The Influence of Sociodemographic Characteristics on Agreement Between Self-Reports and Expert Exposure Assessments

Grace Sembajwe, ScD,[1,2,3]* Margaret Quinn, ScD,[1] David Kriebel, ScD,[1] Anne Stoddard, ScD,[4] Nancy Krieger, PhD,[2] and Elizabeth Barbeau, ScD[2,3]

Background *Often in exposure assessment for epidemiology, there are no highly accurate exposure data and different measurement methods are considered. The objective of this study was to use various statistical techniques to explore agreement between individual reports and expert ratings of workplace exposures in several industries and investigate the sociodemographic influences on this agreement.*

Methods *A cohort of 1,282 employees at 4 industries/14 worksites answered questions on workplace physical, chemical, and psychosocial exposures over the past 12 months. Occupational hygienists constructed job exposure matrices (JEMs) based on worksite walkthrough exposure evaluations. Worker self-reports were compared with the JEMs using multivariable analyses to explore discord.*

Results *There was poor agreement between the self-reported and expert exposure assessments, but there was evidence that agreement was modified by sociodemographic characteristics. Several characteristics including gender, age, race/ethnicity, hourly wage and nativity strongly affected the degree of discord between self-reports and expert raters across a wide array of different exposures.*

Conclusions *Agreement between exposure assessment tools may be affected by sociodemographic characteristics. This study is cross-sectional and therefore, a snapshot of potential exposures in the workplace. Nevertheless, future studies should take into account the social contexts within which workplace exposures occur.* Am. J. Ind. Med. 53:1019–1031, 2010. *© 2010 Wiley-Liss, Inc.*

KEY WORDS: exposure assessment; occupational epidemiology; job exposure matrix; agreement; discord; higher estimation; lower estimation

INTRODUCTION

In exposure assessment for occupational epidemiology, questionnaires are commonly used to collect exposure data, largely because they are simple and cost-efficient [Nieuwenhuijsen, 2005; Perry et al., 2006]. Although other qualitative and quantitative (e.g., biomonitoring) methods for data gathering exist, questionnaire surveys are a major

[1]Department of Work Environment, University of Massachusetts Lowell, Lowell, Massachusetts
[2]Department of Society, Human Development and Health, Harvard School of Public Health, Boston, Massachusetts
[3]Center for Community-Based Research, Dana-Farber Cancer Institute, Boston, Massachusetts
[4]Department of Statistical Analysis and Research, New England Research Institute, Watertown, Massachusetts

None of the authors have an affiliation with an organization that, to their knowledge, has a direct interest, particularly a financial interest, in the subject matter or materials described.

Contract grant sponsor: NIOSH; Contract grant numbers: R01 OH07366-01, R01 OH07366-01S.

*Correspondence to: Dr. Grace Sembajwe, Research Fellow, Department of Society, Human Development and Health, Harvard School of Public Health, Boston, MA 02115.
E-mail: grace_sembajwe@dfci.harvard.edu

Accepted 9 January 2010
DOI 10.1002/ajim.20821. Published online 19 March 2010 in Wiley Online Library (wileyonlinelibrary.com).

tool and a number of validation studies have been conducted to investigate their reliability and accuracy [Kleinman et al., 1986; Kromhout et al., 1987; Hertzman et al., 1988; Teschke et al., 1989, 2000; Owen et al., 1992; Stewart and Stewart, 1994a; Hammond et al., 1995; Benke et al., 1997; Calvert et al., 1997; Bauer et al., 1999; Tielemans et al., 1999; Quinn et al., 2001]. These include agreement analyses, such as those conducted between self-reported exposures, expert assessments, and urine metabolites, as a result of dermal exposures to chemical hazards [Perry et al., 2006; Sembajwe, 2007]. Agreement analyses are usually used for comparing job hazard ratings that have been made by two or more individuals who are all trying to assess the same exposure.

Several studies have found that individual self-reports were better able to estimate exposures to agents that were tactile or easily sensed (like vibrations) [Benke et al., 2001]. A related factor in the accuracy of self-reported exposures is employee awareness of the chemicals with which they work and familiarity with the industrial or conventional names for substances (such as stainless steel) rather than specific chemical compound names (such as chromium) [Hawkins and Evans, 1989; Wiktorin et al., 1996; Calvert et al., 1997; Washington State, 2000]. Validation of self-assessment of occupational exposures has been conducted elsewhere [Kleinman et al., 1986; Kromhout et al., 1987; Hertzman et al., 1988; Teschke et al., 1989, 2000; Owen et al., 1992; Stewart and Stewart, 1994a,b; Hammond et al., 1995; Fritschi et al., 1996; Benke et al., 1997; Siemiatycki et al., 1997; Bauer et al., 1999; Tielemans et al., 1999; Hu et al., 2002; Nieuwenhuijsen, 2005; Perry et al., 2006]. In particular, comparisons between self-assessments and expert ratings of exposures have been evaluated. For example, one study [Perry et al., 2006] found poor agreement when comparing urinalysis deethylatrazine results with self-reported dermal, inhalation, and ingestion exposure. Another two [Kromhout et al., 1987; Fritschi et al., 1996] concluded that self-assessments should be used with a broader exposure assessment strategy that includes qualitative and quantitative information; they placed more confidence in the expert assessments of occupational exposure. Another investigation [Hu et al., 2002] compared self-assessments, expert ratings, and direct air measurements (using personal air samplers) to measure benzene, toluene, styrene, and xylene, in female petrochemical workers in China. They found poor agreement between self-assessments and direct measurements (kappa < 0.4) for all the chemicals. They also found moderate agreement between expert assessments and direct measurements (kappa ~ 0.6). A study from the Netherlands [Tielemans et al., 1999] also discovered poor agreement between the worker self-reports, job exposure matrices (JEMs), and metabolites measured in urine, for exposure to toluene, xylene, glycol ethers, trichloroethylene, and chromium among manufacturing workers. There may be a number of reasons that the different assessment methods did not agree well, including pharmacokinetics (for metabolites) and sampling dates and times for the air monitoring results.

As seen from these examples, there are a number of ways in which to measure agreement between exposure ratings [Cohen, 1960; Goodman, 1979; Fienberg, 1980; Maclure and Willett, 1987; Agresti, 1988; Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990; Posner et al., 1990; Buettner and Garbe, 2000; Carlin et al., 2000; Graham and Jackson, 2000; Nelson and Pepe, 2000; Lester Kirchner and Lemke, 2002] and all of these studies used standard agreement analyses, such as Cohen's Kappa [Cohen, 1960] (often called kappa) even though they collected data representing different levels: some data were collected at a group level (e.g., expert ratings or job assignments) while others were measured at the individual level (e.g., self-reports and urine metabolite levels). Because these data were collected at different levels (expert assessments do not typically evaluate the exposures of each individual, but rather make estimates for jobs or departments in which more than one worker is employed), multilevel analysis, sometimes called hierarchical regression, generally would be considered a necessary tool for correctly analyzing data with this kind of structure [Groves, 1989; Kleinbaum et al., 1998]. Using multilevel models through generalized estimating equations (GEE) in these studies may have yielded further information on agreement by allowing data at various levels within an organizational hierarchy to be examined as a function of several levels of predictors [Groves, 1989; Kleinbaum et al., 1998]. In addition to exploring agreement with multilevel regression analyses, these studies may have discovered supplemental patterns of agreement by accounting for the social context in which these exposures occur.

