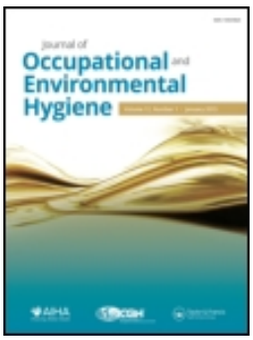

Taylor & Francis
Taylor & Francis Group

Journal of Occupational and Environmental Hygiene

ISSN: 1545-9624 (Print) 1545-9632 (Online) Journal homepage: http://www.tandfonline.com/loi/uoeh20

Using checklists and algorithms to improve qualitative exposure judgment accuracy

Susan F. Arnold, Mark Stenzel, Daniel Drolet & Gurumurthy Ramachandran

To cite this article: Susan F. Arnold, Mark Stenzel, Daniel Drolet & Gurumurthy Ramachandran (2016) Using checklists and algorithms to improve qualitative exposure judgment accuracy, Journal of Occupational and Environmental Hygiene, 13:3, 159-168, DOI: 10.1080/15459624.2015.1053892

To link to this article: http://dx.doi.org/10.1080/15459624.2015.1053892

JOURNAL OF OCCUPATIONAL AND ENVIRONMENTAL HYGIENE
2016, VOL. 13, NO. 3, 159–168
http://dx.doi.org/10.1080/15459624.2015.1053892

Taylor & Francis
Taylor & Francis Group

Using checklists and algorithms to improve qualitative exposure judgment accuracy

Susan F. Arnold[a], Mark Stenzel[b], Daniel Drolet[c], and Gurumurthy Ramachandran[a]

[a]Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, Minnesota; [b]Exposure Assessment Applications, LLC, Arlington, Virginia; [c]Institute de Recherche Robert-Sauvé en santé et an Securité du travail (IRSST), Chemical and Biological Hazards Prevention Research and Expertise Division, Montréal, Québec, Canada

ABSTRACT

Most exposure assessments are conducted without the aid of robust personal exposure data and are based instead on qualitative inputs such as education and experience, training, documentation on the process chemicals, tasks and equipment, and other information. Qualitative assessments determine whether there is any follow-up, and influence the type that occurs, such as quantitative sampling, worker training, and implementing exposure and risk management measures. Accurate qualitative exposure judgments ensure appropriate follow-up that in turn ensures appropriate exposure management. Studies suggest that qualitative judgment accuracy is low. A qualitative exposure assessment Checklist tool was developed to guide the application of a set of heuristics to aid decision making. Practicing hygienists (n = 39) and novice industrial hygienists (n = 8) were recruited for a study evaluating the influence of the Checklist on exposure judgment accuracy. Participants generated 85 pre-training judgments and 195 Checklist-guided judgments. Pre-training judgment accuracy was low (33%) and not statistically significantly different from random chance. A tendency for IHs to underestimate the true exposure was observed. Exposure judgment accuracy improved significantly ($p < 0.001$) to 63% when aided by the Checklist. Qualitative judgments guided by the Checklist tool were categorically accurate or over-estimated the true exposure by one category 70% of the time. The overall magnitude of exposure judgment precision also improved following training. Fleiss' κ, evaluating inter-rater agreement between novice assessors was fair to moderate ($\kappa = 0.39$). Cohen's weighted and unweighted κ were good to excellent for novice (0.77 and 0.80) and practicing IHs (0.73 and 0.89), respectively. Checklist judgment accuracy was similar to quantitative exposure judgment accuracy observed in studies of similar design using personal exposure measurements, suggesting that the tool could be useful in developing informed priors and further demonstrating its usefulness in producing accurate qualitative exposure judgments.

KEYWORDS
Checklist; exposure judgment; qualitative judgments

Introduction

The vast majority of assessments conducted within comprehensive exposure assessment programs are qualitative, i.e., without monitoring data. This is by design and necessity, as the number of exposure scenarios in a workplace may be in the tens or hundreds of thousands, all of which will eventually be assessed under a comprehensive program, in which conducting quantitative exposure assessments (i.e., using monitoring data with sufficient samples to support valid decision making) for every scenario is not feasible. The American Industrial Hygiene Association (AIHA) exposure assessment strategy calls for initial, qualitative assessments of exposures, relative to a reference exposure level, such as an Occupational Exposure Limit, (OEL), Emergency Planning Guideline, or Interim Exposure Limit, based on a No-Observed Adverse Effect Level (NOAEL), respectively. Industrial hygienists (IHs) assess these using a combination of their formal and informal education, professional judgment, personal experience with a given operation, and review of exposures from similar operations to determine the acceptability of exposures for managing engineering controls, medical surveillance, hazard communication, and personal protective equipment programs. Since the

