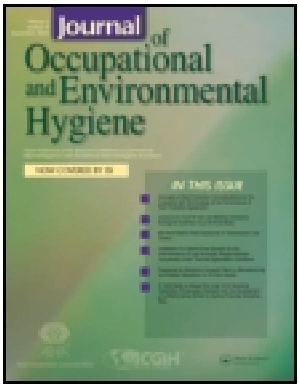

Journal of Occupational and Environmental Hygiene
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uoeh20

Inter-Rater Reliability of Assessed Prenatal Maternal Occupational Exposures to Solvents, Polycyclic Aromatic Hydrocarbons, and Heavy Metals

Carissa M. Rocheleau [a], Christina C. Lawson [a], Martha A. Waters [a], Misty J. Hein [a], Patricia A. Stewart [b], Adolfo Correa [c], Diana Echeverria [d] & Jennita Reefhuis [c]

[a] Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Cincinnati, Ohio

[b] Stewart Exposure Assessments, LLC, Arlington, Virginia

[c] Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, Atlanta, Georgia

[d] Battelle Centers for Public Health and Evaluation, Department of Environmental Health, University of Washington, Seattle, Washington

Available online: 10 Nov 2011

To cite this article: Carissa M. Rocheleau, Christina C. Lawson, Martha A. Waters, Misty J. Hein, Patricia A. Stewart, Adolfo Correa, Diana Echeverria & Jennita Reefhuis (2011): Inter-Rater Reliability of Assessed Prenatal Maternal Occupational Exposures to Solvents, Polycyclic Aromatic Hydrocarbons, and Heavy Metals, Journal of Occupational and Environmental Hygiene, 8:12, 718-728

To link to this article: http://dx.doi.org/10.1080/15459624.2011.627293

Journal of Occupational and Environmental Hygiene, 8: 718–728
ISSN: 1545-9624 print / 1545-9632 online
DOI: 10.1080/15459624.2011.627293

Inter-Rater Reliability of Assessed Prenatal Maternal Occupational Exposures to Solvents, Polycyclic Aromatic Hydrocarbons, and Heavy Metals

Carissa M. Rocheleau,[1] Christina C. Lawson,[1] Martha A. Waters,[1] Misty J. Hein,[1] Patricia A. Stewart,[2] Adolfo Correa,[3] Diana Echeverria,[4] and Jennita Reefhuis[3]

[1]Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Cincinnati, Ohio
[2]Stewart Exposure Assessments, LLC, Arlington, Virginia
[3]Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, Atlanta, Georgia
[4]Battelle Centers for Public Health and Evaluation, Department of Environmental Health, University of Washington, Seattle, Washington

Because direct measurements of past occupational exposures are rarely available in population-based case-control studies, exposure assessment of job histories by multiple expert raters is frequently used; however, the subjective nature of this method makes measuring reliability an important quality control step. We evaluated inter-rater reliability of 7729 retrospective jobs reported in the National Birth Defects Prevention Study. Jobs were classified as exposed, unexposed, or exposure unknown by two independent industrial hygienists; exposed jobs were further evaluated for intensity, frequency, and routes. Exposure prevalence ranged from 0.1–9.8%. Inter-rater reliability for exposure (yes/no), assessed by kappa coefficients, was fair to good for cadmium ($\kappa = 0.46$), chlorinated solvents ($\kappa = 0.59$), cobalt ($\kappa = 0.54$), glycol ethers ($\kappa = 0.50$), nickel compounds ($\kappa = 0.65$), oil mists ($\kappa = 0.63$), and Stoddard Solvent ($\kappa = 0.55$); PAHs ($\kappa = 0.24$) and elemental nickel ($\kappa = 0.37$) had poor agreement. After a consensus conference resolved disagreements, an additional 4962 jobs were evaluated. Inter-rater reliability improved or stayed the same for cadmium ($\kappa = 0.51$), chlorinated solvents ($\kappa = 0.81$), oil mists ($\kappa = 0.63$), PAHs ($\kappa = 0.52$), and Stoddard solvent ($\kappa = 0.92$) in the second job set. Inter-rater reliability varied by exposure agent and prevalence, demonstrating the importance of measuring reliability in studies using a multiple expert rater method of exposure assessment.

Keywords birth defects, exposure assessment, industrial hygienist

Correspondence to: Carissa M. Rocheleau, National Institute for Occupational Safety and Health, 4676 Columbia Pkwy, MS R-15, Cincinnati, OH 45226; e-mail: CRocheleau@cdc.gov.

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

INTRODUCTION

The vast majority of chemicals used in the workplace remain untested for reproductive effects in humans.(1) Among those few chemicals studied in humans, simplistic exposure estimates (based on job titles or self-report) could have biased the results. Where exposure histories must be assessed retrospectively, and thus cannot be directly measured or observed, researchers may rely on self-reported exposures, job-exposure matrices (JEMs), or expert industrial hygienist opinion to assign historical exposures. Expert opinion is costly and time-consuming but generally considered the most accurate of these methods.(2–4)

There remains, however, considerable variation within the methods used by experts to rate exposure. While some techniques can reduce measurement error,(5,6) they cannot fully eliminate it; thus determining the reliability and validity of the exposure assessment method is an important quality control step in any study using indirect methods of estimating exposures.(7) Reliability is the consistency of a rating in measuring an exposure, whereas validity is the accuracy of a rating in measuring an exposure.(8) While validity could not be evaluated in this study, it was possible to assess reliability between raters in assessing occupational exposures.

This study evaluated the assessment of occupational exposures of mothers participating in the National Birth Defects Prevention Study (NBDPS), an ongoing population-based case-control study in the United States.(9) Exposure to cadmium, chlorinated solvents, cobalt, glycol ethers, nickel, oil mists, polycyclic aromatic hydrocarbons (PAHs), and Stoddard solvent was assessed. Etiologic analyses of these data will examine associations between maternal occupational exposures and the risk of birth defects in offspring. This article describes the exposure assessment process used for the NBDPS and provides estimates of inter-rater reliability for the exposure ratings.