The objective of this study was to use a range of statistical modeling techniques to evaluate the agreement between individual self-reports and expert ratings of workplace exposures across a range of industries and jobs, and to assess whether this agreement was modified by individual sociodemographic characteristics.

METHODS

Study Background

These analyses are part of the larger United for Health Study conducted by the Dana Farber Cancer Institute, Harvard School of Public Health and the University of Massachusetts Lowell. The larger study aims and methodology are described elsewhere [Krieger et al., 2005, 2006; Barbeau et al., 2007; Quinn et al., 2007]. In brief, the United for Health study was a cross-sectional survey with the specific aim of looking at a United States population of unionized workers, to characterize the distribution of occupational physical and social hazards, their patterning by race/ethnicity, gender, and wage level, and their contribution

to social inequalities in health. The study was a collaborative with industry managers and labor unions, with the latter being the primary source of participant recruitment [Barbeau et al., 2007]. The study participant response rate was 72% on average per site, with a range of 65–87% [Krieger et al., 2005, 2006; Barbeau et al., 2007; Quinn et al., 2007]. Each research site's institutional review board approved all the protocols, consent forms, and survey instruments for the United for Health Study. Each study participant signed an informed consent form.

Data Collection

Information was collected from four industries, including 14 worksites. Questionnaires collected individual level information; exposure assessment experts (study team occupational hygienists) also conducted worksite walkthrough exposure evaluations at the facilities, collecting group level information about the jobs, departments, worksites, and industries. The occupational hygienists used an exposure evaluation checklist with questions and rating scales similar to those used in the study subjects' questionnaires.

An audio computer-assisted self-interviewing system (ACASI) was used to deliver the questionnaire to study participants in either English or Spanish, depending on the choice of the participant [Barbeau et al., 2007]. Trained survey assistants were available to introduce the main survey, collect anthropometric information (height, weight, and blood pressure), administer a separate job history questionnaire, and answer questions relating to ACASI.

The rationale for the development of the occupational exposure questions and rating scales is described elsewhere, as is the question and scale development for social hazards [Krieger et al., 2005, 2006; Barbeau et al., 2007; Quinn et al., 2007]. The exposures of interest were dust, chemicals, noise, awkward postures of the neck, shoulder and back, heavy lifting, and two components of job strain—demand and control. Simultaneous information on these occupational exposures of interest was recorded by two to three occupational hygienists during worksite walkthrough evaluations using a checklist with similar exposure questions and rating scales as those in the worker questionnaire. The findings from these walkthrough evaluations were used to construct JEMs for each worksite.

Exposure Scoring

From both the questionnaire self-report and walkthrough assessments, information was gathered on the following exposures: dust, chemicals and noise; awkward postures of the neck, shoulder and back; heavy lifting; and psychosocial demand and control.

Dust and chemicals

Respondents were asked to note their exposures to dust and chemicals in the last 12 months in one of four categories: never, rarely, sometimes, and often. For agreement analyses, these were collapsed into three: never (0), rarely or sometimes (1), and often (2). The never category also included respondents who said that they had been exposed to dust at work, but not in the last 12 months.

Noise

Self-reported noise exposures in the last 12 months were categorized as none/low (0) for less than 3 hr of exposure to a noisy area during a work shift; medium (1) for 3–6 hr of exposure per work shift; and high (2) for more than 6 hr of exposure in a work shift.

Neck, shoulder, and back

Self-reported neck, shoulder, and back exposures (awkward postures) in the last 12 months were scored as no/low (0) if there was less than 1 hr of exposure in a work shift; medium (1) for 1–4 hr of exposure per work shift; and high (2) for more than 4 hr of exposure in a work shift. Pictograms were also used to illustrate the awkward postures.

Heavy lifting

Self-reported exposure to heavy lifting in the past 12 months was rated on a three-point scale: light (0), moderate (1), and heavy (2) when rating the heaviest object lifted for the job.

Demand and control

Self-reported job demand and control were defined using four questions for control and three questions for demand from Karasek's Job Content Questionnaire survey [Karasek, 1985]. Demand and control scores were calculated using standard factor analysis methods and were found to be normally distributed; they were then divided into tertiles for the low (0), medium (1), and high (2) categories. Details of the job strain variables and their components are described elsewhere [Quinn et al., 2007; Sembajwe, 2007].

Additional variables of interest

Sociodemographic characteristics were included in the agreement analyses for self-reports and expert walkthrough assessments of occupational exposures, to investigate whether these individual or group characteristics influenced the degree of agreement between self-report and walkthrough exposure scores. The sociodemographic

characteristics used to explore these exposure relationships were gender, race, age, education, wage, nativity, language (the questionnaire could be completed in English or Spanish), industry and industry/job combinations. Further details regarding these variables are presented elsewhere [Quinn et al., 2007].

Occupational Hygiene Evaluations

In the United for Health study, two to three professional occupational hygienists and safety specialists conducted walkthrough assessments of each worksite and documented observed hazards in a JEM. Worksite-specific JEMs were constructed first by defining the duties assigned to the job titles in each department of a specific workplace. Following these assignments, a table listing the departments and job titles along with the type of hazardous exposure (psychosocial stress or job strain, dust, chemicals, noise, musculoskeletal or awkward postures—neck, shoulder, heavy lifting, back) was designed. Based on a consensus of professional judgment [Monge et al., 2005; D'Souza et al., 2007], a value of 0 for no or low exposure, 1 for medium exposure, and 2 for high exposure was given to each job title for each exposure. Because only one walkthrough visit was made to each worksite, the expert assessments represented exposures at just one point in time, unlike the self-reports which asked about exposures over the past 12 months.

Criteria for exposure assignments were developed by the occupational hygienists. Each exposure was rated using the same scale for the JEM as it was in the questionnaire. For example, a noise level assignment of no/low corresponded to a response on the questionnaire that would be expected to also be assigned to the no/low category (less than 3 hr of exposure to a noisy area) by an individual worker. Each job title was reviewed in detail during the walkthrough evaluations and the assignment was a consensus value reached by the occupational hygienists [Monge et al., 2005; D'Souza et al., 2007]. The physical hazards scored were the same as those in the self-report questionnaires. There is a third dimension of job strain considered by the JCQ—social support. This was included in the questionnaire but was not included in the analyses of self-reported responses, because it was not possible to assess this adequately at the job level during the walkthrough evaluations.

Musculoskeletal strain exposures (awkward postures of the back, shoulder, neck; repetitive motions of the hands; and heavy lifting) were evaluated through observation and by using the NIOSH lifting equations as a guide. The NIOSH lifting equation is a tool for measuring single or combinations of lifting tasks and trunk posture to help determine risks associated with the loads being carried [Badger, 1981; Waters et al., 1994].

Dust and chemical exposures were observed directly and through information shared by the supervisors or employees about usual job tasks. Information about the work, materials, practices, unit processes and tasks involved in the job and department also was collected. Detailed notes of the walkthrough observations were recorded and reviewed by the expert assessors.