CONTACT Gurumurthy Ramachandran ramac002@umn.edu. University of Minnesota, Division of Environmental Health Sciences, School of Public Health, 420 Delaware St. S.E., Minneapolis, MN 55455

Supplemental data for this article can be accessed at tandfonline.com/uoeh. AIHA and ACGIH members may also access supplementary material at http://oeh.tandfonline.com/.

type of follow-up that occurs is determined by these initial qualitative judgments, their accuracy is essential.

Research suggests qualitative exposure judgment accuracy, based on subjective professional judgment is low, not statistically different from random chance, and tends to underestimate exposures.[1,2] These findings, indicating qualitative exposure judgments are not only wrong much of the time, but tend to underestimate true exposures are deeply concerning because they lead to ineffective (failing to adequately protect workers) and inefficient (misdirecting resources) exposure assessments, in turn leading to inefficient and ineffective IH programs. Despite the urgent need for better approaches, and a body of literature from psychology, medicine, and aviation safety suggesting they may be helpful,[3–7] the influence of alternate, objective approaches to decision-making on exposure judgment accuracy has not been systematically investigated.

Simple algorithms, requiring just a few inputs have improved health outcomes of neonates,[8] reduced infection rates,[9] and increased airline safety.[4,7] These algorithms, especially useful in low validity environments, i.e., situations with little or no data, and a high degree of uncertainty, focus the decision maker on the most critical inputs, filtering out details that would otherwise distract. In the field of industrial hygiene, simple rules or heuristics, applied consistently, have been shown to improve quantitative judgment accuracy.[1]

We present a checklist (Checklist) that was developed to guide the application of a series of algorithms or heuristics, aiding qualitative exposure assessment judgments, i.e., judgments for which personal exposure measurement data is not available, so the assessment must be conducted using other inputs. The Checklist is applicable to vapor, aerosol, fiber and particulate exposure scenarios, and requires only four readily available pieces of information: the OEL, vapor pressure of the pure chemical (VP) in the case of a vapor, the observed or reported workplace control measures (ObsLC) and the required level of control (ReqLC). While the OEL and VP are truly objective, characterizing the ObsLC is more subjective and subject to interpretation by the IH. This tends to improve with clearly defined criteria coupled with examples to reduce uncertainty, and is further enhanced with diagrams and pictures of engineering controls. The ReqLC is determined as a result of a heuristic, as described later. This article discusses the application of the checklist, and its influence on qualitative exposure judgment accuracy and inter-rater reliability (IRR).

Methods

A qualitative exposure assessment Checklist was developed to guide the application of a set of heuristics developed from empirical observations that are based on physical-chemical principles to systematically improve qualitative exposure judgment accuracy and reliability. For this study, accuracy is defined as categorical agreement between the reference Exposure Control Category (ECC) and the participant's exposure judgment regarding the ECC. The ECC is the category to which the 95th percentile of the exposure distribution ($X_{0.95}$) most likely falls.[10] The boundaries of the four ECCs are presented in Table 1. Reliability is the probability that two or more assessors, evaluating the same scenario, come to the same assessment, i.e., select the same ECC. The Checklist has broad applicability and can be administered quickly, with minimal and readily available inputs. It includes three widely applicable heuristics; the first two, the Rule of 10 and the Vapor Hazard Ratio, apply to scenarios involving pure or relatively pure volatile and semi-volatile compounds. The Particulate Hazard Ratio applies to aerosol, particulate and fiber scenarios.[11] IHs using the Checklist follow these heuristics in a specific order. Although not included in this version of the tool, other heuristics addressing scenarios involving mixtures of chemicals, considering frequency and duration of exposure, quantity of agent, configuration of a vessel opening, system pressure, etc., have been developed and are being added to the next version of the Checklist. The current version is available through the Supplemental Materials which can be found online.

Table 1. AIHA Exposure Control Categories (ECC) with criteria for interpretation.