METHODS

Study Design and Job Information

The NBDPS is an ongoing population-based case-control study, begun in 1997, of more than 30 birth defect types. The details of the study design and methods have been described elsewhere.[9] In brief, affected children and fetuses were identified through active case ascertainment by surveillance programs in eight states (Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, and Texas). Two additional programs (North Carolina and Utah), started in 2002, were excluded from this analysis. Controls were a random sample of live births without major birth defects in the areas covered by the surveillance programs and were identified from hospital delivery logs or birth certificate files.

Mothers completed a computer-assisted telephone interview, in either English or Spanish, between 6 weeks and 2 years after the estimated date of delivery. The first occupational question read: *"The next section is a series of questions about your work experiences—paid, volunteer, or military service. This includes part-time and full-time jobs, jobs at home, and jobs on a farm or outside your home that lasted one month or more. Between (3 months before conception) and (date of infant birth) did you have a job?"* For each job, the mother was asked to provide the employer name, job title, descriptions of the company's product/service, main job activities/duties, chemicals/substances handled, and machines used on the job. Mothers also provided job start and end dates and quantitative information on the usual number of days worked per week and hours worked per day.[9]

Each job was assigned North American Industry Classification System (NAICS) and 2000 Standard Occupational Classification (SOC-2000) codes. The jobs database contained information on 7751 jobs worked from 3 months prior to conception to the end of pregnancy for 6451 mothers with estimated due dates in 1997–2001. Recruitment for the NBDPS continued while these jobs were being evaluated, thus data on an additional 4981 jobs from 4239 mothers (with estimated due dates through 2002) later became available. We evaluated these jobs separately to examine whether rater experience affected inter-rater reliability.

Exposure Ratings

Three industrial hygienists (raters) participated in the occupational exposure assessment. All raters were blinded to the case-control status of subjects. The raters' experience in industrial hygiene monitoring ranged from 17–27 years; each also had at least 10 years of experience in retrospective exposure assessment and participated in a training session prior to reviewing the job histories. During training, the raters were given definitions of the exposure variables and a sample set of 100 example jobs. Each rater independently rated the 100 example jobs, then all raters worked together to examine the rationale and assumptions behind their rating decisions, including discussing mechanisms of exposure and modifying factors. This process was intended to help raters calibrate their ratings.

Exposure agents were selected based on suspected reproductive toxicity: cadmium, chlorinated solvents as a group, cobalt, glycol ethers as a group (limited to ethylene-glycol-mono-ethyl-ethers [EGEE] and ethylene-glycol-mono-methyl-ethers [EGME]), nickel (both elemental and compound forms), oil mists, PAHs, and Stoddard solvent. Industrial PAH exposure was distinguished from exposure to PAHs from work-related second-hand smoke; only the former is reported here.

For each agent, the exposure assessment included three steps: (1) a preliminary evaluation of job codes by a single rater, (2) independent exposure evaluations of all jobs by two raters, and (3) a final consensus rating of each potentially exposed job by all raters. In the first step, an industrial hygienist assigned a preliminary determination of exposure (probably not exposed, probably exposed, or exposure unknown) to each unique NAICS and SOC-2000 code within the database of the 7751 jobs. In the second step, two raters independently reviewed each job in the data set, modifying the exposure status where appropriate, based on the specific job information provided by each mother.

For jobs rated as probably exposed in Step 2, nine characteristics were assigned: (1) whether inhalation exposure was direct, indirect, or both; (2) whether the inhalation exposure was continuous, intermittent, or both; (3) the fraction of total hours worked when exposure was direct; (4) the fraction of total hours worked when exposure was indirect; (5) the intensity of any direct inhalation exposure (on an ordinal scale from 0 to 4, Table I) during the period of direct exposure; (6) the intensity of any indirect inhalation exposure (same scale) during the period of indirect exposure; (7) the possibility of dermal exposure (possible, not likely); (8) the possibility of ingestion (possible, not likely); and (9) an estimate of confidence in the ratings based on the availability of information on similar jobs in the literature and rater experience with similar jobs (best guess, referenced in the literature, knowledgeable, familiar). Direct inhalation exposure was defined as inhalation exposure occurring to a worker that was generated by the job task the worker was performing; whereas indirect inhalation exposure refers to potential inhalation exposure to a worker who is not directly working with the agent. Airborne exposures may be present due to an adjacent process or general room air mixing rather than directly from the process, operation, or tasks the worker is directly involved in.

Quantitative interpretation for the categorical intensity scores are provided in Table I. Jobs with too little information to assign exposure were rated "exposure unknown." The third step of the exposure assessment, the consensus conference, occurred approximately 12 months after the Step 2 ratings were completed. All three raters met, reviewed, and discussed

TABLE I. Exposure Levels Associated with Each Intensity Score during the Period of Exposure, by Agent

Agent	ACGIH® TLV®[A]	Intensity Score 1	2	3	4
Cadmium (μg/m^3)[B]	10	< 1.25	2.5	3.75	> 5.0
Chlorinated solvents[C]	NA	Very low	Low-moderate	Moderate-high	Highest
Cobalt (mg/m^3)	0.02	< 0.005	0.01	0.015	>0.02
Glycol ethers (ppm)[D]	0.1 to 5	< 1.25	2.5	3.75	>5.0
Nickel - compound form (μg/m^3)[E]	0.1 to 0.2	< 0.2	2	20	>200
Nickel - elemental form (μg/m^3)	1.5	< 1.5	15	150[F]	>1500[F]
Oil mists (mg/m^3)	5	< 0.3125	0.625	1.25	>2.5
Polycyclic aromatic hydrocarbons (μg/m^3)[G]	0.2	< 0.1	1	8	>10
Stoddard solvent (ppm)	100	< 12.5	25	37.5	>50

[A]ACGIH threshold limit value. All values listed are 8-hr time-weighted averages.
[B]Inhalable fraction.
[C]NA, not applicable; quantitative exposure levels were not assigned to the intensity scores.
[D]Includes EGME (ethylene-glycol-mono-methyl-ethers, TLV = 0.1 ppm) and EGEE (ethylene-glycol-mono-ethyl-ethers, TLV = 5 ppm) only.
[E]TLV for soluble inorganic compounds is 0.1 μg/m^3; TLV for insoluble inorganic compounds is 0.2 μg/m^3.
[F]No job was assessed as having this intensity.
[G]There is no TLV for polycyclic aromatic hydrocarbons (PAHs) as a group; coal tar pitch volatiles (TLV = 0.2 μg/m^3) were used as the marker for PAHs.

ratings on (1) all jobs considered exposed by at least one rater, (2) any job belonging to the same NAICS or SOC-2000 groups as an exposed job, and (3) any job group where raters had difficulty assigning exposure. Based on the consensus discussion, jobs were re-rated according to the nine exposure characteristics and assigned final consensus ratings. The second job set of 4981 jobs was evaluated separately, allowing us to evaluate the impact of the first consensus process on reliability.