To complete the noise section of the JEM, the occupational hygienists used their observations and direct experience of the noise, comparing it to the Occupational Safety and Health Administration (OSHA) standards for very loud noises and the World Health Organization (WHO) standards for occupational and community noise [WHO, 2001]. These guidelines list the types of activities and equipment or machinery that produce loud (harmful) noise or interfere with speech.

Data Analysis

To explore the agreement between the self-reported exposures and occupational hygiene walkthrough assessments, percent agreement, kappa, loglinear analyses, and generalized linear models (GLM) were used. Each exposure was analyzed in a separate model. In addition, because group and individual levels of exposure and sociodemographic data were collected in the United for Health study, multilevel regression models were used.

A group level variable that combined industry and job was created. This variable consisted of a set of job titles fully nested within the industries. Details about the industry/job variable creation have been described elsewhere [Quinn et al., 2007].

Because the JEM (walkthrough) assignments were made at the job title level, kappa and loglinear analyses were also used to measure agreement between an average of the self-reported exposures at the job title level compared to the expert assessment of the exposure for that job title. For multilevel modeling, this aggregation was not necessary, and the individual and group level scores were given their correct weights and their correlation structure was taken into account.

Agreement models

The following statistics were calculated to assess different aspects of agreement. *Percent agreement*: Initial cross-tabulations of agreement or frequency counts between the different exposures were used to calculate percent agreement. Usually, agreement of less than 70% is considered ''poor'' and greater than 85% is considered ''strong'' [Hunt, 2006]. *Kappa*: Kappa scores were also used to evaluate agreement between the self-assessments and expert ratings. Weighted kappa was used to account for the ordinal rating scale for the exposures. Weighted kappa values less than 0.4 are often considered ''poor'' agreement, while values greater than 0.7 are considered ''good'' agreement

[Fleiss, 1981; Hunt, 2006]. Both kappa and percent agreement were reviewed within stratified groups of sociodemographic variables and industry, industry/job variables, to evaluate whether agreement was dependent upon sociodemographic characteristic. *Loglinear models*: Another way to measure and model agreement uses loglinear modeling. These models are more flexible in assessing agreement and allow for the evaluation of the effect of modifiers or covariates on patterns of agreement [Fienberg, 1980; Maclure and Willett, 1987; Agresti, 1988; Buettner and Garbe, 2000; Graham and Jackson, 2000; Yoder et al., 2001; Lester Kirchner and Lemke, 2002; Schuster, 2002]. Exact and marginal agreement with sociodemographic variables were explored. Loglinear models yield odds ratios of agreement. For data similar to those in this study, odds ratios of less than 5 are considered ''poor'' agreement while those greater than 45 indicate ''strong'' agreement [Hunt, 2006]. Tests of model significance were based on likelihood ratios or Akaike's Information Criterion (AIC) [Fienberg, 1980; Yoder et al., 2001; Lester Kirchner and Lemke, 2002; Schuster, 2002; Stegmann and Lucking, 2007].

Modeling discord with multilevel models

Multilevel regression models were also used to explore agreement between the expert assessments and self-reported exposures by modeling accord (or discord)—the condition of a participant rating his/her exposure similar to or different from the expert rater. Further, the two types of discord can be studied separately—the self-report higher than the expert or vice versa. This approach has the advantage that it allows exploration of patterns in which one party consistently rates exposures higher/lower than the other. This approach was used further to explore whether there were patterns of accord/discord which varied by sociodemographic characteristics of the study participants.

A discord outcome variable was created by assigning a 1 to all ratings that did not match perfectly and a 0 otherwise. This meant that the discord variable captured all levels of assessment that did not agree perfectly. GLM were used for regression analyses. Models that nested individual variables within department had the best fit (using AIC) and consistently converged. After these models showed various patterns of results and some strong sociodemographic predictors of discord (Appendix), it seemed useful to investigate a particular type of discord, that is, self-reported exposure *higher* than the expert assessment. The choice of this type of discord was somewhat arbitrary. The reverse condition, where self-reports were lower than the expert assessment, was also studied and the results were quite similar (but measures of association were inverted). The newly constructed higher discord variable was coded 1 in either of two scenarios: when walkthrough assignments were 0 (no/low) and self-reports were 1 (medium) or 2 (high); and when walkthroughs were 1 (medium) and self-reports were 2 (high). This variable was designed to explore any patterns of association between sociodemographic characteristics and a higher self-report of exposure in comparison to walkthrough assignments.

Because there was a large number of bus drivers and they all were listed under one job title (which would have conferred perfect agreement or perfect discord within for the entire group), the agreement analyses were repeated without this group, to investigate their influence. All data analyses were conducted in SAS® version 9.1.

RESULTS

There were 1,282 eligible respondents in the United for Health study. The population was of a diverse racial/ethnic background with approximately two-thirds male. Half were foreign-born and a majority chose the English language questionnaire (Table I). The study population had an average

TABLE I. Description of Study Population

	N (%)
Number in total study population	1,282 (100)
Gender	
Female	456 (36)
Male	791 (62)
Missing	35 (3)
Race/ethnicity	
Black	505 (39)
Hispanic	292 (23)
White	314 (24)
Other	130 (10)
Missing	41 (3)
Hourly wage	
$6.00–10.54/hr	403 (31)
≥$10.55/hr	774 (60)
Missing	105 (8)
Nativity	
Born in U.S.	607 (47)
Born outside U.S.	616 (48)
Missing	59 (5)
Language of questionnaire	
English	1,060 (83)
Spanish	222 (17)
Missing	0
Education	
Less than 12th grade	296 (23)
High school/GED	470 (36)
Beyond high school	400 (31)
Missing	116 (9)

N, frequency; %, percent.

age of 45 years with almost ten years of tenure in their current job (not shown in the table); a quarter had less than high school education and one-third made below $10.55 per hour—the living wage for the geographic region from which the population was drawn.

Almost three-quarters of the study participants were employed in the school bus service and grocery store retail industries; 12% worked in meat packing and 16% were in electrical light fixture manufacturing. The number of job titles and departments varied by industry; there was also a substantial difference between the number of self-ascribed job titles and official or company-assigned job titles that were recorded by the occupational hygienists during the walkthroughs (Table II). Walkthrough job titles were matched with self-reported job titles based on job task information.

As noted elsewhere [Quinn et al., 2007] only 15% of the study population reported having no high exposures to any of the identified hazards; 85% of the population identified at least one high exposure at their workplace.

Measures of Agreement

There was generally poor agreement between self-reports and walkthrough exposure ratings. Agreement was poor across all exposures, whether measured at the job title level (with average worker scores for comparison), at the individual level (where each worker was assigned the job title level score for comparison), or with/without bus drivers (data not shown). Percent agreement, kappa, and loglinear models were uniformly low, with percent agreement less than 50%, kappa less than 0.4 (with some negative kappa results), and odds ratios of agreement less than 5.0 across all exposures [Fleiss, 1981; Hunt, 2006].

Mixed Models

In multilevel models that nested individuals within departments, there were minimal adjustments in the strengths of the associations found in the non-nested models and the qualitative message remained the same. However, the department level variable was significant in all the models, indicating a degree of clustering by exposure, within departments and possibly job titles, with $P < 0.001$ for the department level variable and an intra-class correlation of 0.4 (for clusters of individuals within department). Each department and job titled was nested within its specific industry.