Exposure Control Category (ECC)	Criteria for Statistical Interpretation
1	$X_{0.95} \leq 0.1 \times OEL$
2	$0.1 \times OEL < X_{0.95} \leq 0.5 \times OEL$
3	$0.5 \times OEL < X_{0.95} \leq OEL$
4	$0.5 \times OEL < X_{0.95} \leq OEL$

The Rule of 10

The Rule of 10 is premised on the incremental reduction in the maximum potential airborne concentration of a volatile chemical resulting from incrementally higher levels of control. For every step change in control (through the use of engineering controls), the maximum concentration for a scenario is reduced by a factor of 10. Engineering control types and their corresponding reduction of the airborne concentrations, expressed as a fraction of the Saturated Vapor Concentration (SVC) are presented in Table 2. The SVC is calculated from the chemical's pure vapor pressure divided by the atmospheric pressure, in mm Hg.[11]

Table 2. Rule of 10 engineering control matrix.

Level of Control	Fraction of the Saturation Vapor Concentration (SVC)
Confined Space – Virtually no circulation	1/10th of Saturation
Poor – Limited Circulation	1/100th of Saturation
Good – General ~ 6 Air Changes/Hour	1/1,000th of Saturation
Capture	1/10,000th of Saturation
Containment	1/100,000th of Saturation

Vapor hazard ratio

The Vapor Hazard Ratio (VHR) is the ratio of the SVC, divided by the OEL. A VHR Scale ranging from 1–6, reflecting ranges of increasing VHRs is used to identify the ReqLC (Table 3). This is the minimum level of control deemed necessary to adequately control the exposure.[11]

Particulate hazard ratio

The Particulate Hazard Ratio (PHR), similar to the VHR, assigns a PHR Scale value ranging from 1–6. The Scale value increases as the OEL value decreases as shown in Table 4.[11]

The checklist

The Checklist (Table 5) provides a prescribed step-by-step process for applying each heuristic. The first two heuristics are appropriate for scenarios involving pure or relatively pure volatile or semi-volatile chemicals. When assessing a volatile or semi-volatile, both heuristics are used independently. If the two heuristics predict ECCs that are not consistent with one another, the highest predicted ECC is used. Using the Rule of 10, the C_{max} acts as a surrogate for the 95th percentile exposure and is compared directly to the OEL to identify the appropriate ECC. With the VHR, a decision logic is applied whereby ObsLC is compared to the ReqLC. If the ObsLC exceeds the ReqLC, then the exposure is most likely a Category 1.

Table 3. Vapor Hazard Ratio (VHR) engineering control matrix.

Vapor Hazard Ratio (VHR)	VHR Scale	Required Level of Control (ReqLC)
< 0.05	1	General Ventilation ~ 3–6 air changes/hr
0.05 to < 1	2	Good general ventilation ~ 6–12 air changes/hr. (GGV)
1 to < 25	3	GGV with capture at emission points
25 to < 500	4	Capture at points of emission with containment wherever practical
500 to < 3000	5	Containment
> 3000	6	Primary and Secondary Containment

Table 4. Particulate Hazard Ratio (PHR) engineering control matrix.

OEL Range (mg/m^3)	PHR Scale	Required Level of Control (ReqLC)
5	1	General ventilation ~ 2–4 air changes/hr
≤ 5 to 1	2	Good – General + fans ~ 4–6 air changes/hr
≤ 1 to 0.1	3	Good – General + fans ~ 6–8 air changes/hr
≤ 0.1–0.01	4	Capture
≤ 0.01–0.001	5	Containment
≤ 0.001	6	Secondary containment

If the ObsLC is equivalent to the ReqLC, then the exposure is most likely a Category 2. If the ObsLC is less stringent than the ReqLC, the exposure is most likely a Category 4 (note that these heuristics bypass Category 3). This finding was validated both empirically over many years by Stenzel, and confirmed by the exposure data corresponding to the scenarios used in our study.

The third heuristic applies to scenarios involving aerosols, fibers and particulates and was derived from the performance based exposure limits used in the pharmaceutical industry. The PHR heuristic is used in the same manner as the VHR, and the same decision logic is used.

Eliciting IH exposure judgments using the checklist

Practicing IHs (n = 39) were recruited for a study evaluating the influence of the Checklist on exposure

Table 5. The checklist - an ordered approach to applying the three heuristics.