Statistical Methods

All statistical analyses were conducted using SAS 9 software (SAS Institute Inc., Cary, N.C.). Jobs rated as exposure unknown by either rater were excluded from the reliability calculations (n = 22–25). Inter-rater reliability was estimated for each agent, using the Step 2 exposure ratings (probably not exposed/probably exposed). Reliability was assessed separately for the compound and elemental forms of nickel. The kappa coefficient (κ) was used to measure inter-rater reliability; κ is a commonly used statistic for measuring inter-rater reliability for categorical data that corrects for agreement that would be expected to occur based on chance alone.(10) Values of κ less than 0.40, between 0.40 and 0.75, and greater than 0.75 are generally interpreted to represent poor, fair to good, and excellent agreement, respectively.(11) Exposure prevalence is typically very low in population-based studies, and in such cases, low κ can occur with high agreement when the data are disproportionately distributed across the agreement cells. For this reason, positive (p_{pos}) and negative (p_{neg}) specific agreement were also computed.(12)

Inter-rater reliability was also assessed for both the direct and indirect ordinal inhalation intensity scores (0–4). In these analyses, jobs rated as not exposed received an intensity score of 0. Since the simple kappa coefficient only considers exact agreement, a weighted kappa coefficient using quadratic weights was used to account for the magnitude of the differences. Inter-rater reliability for direct exposure ratings (not exposed, indirect, direct, and both indirect and direct) and continuous exposure ratings (not exposed, intermittent, continuous, both intermittent and continuous) used the simple kappa coefficient. Inter-rater reliability for the likelihood of dermal exposure combined the "not exposed" and "not likely" categories, which was compared with the "possible" category using the simple or unweighted kappa. Inter-rater reliability for the likelihood of ingestion was calculated similarly.

A weighted intensity score (I_w) was computed from the ratings of ordinal intensity and frequency as:

$$I_w = I_{direct} \times f_{direct} + I_{indirect} \times f_{indirect} + 0 \times [1 - (f_{direct} + f_{indirect})] \quad \textbf{(1)}$$

where I_{direct} is the intensity of direct exposure, f_{direct} is the fraction of time directly exposed, $I_{indirect}$ is the intensity of indirect exposure, and $f_{indirect}$ is the fraction of time indirectly exposed. The sum of the direct and indirect fractions did not exceed 0.9. Reliability of the weighted intensity score was estimated using the intra-class correlation coefficient (ICC), an aggregate measure used for continuous data.(13,14) The ICC was computed using a mixed-effects model that treated rater as a fixed effect and job as a random effect (ICC type 3,1 or consistency-type ICC). A separate model was fit for each agent using the MIXED procedure in SAS. Inter-rater reliability for the chlorinated solvents group, which had the largest number of exposed jobs, was also assessed within broad occupational classes; these classes were defined using the first two digits of the SOC-2000 code. Weighted intensity scores for the final consensus ratings (Step 3) were summarized using the median, range, and percentages.

RESULTS

Exposure prevalence was low (0.1–9.8%) for all agents but varied by agent and rater (Table II). For all agents, most jobs were rated as not exposed by both raters ($p_{neg} = 0.97$–0.999). Agreement between raters on exposed jobs was substantially lower ($p_{pos} = 0.24$–0.65), particularly for elemental nickel and PAHs. Inter-rater reliability based on the kappa coefficient was fair to good for all agents with the exception of elemental nickel ($\kappa = 0.24$) and PAHs ($\kappa = 0.37$). The highest kappa coefficients were for nickel compounds ($\kappa = 0.65$) and oil mists ($\kappa = 0.63$).

Weighted kappa scores for direct intensity ranged from a low of 0.30 (elemental nickel) to a high of 0.69 (Stoddard solvent, Table III). Inter-rater reliability for indirect intensity was generally lower than agreement for direct intensity, with the exception of oil mists (direct intensity $\kappa = 0.39$, indirect intensity $\kappa = 0.62$). Agreement for direct exposure was fair to good for all agents except glycol ethers, elemental nickel, and PAHs. Agreement for continuous exposure was generally similar to agreement for direct exposure and was fair to good for chlorinated solvents, cobalt, nickel compounds, and Stoddard solvent.

Ratings for dermal exposure were somewhat more reliable than the ingestion ratings. Agreement for the likelihood of dermal exposure ranged from 0.27 to 0.71; agreement for the likelihood of ingestion ranged from 0 to 0.73 but could not be calculated for some agents because all jobs were rated as not exposed or not likely exposed by ingestion. The ICC for the weighted intensity score, which incorporated direct and indirect intensity as well as the direct and indirect fraction, was excellent for Stoddard solvent, poor for cobalt and elemental nickel, and fair to good for the remaining agents.