Discord Models

Figure 1 shows results from investigations into the agreement and discord between self-report and expert assessment across exposures. Individuals reported higher exposure than the experts generally about a third of the time for most exposures. The notable exception was job demand, for which higher estimation only occurred for 4% of participants, while 70% reported lower demand than the experts (lower estimation).

There were interesting patterns of higher estimation of exposures in self-reports compared to experts, when the population was stratified by sociodemographic characteristics (Table III). Women were generally more likely than men to report higher estimates of exposure relative to experts. This was particularly pronounced for awkward shoulder posture, prolonged noise exposure, and psychosocial job demands, all of which had prevalence ratios greater than 1.5. Older workers were consistently less likely than younger workers to report higher exposures than the expert raters. Among race/ethnicity groups, Black and Other race were relatively less likely than Whites to report higher exposures. The pattern for Hispanics was inconsistent, with some exposures more frequently reported higher and others lower than Whites. The lower wage workers (earning less than the living wage of $10.55 per hour) were considerably more likely to have relatively higher estimates than those earning more. This was particularly true for the musculoskeletal exposures and high noise for which prevalence ratios were greater than 2.0. Those born outside the U.S. were less likely than the native born to report higher estimates of exposure than the expert raters. The pattern was inconsistent for the variable indicating whether the survey was taken in English

TABLE II. Work Sites and Job Titles Within Industry, United for Health Study, Boston, Massachusetts

Industry	Work sites, N (%)	Departments, N (%)	Job titles (walkthrough), N (%)	Job titles (self-report), N (%)	Individuals, N (%)
School bus service	4 (29)	1 (2)	1 (1)	24 (6)	497 (39)
Meat and meat product (wholesale)	1 (7)	20 (36)	23 (25)	85 (23)	157 (12)
Grocery store (retail)	7 (50)	13 (24)	22 (25)	161 (43)	417 (33)
Light fixtures (manufacturing)	2 (14)	21 (38)	44 (49)	101 (27)	211 (16)
Total	14 (100)	55 (100)	90 (100)	371 (100)	1,282 (100)

N, frequency; %, percent.

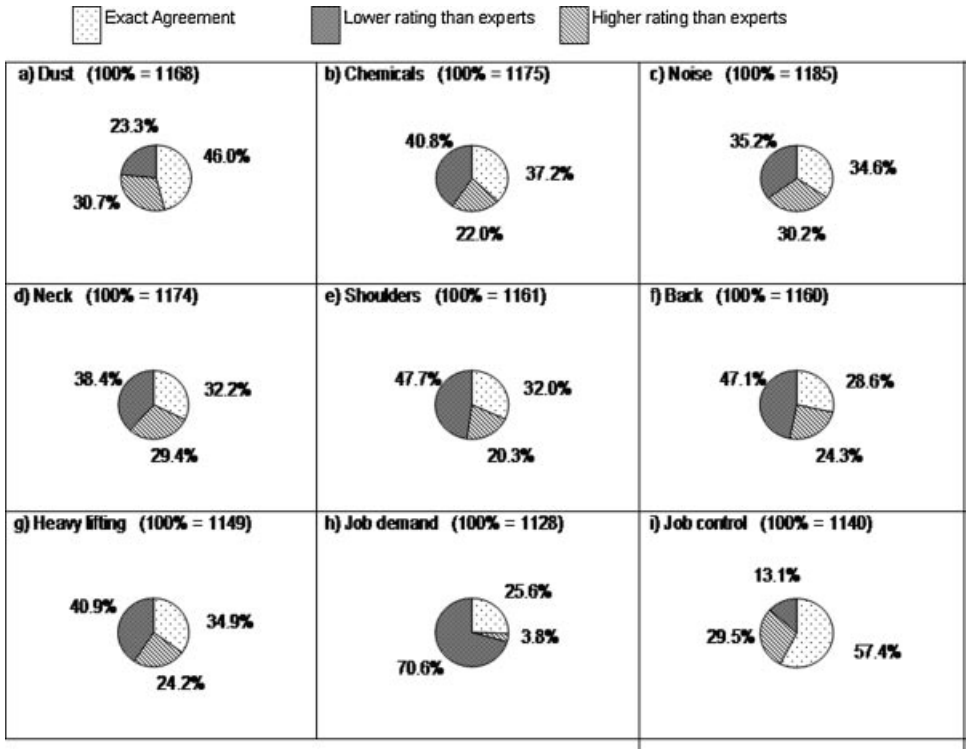

FIGURE 1. Agreement and discord for self-reported exposure ratings compared to expert walkthrough assessments.

or Spanish, and there were few important differences in relative exposure estimation among workers with different levels of education.

DISCUSSION

Comparative studies of expert evaluations and self-reported exposure assessments have recognized that both can be biased [Stewart and Stewart, 1994a]. Several formal validation studies that compared expert assessments to quantitative measurements have found that expert ratings, for example, were highly influenced by the amount of information about the job or worksite available to the experts [Kromhout et al., 1987; Stewart and Stewart, 1994a; Calvert et al., 1997; Tielemans et al., 1999].

Neither walkthrough evaluations (used for the expert assessments in this study) nor self-reports are a ''gold standard'' for exposure assessment and caution should be used when relying upon either form during data collection [Stewart and Stewart, 1994a,b; Fritschi et al., 1996; Siemiatycki et al., 1997]. Self-reported exposures have certain advantages: they are measured at the individual level; they offer a summary of exposure over specific periods of time; they are relatively inexpensive; and they can cover a wide range of exposures. However, from a wide range of industries, self-reports also have some disadvantages: they are subjective; their scale of measurement can be difficult to evaluate; they are open to recall bias; and they can be influenced by individual perception and willingness to report, which may both be socially determined.

Expert walkthrough assessments also have certain advantages: they are generally considered more objective (relative to self-reports); their scale of measurement can be better defined; and they are relatively time- and cost-efficient. There are some disadvantages to expert judgments: they are not truly objective (experts may bring personal as well as professional biases to their assessments); they typically reference only one point in time; and they typically assess exposures at the job level and cannot account for individual employee level variability.

Several validation studies of self-assessment of occupational exposures have been conducted elsewhere [Kleinman et al., 1986; Kromhout et al., 1987; Hertzman et al., 1988; Teschke et al., 1989, 2000; Owen et al., 1992; Stewart and Stewart, 1994a,b; Hammond et al., 1995; Fritschi et al., 1996; Benke et al., 1997; Siemiatycki et al., 1997; Bauer et al., 1999; Tielemans et al., 1999; Hu et al., 2002; Nieuwenhuijsen, 2005; Perry et al., 2006]. In particular, comparisons between self-assessments and expert ratings of exposures often show moderate to poor agreement, with more confidence being placed on expert ratings [Kromhout et al., 1987; Fritschi et al., 1996; Tielemans et al., 1999; Hu et al., 2002; Perry et al., 2006]. In this study, agreement between self-reported and walkthrough assessments varied

TABLE III. Prevalence Ratios for Higher* Self-Reported Exposure Ratings Across Sociodemographic Characteristics

