacerbated by people with bipolar disorder reporting more distress because of their depressive than their manic symptoms (40). Because antidepressant medications can trigger the onset of mania, it is important to screen for a history of bipolar disorder at the initiation of depression treatment. A short and valid screen for manic/hypomanic symptoms has recently been developed that could be used for this purpose (45). It is important for the same reason to include a screen for bipolar disorder in workplace depression screening programs. The prevalence and impairments of subthreshold cases should also be examined. Effectiveness trials are needed to calculate the return on investment from the employer's perspective of coordinated workplace bipolar disorder-major depressive disorder screening and treatment (34, 41).

Received Feb. 4, 2005; revisions received Sept. 19 and Oct. 4, 2005; accepted Nov. 15, 2005. From the Department of Health Care Policy, Harvard Medical School, Boston; the International Mood Center, University of California, San Diego; the VA Psychiatry Service, San Diego; the Analysis Group Inc., Boston; the Department of Psychiatry and Behavioral Sciences, University of Texas Medical Branch, Galveston, Tex.; the Group Health Cooperative, Intramural Research Branch, NIMH, Bethesda, Md.; and the Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston. Address correspondence and reprint requests to Dr. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave., Boston, MA 02115; kessler@hcp.med.harvard.edu (e-mail).

The National Comorbidity Survey Replication (NCS-R) is supported by NIMH (grant U01-MH60220), with supplemental support from the National Institute of Drug Abuse, the Substance Abuse and Mental Health Services Administration, the Robert Wood Johnson Foundation (grant 044780), and the John W. Alden Trust. Additional support for preparation of this article was provided by BristolMyersSquibb. Instrumentation, fieldwork, and consultation on data analysis were supported by the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the U.S. Public Health Service (grants 1R13-MH-066849, R01-MH069864, and R01-DA016558), Eli Lilly and Company, GlaxoSmithKline, Ortho-McNeil Pharmaceutical, Inc., and the Pan American Health Organization.

The authors thank the staff of the World Mental Health Data Collection and Data Analysis Coordination Centers for assistance with instrumentation, fieldwork, and consultation on data analysis.

The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or the U.S. government.

Acknowledgments

Collaborating NCS-R investigators include Ronald C. Kessler (principal investigator, Harvard Medical School), Kathleen Merikangas (co-principal investigator, NIMH), James Anthony (Michigan State University), William Eaton (Johns Hopkins University), Meyer Glantz (NIDA), Doreen Koretz (Harvard University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Greg Simon (Group Health Cooperative), T. Bedirhan Ustun (WHO), Michael Von Korff (Group Health Cooperative), Philip Wang (Harvard Medical School), Kenneth Wells (UCLA), Elaine Wethington (Cornell University), and Hans-Ulrich Wittchen (Max Planck Institute of Psychiatry). A complete list of NCS publications and the full text of all NCS-R instruments can be found at http://www.hcp.med.harvard.edu/ncs. Send correspondence to NCS@hcp.med.harvard.edu. The NCS-R is carried out in conjunction with the WHO World Mental Health (WMH) Survey Initiative. A complete list of WMH publications and instruments can be found at http://www.hcp.med.harvard.edu/wmhcidi.