We found that kappa coefficients for chlorinated solvents varied widely by broad occupational categories (Table IV). Occupations in construction and extraction had excellent agreement between raters. Occupations in architecture and engineering; arts, design, entertainment, sports, and media; farming, fishing and forestry; sales; health care support and practitioners; building and grounds cleaning and maintenance; and personal care and service had low agreement. A substantial

TABLE II. Summary of the Initial Ratings (Step 2) and Measures of Inter-Rater Reliability of Two Industrial Hygienist Raters on 7751 Jobs

Agent	No. Jobs Rated[B]	Exposure Ratings (No. Jobs): Concordant: Exposed	Concordant: Unexposed	Discordant	Assessed Exposure Prevalence, by Rater[A]: Rater A	Rater B	Rater C	Measures of Reliability: P_O^C	P_{neg}^D	P_{pos}^E	κ^F
Cadmium	7728	30	7630	68	0.6%	1.1%	—	0.991	0.996	0.47	0.46
Chlorinated solvents, as a group	7729	336	6966	427	4.4%	—	9.8%	0.945	0.970	0.61	0.59
Cobalt	7726	19	7675	32	0.6%	0.3%	—	0.996	0.998	0.54	0.54
Glycol ethers[G]	7729	11	7696	22	0.4%	0.2%	—	0.997	0.999	0.50	0.50
Nickel - compound form	7727	18	7690	19	0.3%	0.4%	—	0.998	0.999	0.65	0.65
Nickel - elemental form	7727	3	7705	19	0.2%	0.1%	—	0.998	0.999	0.24	0.24
Oil mists	7729	31	7662	36	0.6%	0.6%	—	0.995	0.998	0.63	0.63
PAHs - industrial sources	7728	77	7401	250	1.8%	3.4%	—	0.968	0.983	0.38	0.37
Stoddard solvent	7726	72	7541	113	1.7%	—	1.6%	0.985	0.993	0.56	0.55

[A]Results exclude 22–25 jobs per agent coded as unknown due to insufficient information to make an exposure determination.
[B]Two of the three raters independently assessed exposure for each agent.
[C]Observed agreement (p_o) is the proportion of jobs where the two raters agreed on the exposure status (exposed, not exposed).
[D]Negative specific agreement (p_{neg}) is the ratio of the number of jobs rated as not exposed by both raters and the mean number of jobs rated as not exposed by each rater.
[E]Positive specific agreement (p_{pos}) is the ratio of the number of jobs rated as exposed by both raters and the mean number of jobs rated as exposed by each rater.
[F]Simple (unweighted) kappa coefficient (κ) is a measure of agreement correcting for chance.
[G]Includes EGME (ethylene-glycol-mono-methyl-ethers, TLV = 0.1 ppm) and EGEE (ethylene-glycol-mono-ethyl-ethers, TLV = 5 ppm) only.

TABLE III. Measures of Inter-Rater Reliability for the Initial Ratings (Step 2) of Exposure Characteristics by Two Industrial Hygienist Raters on 7751 Jobs

		Measures of Reliability						
		Kappa Coefficient (κ)^A						ICC^B
Agent	No. Jobs Rated	Intensity Direct	Intensity Indirect	Exposure Direct	Exposure Continuous	Dermal Rating	Ingestion Rating	Weighted Intensity^C
Cadmium	7728	0.59	0.48	0.43	0.39	0.50	0.45	0.59
Chlorinated solvents	7729	0.39	0.26	0.51	0.55	0.63	−0.00	0.54
Cobalt	7726	0.45	0.44	0.53	0.53	0.63	0.48	0.24
Glycol ethers^D	7729	0.53	0.40	0.34	0.34	0.31	—^{E}	0.40
Nickel - compound form	7727	0.57	0.23	0.62	0.56	0.71	0.73	0.56
Nickel - elemental form	7727	0.30	−0.00	0.20	0.20	0.27	0.27	0.21
Oil mists	7729	0.39	0.62	0.48	0.35	0.42	0.42	0.59
Polycyclic aromatic hydrocarbons - industrial sources	7728	0.53	0.16	0.23	0.34	0.32	0.02	0.52
Stoddard solvent	7726	0.69	0.62	0.49	0.49	0.70	—^{E}	0.74

Note: Results exclude 22–25 jobs per agent coded as unknown due to insufficient information to make an exposure determination.

^{A}For direct and indirect intensity ratings (ordinal from 0 to 4), the table entry is the quadratic kappa statistic weighted by frequency (weighting discrepancies with the square of the difference between the paired assessments). The remaining kappa coefficients are simple (unweighted) coefficients. Four categories were considered for direct (not exposed, direct, indirect, both) and continuous (not exposed, continuous, intermittent, both) exposure ratings. Two categories were considered for dermal and ingestion ratings (not exposed/not likely and likely).

^{B}The intra-class correlation coefficient (ICC) was obtained via the MIXED procedure in SAS. The model considered job as a random effect, rater as a fixed effect, and specified an unstructured covariance matrix.

^{C}Weighted intensity = (direct intensity exposure score) × (fraction of time directly exposed) + (indirect intensity exposure score) × (fraction of time indirectly exposed), where intensity score was rated on an ordinal scale from 0 to 4.

^{D}Includes EGME (ethylene-glycol-mono-methyl-ethers, TLV = 0.1 ppm) and EGEE (ethylene-glycol-mono-ethyl-ethers, TLV = 5 ppm) only.

^{E}Kappa could not be computed because there was no variation in exposure (i.e., all jobs were rated as not exposed by ingestion).

difference between the two raters was observed, with Rater C considering many more jobs exposed to chlorinated solvents than Rater A.

This study was able to observe the impact of the consensus conference (Step 3) on exposure ratings. All three raters were present for the consensus conference. Based on discussions during the consensus conference, several job classes were re-evaluated for exposure. This led to rating changes for several jobs that had previously been concordant (e.g., 149 jobs initially rated "exposed" to chlorinated solvents by both Raters A and C were assigned to "not exposed" after the consensus process, Table V).

We also observed variations by agent with regard to exposure status, weighted intensity score, and the other characteristics of exposure as rated by the consensus conference (direct/indirect, continuous/intermittent, likelihood of dermal exposure, and likelihood of ingestion, Table VI). For most agents, few exposed jobs had only indirect exposure. For most agents, exposed jobs were intermittently exposed instead of continuously exposed, with the exception of oil mists (74% of exposed jobs were rated as continuously exposed). Dermal and ingestion exposure was most likely for cadmium, nickel (both forms), and PAHs.

In the second job set, reliability between the raters generally improved for agents with higher exposure prevalence (chlorinated solvents, PAHs, and Stoddard solvent) but stayed the same for some agents with relatively moderate exposure prevalence (cadmium and oil mists) (Table VII). Inter-rater reliability decreased for cobalt, glycol ethers, and nickel (both forms); these agents had the lowest exposure prevalence.