	Neck PR (95% CI) (N = 1,174)	Shoulder PR (95% CI) (N = 1,161)	Back PR (95% CI) (N = 1,160)	Heavy lifting PR (95% CI) (N = 1,149)	Dust PR (95% CI) (N = 1,168)	Chemical PR (95% CI) (N = 1,175)	Noise PR (95% CI) (N = 1,185)	Demand PR (95% CI) (N = 1,128)	Control PR (95% CI) (N = 1,140)
Gender									
Female	**1.43 (1.20–1.71)**	**1.73 (1.38–2.16)**	**1.25 (1.02–1.53)**	1.18 (0.96–1.45)	1.17 (0.98–1.40)	**1.38 (1.11–1.71)**	**1.57 (1.33–1.86)**	0.57 (0.28–1.14)	**1.68 (1.41–2.00)**
Male[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Age									
≥45	**0.59 (0.49–0.71)**	**0.61 (0.48–0.77)**	**0.59 (0.48–0.73)**	**0.61 (0.49–0.75)**	**0.71 (0.60–0.85)**	**0.64 (0.51–0.80)**	**0.57 (0.47–0.68)**	**0.32 (0.16–0.63)**	**0.64 (0.53–0.78)**
<45[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Race/ethnicity									
Hispanic	1.08 (0.90–1.30)	0.96 (0.72–1.28)	1.19 (0.46–1.48)	**0.74 (0.58–0.96)**	**0.70 (0.57–0.87)**	**0.50 (0.36–0.69)**	1.14 (0.95–1.37)	0.72 (0.36–1.45)	**1.89 (1.47–2.43)**
Black	**0.28 (0.21–0.36)**	**0.47 (0.35–0.64)**	**0.29 (0.22–0.40)**	**0.37 (0.28–0.48)**	**0.39 (0.31–0.48)**	**0.45 (0.35–0.57)**	**0.31 (0.24–0.40)**	**0.09 (0.03–0.30)**	1.11 (0.87–1.43)
Other	**0.56 (0.40–0.78)**	0.82 (0.56–1.19)	**0.57 (0.39–0.83)**	**0.56 (0.39–0.82)**	**0.56 (0.41–0.76)**	**0.51 (0.34–0.76)**	**0.58 (0.42–0.81)**	0.45 (0.16–1.29)	1.31 (0.94–1.82)
White[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hourly wage									
$6.00–10.54/hr	**2.25 (1.88–2.68)**	**2.45 (1.95–3.10)**	**2.09 (1.71–2.57)**	**1.52 (1.24–1.87)**	1.14 (0.95–1.37)	**1.51 (1.22–1.88)**	**2.08 (1.75–2.47)**	1.15 (0.63–2.12)	**1.88 (1.57–2.26)**
≥$10.55/hr[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Nativity									
Born outside U.S.	**0.64 (0.53–0.77)**	**0.58 (0.45–0.74)**	**0.71 (0.58–0.88)**	**0.46 (0.37–0.58)**	**0.50 (0.41–0.60)**	**0.37 (0.29–0.48)**	**0.55 (0.46–0.67)**	0.53 (0.29–1.00)	1.07 (0.89–1.28)
Born in U.S.[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Language									
Spanish	**2.10 (1.75–2.51)**	1.15 (0.84–1.59)	**2.22 (1.80–2.74)**	1.19 (0.90–1.57)	1.01 (0.78–1.30)	**0.50 (0.32–0.78)**	**1.76 (1.46–2.13)**	0.91 (0.36–2.26)	**1.88 (1.55–2.28)**
English[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Education									
<High school	1.16 (0.0.92–1.47)	0.94 (0.68–1.29)	0.96 (0.73–1.26)	**0.67 (0.50–0.90)**	0.98 (0.77–1.23)	**0.65 (0.47–0.89)**	0.92 (0.73–1.16)	0.53 (0.22–1.24)	1.22 (0.96–1.55)
High school/GED	1.05 (0.85–1.31)	1.07 (0.81–1.40)	0.91 (0.72–1.16)	0.90 (0.71–1.13)	0.98 (0.80–1.20)	1.04 (0.82–1.32)	0.94 (0.77–1.16)	0.85 (0.45–1.61)	1.01 (0.80–1.27)
>High School[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

PR, prevalence ratio; 95% CI, 95% confidence interval.
Bold denotes a significant association ($P \leq 0.05$); N, frequency; %, percent.
Each exposure was analyzed in a separate model.
*Compared to walkthrough exposure ratings.
[a]Reference.

by exposure, but in general agreement was poor across all exposures. There are a number of reasons that could have contributed to this discord.

The scales for the survey response options varied depending on the exposure. Individuals responding to the musculoskeletal questions (neck, shoulder, back exposures), were asked to indicate the number of hours in a day that they were exposed (<1 hr, 1–4 hr, or >4 hr). Whereas the heavy lifting question asked workers to rate the weight of the heaviest object routinely lifted on the job, in the past year (light, moderate, or heavy); and the questions on dust and chemical exposure asked if one was ever exposed, with a follow-up question about exposure in the past year (never, rarely, sometimes, or often). The question on noise was also in two parts, with an initial question on working in a noisy area in the past year, followed by a query about the number of hours of exposure in a day (>3 hr, 3–6 hr, or >6 hr). Although the walkthrough assessments tried to use similar scales, individual workers may have been more able to accurately identify exposure patterns (intensity and frequency) that could not be observed at one or two work site visits. The constellation and subjectivity of questions for the demand and control exposures are examples of how the questionnaire scale could not always be reproduced for the walkthrough assessments.

The questionnaire was administered by computer using audio and visual cues in the participant's language of choice. Pictograms for the musculoskeletal questions were included. For these reasons, this questionnaire may have collected more accurate exposure data than the more common paper and pencil questionnaire.

Whether an exposure was continuous or intermittent and its physical form may have also contributed to discord in self-reports and walkthrough ratings. Musculoskeletal exposure questions may have been more tactile than questions about dust and chemical exposure (the graphics used for the awkward posture questions would also have helped with relaying the survey's intent). Intermittent exposures may have been missed during a site visit, or if noticed during the walkthrough, may have been given more importance than assessments from self-reports.

Although symptom questions were not investigated in this report, individual symptoms or pain may have influenced self-reported responses to exposure questions (common variable bias). For example, a worker experiencing neck pain may have indicated high exposures (>4 hr in a shift) to awkward neck postures when expert observations of the job tasks may have rated the exposure as low. The subjectivity of self-reports has been addressed elsewhere and has often been used as a reason to prefer expert assessments of exposure [Teschke et al., 1989; Fritschi et al., 1996; Benke et al., 1997; Stewart et al., 2000; Hu et al., 2002].

During the walkthrough exposure assessments, experts were able to rate jobs based on comparisons within industry, worksite, and department as well as draw upon their experiences across industries, worksites, and departments. This should have reduced subjectivity on the part of the experts, in comparison to the self-reports, but may also have contributed to the observed discord with workers' self-reports.

For this study, kappa, loglinear, GLM, and multilevel models were used. In order to work well, all of these statistical models rely on specific, underlying distributions for the data. There were significant differences in the way exposure scores were assigned by the self-report and walkthrough procedures. Thus, there were unequal margins (or distributions of overall exposure scores) for tables that compared self-reported exposures to occupational hygiene assessments. This imbalance in our data compromised agreement analyses using kappa, since the calculation of kappa relies so heavily on the assumption of balanced margins [Cohen, 1960; Goodman, 1979; Maclure and Willett, 1987; Agresti, 1988; Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990; Posner et al., 1990; Buettner and Garbe, 2000; Carlin et al., 2000; Graham and Jackson, 2000; Nelson and Pepe, 2000; Lester Kirchner and Lemke, 2002]. Thus it may not be so surprising that the kappa scores in our study were poor.