References

1. Kessler RC, Rubinow DR, Holmes C, Abelson JM, Zhao S: The epidemiology of DSM-III-R bipolar I disorder in a general population survey. Psychol Med 1997; 27:1079–1089
2. Robins LN, Helzer JE, Weissman MM, Orvaschel H, Gruenberg E, Burke JD Jr, Regier DA: Lifetime prevalence of specific psychiatric disorders in three sites. Arch Gen Psychiatry 1984; 41:949–958
3. Angst J, Gamma A, Benazzi F, Ajdacic V, Eich D, Rossler W: Toward a re-definition of subthreshold bipolarity: epidemiology and proposed criteria for bipolar-II, minor bipolar disorders and hypomania. J Affect Disord 2003; 73:133–146
4. Benazzi F, Koukopoulos A, Akiskal HS: Toward a validation of a new definition of agitated depression as a bipolar mixed state (mixed depression). Eur Psychiatry 2004; 19:85–90
5. Judd LL, Akiskal HS: The prevalence and disability of bipolar spectrum disorders in the US population: re-analysis of the ECA database taking into account subthreshold cases. J Affect Disord 2003; 73:123–131
6. Akiskal HS, Bourgeois ML, Angst J, Post R, Moller H, Hirschfeld R: Re-evaluating the prevalence of and diagnostic composition within the broad clinical spectrum of bipolar disorders. J Affect Disord 2000; 59(suppl 1):S5–S30
7. Judd LL, Akiskal HS, Schettler PJ, Endicott J, Maser J, Solomon DA, Leon AC, Rice JA, Keller MB: The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry 2002; 59:530–537
8. Calabrese JR, Hirschfeld RM, Frye MA, Reed ML: Impact of depressive symptoms compared with manic symptoms in bipolar disorder: results of a US community-based sample. J Clin Psychiatry 2004; 65:1499–1504
9. Hirschfeld RM, Vornik LA: Recognition and diagnosis of bipolar disorder. J Clin Psychiatry 2004; 65(suppl 15):5–9
10. Greenberg PE, Stiglin LE, Finkelstein SN, Berndt ER: The economic burden of depression in 1990. J Clin Psychiatry 1993; 54:405–418
11. Greenberg PE, Kessler RC, Nells TL, Finkelstein SN, Berndt ER: Depression in the workplace: an economic perspective, in Perspectives in Psychiatry: Selective Serotonin Re-uptake Inhibitors, 2nd ed, vol 5. Edited by Feighner JP, Boyer WF. New York, John Wiley & Sons, 1996, pp 327–363
12. Greenberg PE, Kessler RC, Birnbaum HG, Leong SA, Lowe SW, Berglund PA, Corey-Lisle PK: The economic burden of depression in the United States: how did it change between 1990 and 2000? J Clin Psychiatry 2003; 64:1465–1475