DISCUSSION

The reliability of expert exposure assessments, measured by agreement beyond chance (kappa statistic), was generally fair to good for most agents and exposure metrics, although reliability was consistently poor for elemental nickel. For all agents, raters had high negative agreement but lower positive agreement. After the consensus conference, reliability generally improved for agents with the highest exposure prevalence. Percentage agreement between each rater's initial (Step 2) rating and the final consensus (Step 3) rating varied across agents, suggesting that no one rater was systematically assigning unlikely exposures (i.e., low rater bias). Rather, each rater performed better for some agents and poorer for others, possibly reflecting their experience with each agent.

TABLE IV. Measures of Inter-Rater Reliability of the Initial Ratings (Step 2) for Chlorinated Solvents (Exposed/Not Exposed) by Two Industrial Hygienist Raters on 7729 Jobs by Major Occupational Group, Sorted in Order of Decreasing Kappa

				Raters Concordant[A]		Raters Discordant	
SOC 2000	Occupation Group	No. Jobs Rated	Kappa Coefficient (κ)	A_{NE}, C_{NE}	A_E, C_E	A_E, C_{NE}	A_{NE}, C_E
—	Unknown	21	1	19	2	0	0
25	Education, Training, and Library Occupations	544	1	542	2	0	0
15	Computer and Mathematical Occupations	115	—[B]	115	0	0	0
23	Legal Occupations	84	—[B]	84	0	0	0
33	Protective Service Occupations	53	—[B]	53	0	0	0
55	Military Specific Occupations	10	—[B]	10	0	0	0
47	Construction and Extraction Occupations	17	0.88	8	8	0	1
19	Life, Physical, and Social Science Occupations	83	0.65	57	15	0	11
11	Management Occupations	619	0.60	564	25	0	30
51	Production Occupations	370	0.55	159	124	1	86
43	Office and Administrative Support Occupations	1700	0.52	1680	7	0	13
49	Installation, Maintenance, and Repair Occupations	12	0.44	7	2	0	3
53	Transportation and Material Moving Occupations	175	0.41	142	10	0	23
17	Architecture and Engineering Occupations	34	0.39	27	2	0	5
27	Arts, Design, Entertainment, Sports, and Media Occupations	143	0.35	130	3	1	9
45	Farming, Fishing, and Forestry Occupations	116	0.31	106	2	0	8
41	Sales and Related Occupations	1056	0.31	1014	8	0	34
31	Healthcare Support Occupations	351	0.22	343	1	0	7
29	Healthcare Practitioners and Technical Occupations	551	0.18	522	3	0	26
37	Building and Grounds Cleaning and Maintenance Occupations	194	0.17	10	122	0	62
13	Business and Financial Operations Occupations	337	—[C]	335	0	0	2
21	Community and Social Services Occupations	127	—[C]	126	0	0	1
35	Food Preparation and Serving Related Occupations	648	—[C]	640	0	0	8
39	Personal Care and Service Occupations	369	—[C]	273	0	0	96
	Overall	7729	0.59	6966	336	2	425

[A]Rater A, Rater C; NE, not exposed; E, exposed.
[B]Perfect agreement. Kappa could not be computed because there was no variation in exposure (i.e., all jobs rated as not exposed).
[C]Kappa could not be computed because there was no variation in exposure for Rater A (i.e., Rater A rated all jobs as not exposed).

Variation in inter-rater reliability may also reflect the inherent difficulty of evaluating some job descriptions. Information on hygiene practices, ventilation, and engineering controls are important to determine worker exposure but was not available to raters;[(15–20)] therefore raters were forced to make assumptions about workplaces, and these assumptions may not be identical between raters. Where the potential frequency and intensity of exposure are very low, even small differences in raters' assumptions may result in large differences in overall inter-rater reliability.

Overall, direct intensity ratings were characterized by higher reliability than indirect intensity ratings, with the exception of oil mists. Dermal exposures were assessed no better than inhalation exposures but somewhat more reliably than ingestion exposures, which can be influenced by smoking behaviors—information not available to the expert raters. Differences between the raters' and the consensus estimates do not appear to follow any clear pattern. There was, however, some suggestion that the consensus meeting improved reliability for more prevalent exposures.

TABLE V. Number of Initial (Step 2) and Consensus (Step 3) Ratings by Two Industrial Hygienist Raters

	Changes from Initial to Consensus (initial→consensus) Exposure Ratings				Percentage Agreement Between Initial and Consensus Ratings for Discordant Jobs[A]		
	Concordant Jobs[B]		Discordant Jobs[B]				
	E→ NE	NE→ E	D→ E	D→ NE	Rater A	Rater B	Rater C
Cadmium	1	1	30	38	61.8	38.2	—
Chlorinated solvents	149	0	81	344	80.9	—	19.0
Cobalt	0	0	22	10	62.5	37.5	—
Glycol ethers	0	0	7	15	27.3	72.7	—
Nickel - compound form	0	0	7	12	63.2	36.8	—
Nickel - elemental form	0	0	4	15	42.1	57.9	—
Oil mists	0	0	31	5	52.8	47.2	—
Polycyclic aromatic hydrocarbons - industrial sources	3	86	170	80	46.8	53.2	—
Stoddard solvent	5	0	54	59	69.9	—	30.1

[A]E = exposed, NE = not exposed, D = discordant.
[B]Only two raters completed the Step 2 ratings, while all three raters participated in the Step 3 consensus ratings.