Loglinear models assisted in accounting for the marginal distributions and allowed for a little more flexibility in the agreement analyses [Cohen, 1960; Goodman, 1979; Maclure and Willett, 1987; Agresti, 1988; Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990; Posner et al., 1990; Buettner and Garbe, 2000; Carlin et al., 2000; Graham and Jackson, 2000; Nelson and Pepe, 2000; Lester Kirchner and Lemke, 2002]. These models confirmed that there was poor agreement between the self-reported exposure ratings and the walkthrough assessments. These model results also suggested, as did the negative kappa scores for some of the exposures, that there was a higher degree of discord than would be found by chance.

Loglinear, multilevel, and GLM techniques were used to model and highlight the sociodemographic characteristics that were influencing the discord. These techniques were very useful for modeling discord and through these statistical methods each exposure score with additional covariates was investigated, to check the magnitude of any associations that influenced the lack of agreement. Multilevel modeling was particularly helpful when accounting for the nesting of individuals within job titles and departments.

Discord, Higher Estimation, and Lower Estimation

By design, workers and expert exposure assessors used similar scales of measurement in evaluating exposures. But this consistency of form (usually a 3- or 4-point scale) appears to have been less important than other aspects of the

exposure assessment processes which evidently were quite different between workers and experts. As noted above, there are many possible explanations for this discrepancy including different periods of time (workers were asked about the past year; experts could only evaluate what they saw during their brief visits), different implicit comparisons (workers perhaps referred to their own previous jobs or those of their friends or colleagues; experts relied upon their professional practice and standards in the field), and various kinds of bias on the part of both types of assessors.

Initial investigations of the low level of agreement focused on *discord* and the simple fact that the worker and expert ratings were different. Mathematically, discord is represented by any pattern of responses that is not on the main diagonal of a 3×3 matrix of the worker and expert exposure scores. When this approach seemed to explain very little about the data, it seemed possible that perhaps higher estimation and lower estimation—lumped together when studying discord—were different phenomena with different determinants. This appears to be the case. There were much stronger patterns of differences in who reported higher estimates of exposure (one corner of the 3×3 matrix) than in who was simply not in accord with the experts (similar, inverse patterns occurred for lower estimations, the opposite corner of the 3×3 matrix). While there is no definite explanation for the observed patterns, certain observations may be useful.

(1) There was a strong tendency for the same pattern of relative higher or lower estimation to occur across most or all exposures. That is, how exposures were reported was consistent across sociodemographic characteristics and seemed little affected by the specific exposure in question.
(2) Women were more likely than men to report higher estimates for most exposures, compared to experts. This consistent gender difference could be due to gender-based differences in tasks within jobs, or it could be due to differences in how men and women perceived and rated their exposures.
(3) Older workers were less likely than younger workers to report higher estimates for all exposures, compared to experts. This pattern was quite consistent across the diverse types of exposures. One or both of the same broad explanations used for gender may apply.
(4) Blacks and the other race group were less likely than Whites to report higher exposures than the expert raters.
(5) Lower wage workers were more likely than higher wage workers to report higher exposures relative to experts.
(6) Those born outside the U.S. were less likely than native-born workers to report higher exposures than expert raters.

It is difficult to find a simple explanation for these findings. It appears though that sociodemographic characteristics do affect agreement between self-reports and expert raters across a wide array of different types of exposures, and a range of sociodemographic characteristics. These results also indicate a need to employ an ecosocial framework during exposure assessment for occupational epidemiology, by linking social, biological, and historical (time) when assessing population experiences of factors related to health (and ultimately linking these with health outcomes) [Krieger, 1994, 2001].

One direct way to reduce the discrepancy between self-reported exposures and expert walkthrough assessments at worksites would be during exposure assessments, for experts to be more aware of and document the individual level profiles that make up a job title (job group) or department being assessed. Similarly, individual self-reported data (surveys) should always include information on sociodemographic characteristics to facilitate researcher understanding of the complex social, ecological, and historical relationships that make up the realities of exposure and ultimately how these perceptions and experiences affect health. In addition, during analyses, the sociodemographic characteristics of employees that make up the assessed jobs and/or departments should be included, necessarily. This would require the use of multilevel modeling strategies to evaluate the complex associations that make up individual and group level variability, especially when exposures are linked to health outcomes.

The ultimate goal of these improved exposure assessment strategies would be to increase measurement precision for worksite interventions and most importantly, reduce harmful exposures at the population (or group) level by also reducing individual level (within group) variability in exposure.

CONCLUSIONS

In this study there was poor agreement between self-reported exposures and expert assessments. A number of possible reasons for the poor agreement (discord) between the walkthrough and self-reported assessments are important and include: precision of the survey questions; the time period they ask about; measurement scales; nature of exposure (whether it is intermittent, perceptible); number of individuals per job title or department and variability of the individual exposures within job title; and relative subjectivity of the instruments being used. An analysis of the lack of agreement or discord showed definite patterns in particular types of discord. These patterns of discord were associated with the sociodemographic characteristics of the survey participants, across a wide range of exposures.

These discord analyses are important because often studies of agreement in exposure assessment attribute poor agreement to misclassification by one type of measure or another and do not consider sociodemographic characteristics [Fritschi et al., 1996; Siemiatycki, 1996; Benke et al., 2001; Quinn et al., 2007]. Individual attributes such as age, race/ethnicity, gender, and wage may add important information about perceptions or possible individual sensitivities to workplace exposures [Kennedy and Koehoorn, 2003].

These analyses are a useful contribution to the exploration of agreement between exposure assessment tools; these results suggest and the authors recommend that sociodemographic characteristics should be an important part of exposure assessment for epidemiology, through qualitative and quantitative measurement strategies [Krieger, 1994, 2001; Kennedy and Koehoorn, 2003].

ACKNOWLEDGMENTS

The authors thank other members of our study team (in alphabetical order) Lisa Bates, Gary Bennett, Vanessa Costa, Ruth Lederman, Maribel Melendez, Deepa Naishadham, Michael Ostler, Glorian Sorensen, and Richard Youngstrom. We thank Rebecca Gore for consultation on the data analysis, Judith Gold for consultation on the musculoskeletal strain measures, and Manuel Cifuentes, Bong Kyoo Choi, and Sean Collins for consultation on the job strain measures. We are grateful to our union and worksite collaborators; and most especially, the workers who shared their experiences with us by participating in the study.

REFERENCES

Agresti A. 1988. A model for agreement between ratings on an ordinal scale. Biometrics 44(2):539–548.

Badger D. 1981. Work practices guide for manual lifting: U.S. Department of Health and Human Services. Cincinnati, Ohio: National Institute for Occupational Safety and Health, Publication No. 81-122.

Barbeau EM, Hartman C, Quinn MM, Stoddard AM, Krieger N. 2007. Methods for recruiting white, black, and hispanic working-class women and men to a study of physical and social hazards at work: The united for health study. Int J Health Serv 37(1):127–144.