13. Rice DP, Miller LS: The economic burden of affective disorders. Adv Health Econ Health Serv Res 1993; 14:37–53
14. Wyatt RJ, Henter I: An economic evaluation of manic-depressive illness—1991. Soc Psychiatry Psychiatr Epidemiol 1995; 30:213–219
15. Begley CE, Annegers JF, Swann AC, Lewis C, Coan S, Schnapp WB, Bryant-Comstock L: The lifetime cost of bipolar disorder in the US: an estimate for new cases in 1998. Pharmacoeconomics 2001; 19:483–495
16. Das Gupta R, Guest JF: Annual cost of bipolar disorder to UK society. Br J Psychiatry 2002; 180:227–233
17. Dean BB, Gerner D, Gerner RH: A systematic review evaluating health-related quality of life, work impairment, and healthcare costs and utilization in bipolar disorder. Curr Med Res Opin 2004; 20:139–154
18. Kleinman L, Lowin A, Flood E, Gandhi G, Edgell E, Revicki D: Costs of bipolar disorder. Pharmacoeconomics 2003; 21:601–622
19. Kessler RC, Merikangas KR: The National Comorbidity Survey Replication (NCS-R): background and aims. Int J Methods Psychiatr Res 2004; 13:60–68
20. Kessler RC, Abelson J, Demler O, Escobar JI, Gibbon M, Guyer ME, Howes MJ, Jin R, Vega WA, Walters EE, Wang P, Zaslavsky A, Zheng H: Clinical calibration of DSM-IV diagnoses in the world mental health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMH-CIDI). Int J Methods Psychiatr Res 2004; 13:122–139
21. Kessler RC, Ustun TB: The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 2004; 13:93–121
22. First MB, Spitzer RL, Gibbon M, Williams JBW: Structured Clinical Interview for DSM-IV Axis I Disorders, Research Version, Non-Patient Edition (SCID-I/NP). New York, New York State Psychiatric Institute, Biometrics Research, 2002
23. Keller MB, Hirschfeld RM, Hanks D: Double depression: a distinctive subtype of unipolar depression. J Affect Disord 1997; 45:65–73
24. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB: The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003; 54:573–583
25. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, Pronk N, Simon G, Stang P, Ustun TB, Wang P: The World Health Organization Health and Work Performance Questionnaire (HPQ). J Occup Environ Med 2003; 45:156–174
26. Kessler RC, Ames M, Hymel PA, Loeppke R, McKenas DK, Richling DE, Stang PE, Ustun TB: Using the World Health Organization Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. J Occup Environ Med 2004; 46:S23–S37
27. Laurent A, Cannell CF, Marquis KH: Reporting health events in household interviews: effects of an extensive questionnaire and a diary procedure. Vital Health Stat 1 1972; 2:1–80
28. Wolter KM: Introduction to Variance Estimation. New York, Springer-Verlag, 1985
29. Research Triangle Institute: SUDAAN: Professional Software for Survey Data Analysis. Research Triangle Park, NC, Research Triangle Institute, 2002
30. Bauer MS, Simon GE, Ludman E, Unutzer J: "Bipolarity" in bipolar disorder: distribution of manic and depressive symptoms in a treated population. Br J Psychiatry 2005; 187:87–88
31. Stewart WF, Ricci JA, Chee E, Hahn SR, Morganstein D: Cost of lost productive work time among US workers with depression. JAMA 2003; 289:3135–3144
32. Kessler RC, Barber CB, Birnbaum HG, Frank RG, Greenberg PE, Rose RM, Simon GE, Wang PS: Depression in the workplace: effects on short-term work disability. Health Aff (Millwood) 1999; 18:163–171
33. Simon GE, Barber C, Birnbaum HG, Frank RG, Greenberg PE, Rose RM, Wang P, Kessler RC: Depression and work productivity: the comparative costs of treatment versus nontreatment. J Occup Environ Med 2001; 43:2–9
34. Rost K, Smith JL, Dickinson M: The effect of improving primary care depression management on employee absenteeism and productivity: a randomized trial. Med Care 2004; 42:1202–1210
35. Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W, Pickering RP, Kaplan K: Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Arch Gen Psychiatry 2004; 61:807–816
36. Hirschfeld RM, Calabrese JR, Weissman MM, Reed M, Davies MA, Frye MA, Keck PE Jr, Lewis L, McElroy SL, McNulty JP, Wagner KD: Screening for bipolar disorder in the community. J Clin Psychiatry 2003; 64:53–59
37. Judd LL, Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W, Paulus MP, Kunovac JL, Leon AC, Mueller TI, Rice JA, Keller MB: A prospective 12-year study of subsyndromal and syndromal depressive symptomatology in unipolar depressive disorders. Arch Gen Psychiatry 1998; 55:694–701
38. Pezawas L, Angst J, Gamma A, Ajdacic V, Eich D, Rossler W: Recurrent brief depression—past and future. Prog Neuropsychopharmacol Biol Psychiatry 2003; 27:75–83
39. Rapaport MH, Judd LL: Minor depressive disorder and subsyndromal depressive symptoms: functional impairment and response to treatment. J Affect Disord 1998; 48:227–232
40. Calabrese JR, Hirschfeld RM, Reed M, Davies MA, Frye MA, Keck PE, Lewis L, McElroy SL, McNulty JP, Wagner KD: Impact of bipolar disorder on a US community sample. J Clin Psychiatry 2003; 64:425–432
41. Wang PS, Simon GE, Kessler RC: The economic burden of depression and the cost-effectiveness of treatment. Int J Methods Psychiatr Res 2003; 12:22–33
42. Bauer MS, Kirk GF, Gavin C, Williford WO: Determinants of functional outcome and healthcare costs in bipolar disorder: a high-intensity follow-up study. J Affect Disord 2001; 65:231–241
43. Judd LL, Akiskal HS: Depressive episodes and symptoms dominate the longitudinal course of bipolar disorder. Curr Psychiatry Rep 2003; 5:417–418
44. Hirschfeld RM: Bipolar depression: the real challenge. Eur Neuropsychopharmacol 2004; 14(suppl 2):S83–S88
45. Hirschfeld RM, Holzer C, Calabrese JR, Weissman M, Reed M, Davies M, Frye MA, Keck P, McElroy S, Lewis L, Tierce J, Wagner KD, Hazard E: Validity of the mood disorder questionnaire: a general population study. Am J Psychiatry 2003; 160:178–180