Comparison With Other Studies of Inter-Rater Reliability

We observed wide variation in inter-rater reliability by agent but generally within the range described by other retrospective occupational exposure studies. Inter-rater reliability for chlorinated solvents ($\kappa = 0.59$) and Stoddard solvent (0.55) in our study was higher than a previous comparison of two industrial hygienists rating all solvent exposures (as a group) within five factories where the kappa statistics for inter-estimator agreement, per factory, ranged from 0.23–0.50.[(4)]

In another community-based case-control study, unweighted kappas for pairwise inter-rater comparisons among five raters ranged from 0.02 to 0.81, depending on the agent (21 separate chemicals) and the pair of raters being compared;[(5)] their reported inter-rater reliability for PAHs ($\kappa = 0.22$) was low but similar to ours ($\kappa = 0.37$). Inter-rater reliability in our

TABLE VI. Summary Estimates of Exposure for Jobs Rated by Consensus (Step 3) as Exposed

		Weighted Intensity Score[A]					
Agent	Exposed Jobs[B]	Median	Range	Exposure Direct[C]	Exposure Continuous[D]	Dermal Likely	Ingestion Likely
Cadmium	60 (0.78%)	0.30	0.020–3.2	88%	23%	88%	88%
Chlorinated solvents	267 (3.5%)	0.90	0.020–2.7	81%	49%	59%	0%
Cobalt	41 (0.53%)	0.75	0.050–1.7	95%	15%	51%	83%
Glycol ethers	18 (0.23%)	0.95	0.10–1.9	56%	28%	67%	0%
Nickel - compound form	25 (0.32%)	0.90	0.050–3.3	88%	40%	88%	88%
Nickel - elemental form	7 (0.09%)	0.75	0.25–1.6	71%	29%	71%	71%
Oil mists	62 (0.80%)	0.90	0.10–3.6	48%	74%	29%	2%
Polycyclic aromatic hydrocarbons - industrial sources	330 (4.3%)	0.20	0.050–3.0	92%	2%	87%	85%
Stoddard solvent	121 (1.6%)	0.95	0.030–3.2	91%	21%	67%	0%

[A]Weighted exposure intensity score = (direct intensity exposure score) × (fraction of time directly exposed) + (indirect intensity exposure score) × (fraction of time indirectly exposed), where intensity score was rated on an integer scale from 0 to 4.
[B]Number (percentage) of jobs rated by consensus as exposed.
[C]Jobs were rated as direct, indirect, or both. Exposure was considered as direct here if it was rated as direct or both.
[D]Jobs were rated as continuous, intermittent, or both. Exposure was considered as continuous here if it was rated as continuous or both.

TABLE VII. Measures of Inter-Rater Reliability for Initial (Step 2) and Consensus (Step 3) Exposure Ratings by Two Industrial Hygienist Raters by Job Set

			Measures of Reliability							
			Kappa Coefficient (κ)[A]						ICC[B]	
	No. Jobs Rated		Exposure Rating		Intensity Direct		Intensity Indirect		Weighted Intensity	
Agent	**Set 1**	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**
Cadmium	7728	4959	0.46	0.51	0.59	0.42	0.48	0.47	0.59	0.45
Chlorinated solvents	7729	4959	0.59	0.81	0.39	0.67	0.26	0.67	0.54	0.54
Cobalt	7726	4953	0.54	0.35	0.45	0.38	0.44	0.24	0.24	0.36
Glycol ethers	7729	4958	0.50	0.32	0.53	0.30	0.40	0.27	0.40	0.31
Nickel - compound form	7727	4961	0.65	0.30	0.57	0.44	0.23	0.08	0.56	0.28
Nickel - elemental form	7727	4961	0.24	0.15	0.30	0.00	−0.00	0.18	0.21	0.01
Oil mists	7729	4961	0.63	0.63	0.39	0.22	0.62	0.55	0.59	0.52
Polycyclic aromatic hydrocarbons - industrial sources	7728	4960	0.37	0.52	0.53	0.65	0.16	0.25	0.52	0.53
Stoddard solvent	7726	4962	0.55	0.92	0.69	0.90	0.62	0.94	0.74	0.91

Note: Set 1 represents the first set of jobs that were rated; Set 2 represents a second set of jobs that were rated after the Set 1 consensus process was completed. Results exclude 22–25 (Set 1) and 19–28 (Set 2) jobs that were coded as unknown due to insufficient information to make an exposure determination.
[A]For direct and indirect intensities, the cell entry is the quadratic weighted kappa statistic (weighting discrepancies with the square of the difference between the paired assessments), which has been shown to be equivalent to the intra-class correlation coefficient (ICC).
[B]The intra-class correlation coefficient (ICC) was obtained via the MIXED procedure in SAS. The model considered job as a random effect, rater as a fixed effect, and specified an unstructured covariance matrix.

study was lower than observed by Siemiatycki et al.[21] for two experts rating exposures to 294 separate agents (weighted $\kappa = 0.70$). These raters, however, had participated in consensus ratings for the same jobs 4 years earlier, and their high agreement likely reflects these prior conferences. A study of formaldehyde workers demonstrated that inter-rater reliability improves with practice, and varies widely depending on agent, which is consistent with our findings.[22] Our finding that inter-rater agreement was generally higher for direct intensity than indirect intensity is consistent with a study by Correa and associates[23] ($\kappa = 0.35$ and 0.10, respectively, for direct and indirect lead exposure probability).

Both rater bias and variations in raters' familiarity with a given industry or agent can cause misclassification of exposure. In a study by Rybicki and associates,[20] discordant ratings tended to be resolved as exposed; this may reflect a tendency of the raters to under-assign exposure. Benke and associates found that the prevalence of assigned exposure varied greatly by rater (a measure of rater bias); low exposure prevalence and high rater bias tended to decrease kappas.[5]

In contrast to these studies, our study did not observe pervasive patterns indicative of rater bias across agents. None of the raters consistently assigned a higher rate of exposure than the other (Table II), and we observed no pattern of raters under-assigning exposure across agents compared with the final ratings (Table V) as Rybicki described. As Table V shows, only oil mists and PAHs were more likely to be resolved as exposed than not exposed. We did, however, observe variations in assigned prevalence of exposure within occupational classes for specific solvents (Table IV); this indicates that even in the absence of pervasive rater bias, it is important to use multiple raters.

These previous studies suggest that variations in expert ratings may lead to exposure misclassification that could bias any observed associations toward the null. Some raters' exposure assessments may be more valid than others. When the true validity of exposure assessment cannot be observed, inter-rater reliability can be assessed as a partial proxy for validity.[8] Inter-rater reliability has varied widely by both rater and agent in previous studies, ranging from poor to very good. Inter-rater reliability observed in our study is generally comparable to or slightly higher than those observed in similar studies.