Bauer E, Romitti P, Reynolds S. 1999. Evaluation of reports of periconceptual occupational exposure: Maternal-assessed versus industrial hygienist-assessed exposure. Am J Ind Med 36(5):573–578.

Benke G, Sim M, Forbes A, Salzberg M. 1997. Retrospective assessment of occupational exposure to chemicals in community-based studies: Validity and repeatability of industrial hygiene panel ratings. Int J Epidemiol 26(3):635–642.

Benke G, Sim M, Fritschi L, Aldred G, Forbes A, Kauppinen T. 2001. Comparison of occupational exposure using three different methods: Hygiene panel, job exposure matrix (JEM), and self reports. Appl Occup Environ Hyg 16(1):84–91.

Buettner P, Garbe C. 2000. Agreement between self-assessment of melanocytic nevi by patients and dermatologic examination. Am J Epidemiol 151(1):72–77.

Calvert G, Mueller C, O'Neill V, Fajen J, Briggle T, Fleming L. 1997. Agreement between company recorded self-reported estimates of duration and frequency to occupational fumigant exposure. Am J Ind Med 32(4):364–368.

Carlin J, Ryan L, Harvey E, Holmes L. 2000. Anticonvulsant teratogenesis 4: Inter-rater agreement in assessing minor physical features related to anticonvulsant therapy. Teratology 62(6):406–412.

Cicchetti D, Feinstein A. 1990. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43(6):551–558.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.

D'Souza JC, Keyserling WM, Werner RA, Gillespie B, Franzblau A. 2007. Expert consensus ratings of job categories from the third national health and nutrition examination survey (NHANES III). Am J Ind Med 50(8):608–616.

Feinstein A, Cicchetti D. 1990. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43(6):543–549.

Fienberg S. 1980. The analysis of cross-classified data. Cambridge, MA: The MIT Press.

Fleiss JL. 1981. Statistical methods for rates and proportions, 2nd edition. New York: John Wiley and Sons. p 216, 218.

Fritschi L, Siemiatycki J, Richardson L. 1996. Self-assessed versus expert-assessed occupational exposures. Am J Epidemiol 144(5):521–527.

Goodman L. 1979. Simple models for the analysis of association in cross-classifications having ordered categories. J Am Stat Assoc 74:537–552.

Graham P, Jackson R. 2000. A comparison of primary and proxy respondent reports of habitual physical activity, using kappa statistics and log-linear models. J Epidemiol Biostat 5(4):255–265.

Groves R. 1989. Survey errors and survey costs. New York, NY: John Wiley & Sons.

Hammond S, Hines C, Hallock M, Hallock M, Woskie S, Abdollahzadeh S, Iden C, Anson E, Ramsey F, Schenker M. 1995. Tiered exposure-assessment strategy in the semiconductor health study. Am J Ind Med 28(6):661–680.

Hawkins N, Evans J. 1989. Subjective estimation of toluene exposures: A calibration study of industrial hygienists. Appl Ind Hyg J 4:61–68.

Hertzman C, Teschke K, Dimich-Ward H, Ostry A. 1988. Validity and reliability of a method for retrospective evaluation of chlorophenate exposure in the lumber industry. Am J Ind Med 14(6):703–713.

Hu Y, Smith T, Xu X, Wang L, Watanabe H, Christiani D. 2002. Comparison of self-assessment of solvent exposure with measurement and professional assessment for female petrochemical workers in china. Am J Ind Med 41(6):483–489.

Hunt P. 2006. Loglinear modeling of agreement. Doctoral Dissertation: University of Massachusetts, Lowell.

Karasek R. 1985. Job content questionnaire and user's guide. Lowell: University of Massachusetts.

Kennedy SM, Koehoorn M. 2003. Exposure assessment in epidemiology: Does gender matter? Am J Ind Med 44(6):576–583.

Kleinbaum D, Kupper L, Muller K, Nizam A. 1998. Applied regression analysis and other multivariable methods. Pacific Grove, CA: Duxbury Press.

Kleinman G, Horstman S, Kalman D, McKenzie J, Stansel D. 1986. Industrial hygiene, chemical and biological assessment of exposures to a chlorinated phenolic sapstain control agent. Am Ind Hyg Assoc J 47(12):731–741.

Krieger N. 1994. Epidemiology and the web of causation: Has anyone seen the spider? Soc Sci Med 39(7):887–903.

Krieger N. 2001. Theories for social epidemiology in the 21st century: An ecosocial perspective. Int J Epidemiol 30(4):668–677.

Krieger N, Smith K, Naishadham D, Hartman C, Barbeau EM. 2005. Experiences of discrimination: Validity and reliability of a self-report measure for population health research on racism and health. Soc Sci Med 61(7):1576–1596.

Krieger N, Waterman PD, Hartman C, Bates LM, Stoddard AM, Quinn MM, Sorensen G, Barbeau EM. 2006. Social hazards on the job: Workplace abuse, sexual harassment, and racial discrimination—A study of Black, Latino, and White low-income women and men workers in the United States. Int J Health Sci 36(1):51–85.

Kromhout H, Oostendorp Y, Heederik D, Boleij J. 1987. Agreement between qualitative exposure estimates and quantitative exposure measurements. Am J Ind Med 12(5):551–562.

Lester Kirchner H, Lemke J. 2002. Simultaneous estimation of intrarater and interrater agreement for multiple raters under order restrictions for a binary trait. Stat Med 21(12):1761–1772.

Maclure M, Willett W. 1987. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126(2):161–169.

Monge P, Partanen T, Wesseling C, Bravo V, Ruepert C, Burstyn I. 2005. Assessment of pesticide exposure in the agricultural population of Costa Rica. Ann Occup Hyg 49(5):375–384.

Nelson J, Pepe M. 2000. Statistical description of interrater variability in ordinal ratings. Stat Methods Med Res 9(5):475–496.

Nieuwenhuijsen M. 2005. Design of exposure questionnaires for epidemiological studies. Occup Eviron Med 62(4):272–280.

Owen C, Acquavella J, Lynch J, Bird M. 1992. An industrial hygiene methodology developed in support of a retrospective morbidity case-control study. Am Ind Hyg Assoc J 53(9):540–547.

Perry M, Marbella A, Layde P. 2006. Nonpersistent pesticide exposure self-report versus biomonitoring in farm pesticide applicators. Ann Epidemiol 16(9):701–707.

Posner K, Sampson P, Caplan R, Ward R, Cheney F. 1990. Measuring interrater reliability among multiple raters: An example of methods for nominal data. Stat Med 9(9):1103–1115.

Quinn MM, Smith TJ, Youk AO, Marsh GM, Stone RA, Buchanich JM, Gula MJ. 2001. Historical cohort study of us man-made vitreous fiber production workers: viii. Exposure-specific job analysis. J Occup Environ Med 43(9):824–834.

Quinn M, Sembajwe G, Stoddard A, Kriebel D, Krieger N, Sorensen G, Hartman C, Naishadham D, Barbeau E. 2007. Social disparities in the burden of occupational exposures: Results of a cross-sectional study. Am J Ind Med 50(12):861–875.

Schuster C. 2002. A mixture model approach to indexing rater agreement. Br J Math Stat Psychol 55(Pt 2):289–303.

Sembajwe G. 2007. Common variable bias in occupational epidemiology. Doctoral Dissertation: University of Massachusetts, Lowell.

Siemiatycki J. 1996. Exposure assessment in community-based studies of occupational cancer. Occup Hyg 3:41–58.

Siemiatycki J, Fritschi L, Nadon L, Gérin M. 1997. Reliability of an expert rating procedure for retrospective assessment of occupational exposures in community-based case-control studies. Am J Ind Med 31(3):280–286.

Stegmann J, Lucking A. 2007. Assessing reliability on annotations (1): Theoretical considerations. Collaborative research center. Technical report. Bielfield, Germany: University of Bielfield. Accessed July 2007.

Stewart P, Stewart W. 1994a. Occupational case-control studies: II. Recommendations for exposure assessment. Am J Ind Med 26(3):313–326.

Stewart W, Stewart P. 1994b. Occupational case-control studies: I. Collecting information on work histories and work-related exposures. Am J Ind Med 26(3):297–312.

Stewart P, Carel R, Schairer C, Blair A. 2000. Comparison of industrial hygienists' exposure evaluations for an epidemiologic study. Scand J Work Environ Health 26(1):44–51.

Teschke K, Hertzman C, Dimich-Ward H, Ostry A, Blair J, Hershler R. 1989. A comparison of exposure estimates by worker raters and industrial hygienists. Scand J Work Environ Health 15(6):424–429.

Teschke K, Smith J, Olshan A. 2000. Evidence of recall bias in volunteered vs. prompted responses about occupational exposures. Am J Ind Med 38(4):385–388.

Tielemans E, Heederik D, Burdorf A, Vermeulen R, Veulemans H, Kromhout H, Hartog K. 1999. Assessment of occupational exposures in a general population: Comparison of different methods. Occup Environ Med 56(3):145–151.

Washington State. 2000. Ergonomic rule (wac 296-62-05174). Appendix B of the Washington State Department of Labor and Industries (L&I). Seattle, Washington: Washington Department of Labor and Industries.

Waters T, Putz-Anderson V, Garg A. 1994. Applications manual for the revised NIOSH lifting equation, Vol. 94-110 NIOSH CDC. Atlanta, GA: Centers for Disease Control & Prevention (NIOSH).

World Health Organization. 2001. Occupational and community noise. Geneva, Switzerland: World Health Organization.

Wiktorin C, Hjelm E, Winkel J, Köster M. 1996. Reproducibility of a questionnaire for assessment of physical load during work and leisure time. Stockholm music in study group. Musculoskeletal intervention center. J Occup Environ Med 38(2):190–201.

Yoder P, Bruce P, Tapp J. 2001. Comparing sequential associations within a single dyad. Behav Res Methods Instrum Comput 33(3):331–338.

APPENDIX: Prevalence ratios for the significant associations between the discord variable* and sociodemographic characteristics

	Neck PR (95% CI) (N = 1174)	Shoulder PR (95% CI) (N = 1161)	Back PR (95% CI) (N = 1160)	Heavy Lifting PR (95% CI) (N = 1149)	Dust PR (95% CI) (N = 1168)	Chemical PR (95% CI) (N = 1175)	Noise PR (95% CI) (N = 1185)	Demand PR (95% CI) (N = 1128)	Control PR (95% CI) (N = 1140)
Gender									
Female	**0.87 (0.78–0.96)**	**0.86 (0.78–0.94)**	0.97 (0.89–1.05)	1.03 (0.95–1.11)	0.97 (0.90–1.04)	0.99 (0.92–1.06)	**0.90 (0.83–0.98)**	**1.08 (1.01–1.15)**	**1.41 (1.26–1.58)**
Male[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Age									
≥ 45	1.05 (0.96–1.16)	**1.09 (1.00–1.18)**	1.03 (0.95–1.11)	1.00 (0.93–1.07)	**1.08 (1.00–1.16)**	1.00 (0.93–1.06)	1.08 (1.00–1.16)	1.04 (0.98–1.11)	**0.84 (0.75–0.94)**
< 45[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Race/ethnicity									
Hispanic	0.98 (0.85–1.12)	**1.15 (1.02–1.31)**	**1.23 (1.10–1.37)**	**1.10 (1.00–1.21)**	**1.11 (1.01–1.22)**	**1.20 (1.10–1.32)**	1.05 (0.93–1.18)	**1.15 (1.04–1.26)**	**1.61 (1.38–1.89)**
Black	1.03 (0.92–1.16)	**1.29 (1.15–1.43)**	1.07 (0.96–1.18)	0.92 (0.84–1.01)	0.99 (0.90–1.09)	1.06 (0.97–1.16)	**1.13 (1.03–1.25)**	**1.21 (1.11–1.32)**	1.10 (0.93–1.30)
Other	0.96 (0.81–1.15)	1.03 (0.86–1.22)	1.05 (0.90–1.22)	1.03 (0.90–1.17)	1.07 (0.95–1.22)	1.02 (0.89–1.16)	1.07 (0.93–1.24)	**1.13 (1.00–1.27)**	1.20 (0.97–1.51)
White[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hourly wage									
$6.00–$10.54/hr	**0.62 (0.54–0.70)**	**0.71 (0.64–0.79)**	**0.83 (0.75–0.91)**	**1.09 (1.01–1.18)**	1.00 (0.92–1.08)	1.02 (0.95–1.10)	**0.83 (0.76–0.91)**	0.94 (0.88–1.01)	**1.30 (1.15–1.46)**
≥ $10.55/hr[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Nativity									
Born outside U.S.	**1.10 (1.00–1.21)**	**1.20 (1.11–1.31)**	**1.15 (1.06–1.25)**	**1.07 (1.00–1.15)**	1.03 (0.96–1.11)	**1.14 (1.07–1.22)**	**1.16 (1.07–1.21)**	**1.14 (1.08–1.22)**	1.06 (0.94–1.19)
Born in U.S.[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Language									
Spanish	0.90 (0.79–1.03)	0.98 (0.88–1.09)	**1.15 (1.06–1.25)**	**1.22 (1.13–1.31)**	**1.18 (1.10–1.27)**	**1.22 (1.14–1.30)**	1.00 (0.91–1.10)	1.04 (0.97–1.12)	**1.63 (1.46–1.81)**
English[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Education									
< High school	1.00 (0.88–1.14)	0.97 (0.86–1.09)	0.96 (0.86–1.06)	1.09 (0.99–1.21)	1.03 (0.93–1.14)	1.05 (0.96–1.16)	0.98 (0.88–1.09)	1.03 (0.95–1.12)	**1.22 (1.04–1.43)**
High school/GED	1.06 (0.94–1.18)	1.06 (0.96–1.17)	0.98 (0.89–1.07)	1.08 (0.98–1.18)	1.06 (0.97–1.16)	1.06 (0.97–1.15)	1.02 (0.94–1.12)	1.05 (0.97–1.13)	1.06 (0.92–1.23)
>High school[a]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

PR, prevalence ratio; 95% CI, 95% confidence interval.
Bold denotes a significant association ($P \leq 0.05$); N, frequency; %, percent.
*Overall discord between self-reports and walkthrough exposure ratings.