Limitations and Strengths

Due to the rarity of birth defects, this study followed a case-control design and relied on information collected from respondents during computer-assisted telephone interviews. This information consisted of a work history around the time of pregnancy, which often covered more than one job or work site. Using retrospective job information precluded measuring exposures directly or even obtaining exposure measurement data from companies where mothers worked. In addition, mothers can only report exposure information they know, yet raters often need additional technical information. The collection of more detailed job histories, specific to the exposure(s) of interest, may improve the quality of occupational exposure assessments.[22]

Though typical of occupational exposures in population-based case-control studies, the low prevalence of occupational exposures in our study was a major limitation. Indeed, while negative specific agreement was very high, positive specific agreement was much lower. The expert panel method of exposure assessment reportedly works well for relatively common exposures (i.e., prevalence $> 5\%$);[5] however, in our study, the exposure prevalence ranged from 0.1% to 4.3%.

Reliability and validity both contribute to measurement error. To evaluate validity, exposure estimates must be compared with actual exposure measurements. It was not possible to evaluate the validity of the exposure estimates in this study due to the lack of a gold standard or validation data. Without validity data, we cannot estimate the extent to which odds ratios might be over- or underestimated in future exposure-response analyses. A recent review of exposure assessment methods for case-control studies, however, suggests that expert opinion has higher validity than self-reports.[2]

Since multiple ratings of the same job from the same rater were not available, intra-rater reliability could not be assessed. In addition, the assessment of inter-rater reliability was complicated by the fact that both raters assigned exposure estimates using the same preliminary determination of exposure (i.e., Step 1). Thus, the ratings in Step 2 were not completely independent. Each rater was instructed, however, to rate all jobs (not just jobs rated possibly exposed in the preliminary assessment) and had the opportunity to revise the Step 1 exposure status as part of the Step 2 assessment.

Despite these limitations, this study benefits from several unique strengths. Several steps were taken to improve the quality of job information: the study used a computer-assisted telephone interview, interviewers were extensively trained to elicit and transcribe job histories in a standardized manner, and probes were programmed into the interview to solicit complete responses. These steps may reduce variations in data quality.[24,25]

Second, subjects completed a pregnancy calendar at the start of the NBDPS interview to relate periods of their pregnancy to calendar dates. Creating a timeline may improve subjects' recall.[26] Finally, NBDPS respondents reported only their occupational history during a relatively short exposure window; this contrasts with many case-control studies that assess lifetime jobs or a long exposure period. This decreased the number of jobs (average of 1.2 compared with four or more for many studies) that subjects had to distinguish from one another and also the number of jobs to be evaluated. This decrease may have reduced estimation errors due to tedium during the rating process. Furthermore, the short duration (less than 4 years) between job end dates and questionnaire administration may have decreased recall errors. All jobs occurred in relatively recent time periods; reliability of exposure estimates decreases for more distant time periods compared with recent time periods.[27]

Previous work suggests that expert assessments can be improved by increasing the number of raters, training or calibrating the raters to one another, using raters with broad industry experience and prior retrospective rating experience, and using procedures to either develop consensus among raters or eliminate aberrant ratings.[4,5,28–32] All these techniques were used, which may have improved accuracy and reliability in this study; however, these techniques are resource-intensive and may not be feasible for all investigators.

A final, particular strength of this study is its ability to observe both the consensus process and the impact of consensus on subsequent assessments. Discrepancies in exposure assessment between raters may be based on either random variations in judgment, differences in knowledge of a particular industry or agent, or a personal (rater) bias. Rater bias would have occurred if one rater had consistently over- or under-assigned exposure across multiple agents and/or industries. If the consensus results had corrected for rater bias, one might expect to see that one rater had particularly low agreement with the consensus ratings, perhaps across all categories of exposure. Yet in Table V we saw no overall trends by rater in the changes from the initial to the consensus ratings across agents. This suggests that inter-rater disagreements were probably not due to rater bias.

During the consensus conference, the raters decided that additional information on changes in cleaning solvents over time were needed to accurately rate this exposure. Based on additional investigation by the raters, several additional jobs were rated as exposed to PAHs (Table V). This demonstrates the immense importance of outside information, such as results of monitoring studies or historical ingredient lists, to even expert raters.

After the consensus conference, the raters went on to independently rate a second set of jobs. We expected that the consensus conference might improve inter-rater reliability for the second set of jobs; this did not occur for all agents (Table VII). Reliability generally improved for the agents with the highest prevalence, stayed the same for the agents with the moderate prevalence, and declined for the agents with low prevalence in the second set. This may be due to the low prevalence of exposure overall in the population. Even a few disagreements between raters may heavily influence the kappa coefficients when exposure prevalence is low.

CONCLUSIONS

We obtained exposure estimates with fair to good agreement for cadmium, chlorinated solvents, cobalt, glycol ethers, nickel compounds, oil mists, and Stoddard solvent, but poor agreement for elemental nickel. Inter-rater reliability was poor for PAHs in the initial job set but fair to good for jobs in the second job set. In the second job set, agreement for chlorinated and Stoddard solvents also increased to excellent. PAHs, chlorinated solvents, and Stoddard solvent were also the most prevalent agents. This suggests that consensus discussions can improve inter-rater reliability as measured by kappa, particularly for exposures with higher prevalence. Because raters showed lower agreement for indirectly and intermittently exposed jobs, as well as exposure through dermal and ingestion routes, we suggest that rater training should pay special attention to these areas. Overall, this study shows that reliable retrospective assessment of maternal occupational exposures is feasible in case-control studies of birth defects.

ACKNOWLEDGMENTS

The authors would like to recognize Marianne Yencken and James Catalano of the Battelle Centers for Public Health and Evaluation for their contributions to the ratings process and interpretation. We particularly thank the 10 NBDPS centers who collected the data. This work was supported by the Centers for Disease Control and Prevention (U50/CCU 713238; U01/DD000492); and the National Institute for Occupational Safety and Health (contract 200-2000-08018).

REFERENCES

1. **Endocrine Disruptor Screening and Testing Advisory Committee:** "Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report." [Online] Available at http://www.epa.gov/endo/pubs/edspoverview/finalrpt.htm (accessed March 14, 2011).
2. **Teschke, K., A.F. Olshan, J.L. Daniels, et al.:** Occupational exposure assessment in case-control studies: Opportunities for improvement. *Occup. Environ. Med. 59*(*9*):575–93; discussion 594 (2002).
3. **Bauer, E.P., P.A. Romitti, and S.J. Reynolds:** Evaluation of reports of periconceptual occupational exposure: Maternal-assessed versus industrial hygienist-assessed exposure. *Am. J. Ind. Med. 36*(*5*):573–578 (1999).
4. **Kromhout, H., Y. Oostendorp, D. Heederik, and J.S. Boleij:** Agreement between qualitative exposure estimates and quantitative exposure measurements. *Am. J. Ind. Med. 12*(*5*):551–562 (1987).
5. **Benke, G., M. Sim, A. Forbes, and M. Salzberg:** 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 (1997).
6. **Hawkins, N.C., and J.S. Evans:** Subjective estimation of toluene exposures: A calibration study of occupational hygienists. *Appl. Ind. Hyg. J. 4*(*3*):61–68 (1989).
7. **Correa, A., W.F. Stewart, H.C. Yeh, and C. Santos-Burgoa:** Exposure measurement in case-control studies: Reported methods and recommendations. *Epidemiol. Rev. 16*(*1*):18–32 (1994).
8. **White, E., B.K. Armstrong, and R. Saracci:** *Principles of Exposure Measurement in Epidemiology*, 2nd ed. New York: Oxford University Press, 2008.
9. **Yoon, P.W., S.A. Rasmussen, M.C. Lynberg, et al.:** The national birth defects prevention study. *Public Health Rep. 116*(*Suppl 1*):32–40 (2001).
10. **Cohen, J.A.:** A coefficient of agreement for nominal scales. *Educ. Psychol. Meas. 30*:37–46 (1960).
11. **Fleiss, J.L.:** *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons, 1981.
12. **Cicchetti, D.V., and A.R. Feinstein:** High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol. 43*(*6*):551–558 (1990).
13. **Bartko, J.J.:** The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep. 19*(*1*):3–11 (1966).
14. **Fleiss, J.L., M.J. Slakter, S.L. Fischman, M.H. Park, and N.W. Chilton:** Inter-examiner reliability in caries trials. *J. Dent. Res. 58*(*2*):604–609 (1979).

15. **Burstyn, I., and K. Teschke:** Studying the determinants of exposure: A review of methods. *Am. Ind. Hyg. Assoc. J. 60*:57–72 (1999).
16. **Kromhout, H., E. Symanski, and S.M. Rappaport:** A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Ann. Occup. Hyg. 37*(*3*):253–270 (1993).
17. **Park, D., P.A. Stewart, and J.B. Coble:** Determinants of exposure to metalworking fluid aerosols: A literature review and analysis of reported measurements. *Ann. Occup. Hyg. 53*(*3*):271–288 (2009).
18. **Pronk, A., J. Coble, and P.A. Stewart:** Occupational exposure to diesel engine exhaust: A literature review. *J. Expo. Sci. Environ. Epidemiol. 19*(*5*):443–457 (2009).
19. **Rappaport, S.M.:** Assessment of long-term exposures to toxic substances in air. *Ann. Occup. Hyg. 35*(*1*):61–121 (1991).
20. **Rybicki, B.A., E.L. Peterson, C.C. Johnson, G.X. Kortsha, W.M. Cleary, and J.M. Gorell:** Intra- and inter-rater agreement in the assessment of occupational exposure to metals. *Int. J. Epidemiol. 27*(*2*):269–273 (1998).
21. **Siemiatycki, J., L. Fritschi, L. Nadon, and M. Gerin:** 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 (1997).
22. **Stewart, P.A., R. Carel, C. Schairer, and A. Blair:** Comparison of industrial hygienists' exposure evaluations for an epidemiologic study. *Scand. J. Work Environ. Health 26*(*1*):44–51 (2000).
23. **Correa, A., Y.I. Min, P.A. Stewart, et al.:** Inter-rater agreement of assessed prenatal maternal occupational exposures to lead. *Birth Defects Res. A Clin. Mol. Teratol. 76*(*11*):811–624 (2006).
24. **Stewart, P.A., and W.F. Stewart:** Occupational case-control studies: II. Recommendations for exposure assessment. *Am. J. Ind. Med. 26*(*3*):313–326 (1994).
25. **Clavel, J., D.C. Glass, S. Cordier, and D. Hemon:** Standardization in the retrospective evaluation by experts of occupational exposure to organic solvents in a population-based case-control study. *Int. J. Epidemiol. 22*(*Suppl 2*):S121–S126 (1993).
26. **Bradburn, N.M., L.J. Rips, and S.K. Shevell:** Answering autobiographical questions: The impact of memory and inference on surveys. *Science 236*(*4798*):157–161 (1987).
27. **Teschke, K., S.A. Marion, A. Ostry, et al.:** Reliability of retrospective chlorophenol exposure estimates over five decades. *Am. J. Ind. Med. 30*(*5*):616–622 (1996).
28. **McGuire, V., L.M. Nelson, T.D. Koepsell, H. Checkoway, and W.T. Longstreth Jr.:** Assessment of occupational exposures in community-based case-control studies. *Ann. Rev. Public Health 19*:35–53 (1998).
29. **Macaluso, M., E. Delzell, V. Rose, J. Perkins, and K. Oestenstad:** Inter-rater agreement in the assessment of solvent exposure at a car assembly plant. *Am. Ind. Hyg. Assoc. J. 54*:351–359 (1993).
30. **De Cock, J., H. Kromhout, D. Heederik, and J. Burema:** Experts' subjective assessment of pesticide exposure in fruit growing. *Scand. J. Work Environ. Health 22*(*6*):425–432 (1996).
31. **Walker, K.D., P. Catalano, J.K. Hammitt, and J.S. Evans:** Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene. *J. Expo. Anal. Environ Epidemiol. 13*(*1*): 1–16 (2003).
32. **Walker, K.D., J.S. Evans, and D. MacIntosh:** Use of expert judgment in exposure assessment. Part I. Characterization of personal exposure to benzene. *J. Expo. Anal. Environ. Epidemiol. 11*(*4*):308–322 (2001).