Rule of 10	1. Select the appropriate Occupational Exposure Limit (OEL) 2. Determine the Vapor Pressure & Saturated Vapor Concentration (SVC) 3. Identify the Observed Or Reported Level of Control (ObsLC) 4. Estimate the fraction of the SVC 5. Calculate the maximum concentration (C_{max}) 6. Compare the C_{max} to the OEL 7. Determine the predicted Exposure Control Category (ECC)
Vapor Hazard Ratio (VHR)	1. Divide VP/OEL to determine VHR Score 2. Identify Required Level of Control (ReqLC) from VHR matrix 3. Compare ReqLC with ObsLC 4. Determine ECC: If ObsLC > ReqLC = 1 If ObsLC = ReqLC = 2 If ObsLC < ReqLC = 4 5. If the ECC's based on Rule of 10 & VHR differ, use the highest ECC
Particulate Hazard Ratio (PHR)	1. Identify OEL 2. Identify ReqLC from PHR matrix 3. Compare ReqLC with ObsLC 4. Determine ECC: If ObsLC > ReqLC = 1 If ObsLC = ReqLC = 2 If ObsLC < ReqLC = 4

Table 6. Exposure scenario details, showing the scenario number, agent of concern (chemical agent), the relevant OEL, the primary task or work process from which the exposure occurred (process), the number of personal exposure samples collected, from which the reference ECC was calculated (Reference ECC data set) and the corresponding reference ECC.

Scenario	Chemical Agent	OEL	Process	Number of measurements in Reference ECC data set (n)	ECC
1	Glutaraldehyde	Cal OSHA ceiling: 0.05 ppm	Neutralizing, dumping and pouring liquid	6	2
2	Mannitol	NIOSH REL: 8 hour TWA: 0.25 $\mu g/m^3$	Potent compound transfer	6	2
3	Trichloro-ethylene	ACGIH TLV TWA: 5 ppm	Spray cleaning degreaser	11	4
4	Asbestos	OSHA PEL: 8 hour TWA: 0.1 f/cc	Locomotive steamline repair	9	2
5	Isopropanol	OSHA PEL: 8 hour TWA 400 ppm	Cleaning printing presses	8	2
6	Hexavalent chromium	OSHA PEL: 8 hr. TWA: 0.5 mg/m3	Repair – welding railroad frog	8	4
7	Asbestos	OSHA PEL: 30 min. excursion limit: 1 f/cc	Bystander exposure in locomotive repair shop	29	1
8	Acetone	OSHA PEL: 8 hour TWA: 1000 ppm	Cleaning printing presses	8	1
9	Phenol	ACGIH TLV: TWA: 5 ppm	Foundry shell core – mold making	8	1
10	Quartz	ACGIH TLV TWA for α-quartz = 0.025 mg/m^3	Foundry shake out – breaking molded parts	8	4
11	Methylene chloride	OSHA PEL: 15 min. STEL: 125 ppm	Collecting a sample from a vessel	10	4

judgment accuracy. Personal determinants (experience, training and education) were collected from this group. Novice IHs (n = 8 Master's degree students in IH) were also recruited, and their personal determinants were recorded. A summary of the profile of this cohort is presented in Table SIa and SIb. Each group was assigned several exposure scenarios and asked to assess worker exposures, before and after receiving the Checklist training. Informed consent was obtained from all participants and human subject research approval for the study granted by the University of Minnesota Institutional Review Board (IRB Code 1212M25182).

Scenarios were developed from information and data voluntarily submitted by a number of companies and organizations. An Industrial Hygiene Exposure Scenario Tool (IHEST) was developed to facilitate consistent collection and reporting of exposure scenario details, determinants, and personal exposure data (available through the Supplemental Materials). Each exposure scenario was described in a two-page narrative, systematically presenting exposure related information and providing details regarding the workplace, work tasks, chemical agent, and OEL. An example of a scenario narrative is available in the Supplemental Materials. A list of the scenarios developed for this study, the agent of interest, and ECC are presented in Table 6. Quantitative personal exposure monitoring data were excluded from the narratives, and were used only to determine. the reference ECC, against which exposures were compared, was the surrogate for the "true" ECC. Reference ECCs were calculated from a minimum sample size of six personal exposure measurements to ensure a reasonable degree of confidence in these reference values. The measurements were used in the IHDA Lite software (oesh.com) in which uniform priors were assumed and the "likelihood" decision chart produced by this Bayesian Decision Analysis software was used as the Reference ECC. Specifically, the ECC with the highest probability was identified as the Reference ECC. In two separate workshops, practicing IHs were randomly assigned four scenarios from a database comprising 11 exposure scenarios: five vapor-related scenarios and six involving aerosols, fibers, and particulates. Each IH evaluated two scenarios at the beginning of the study, before training was conducted, providing data on the participant's exposure assessment proficiency from which baseline accuracy was determined. Seven very enthusiastic study participants assessed more than the two pre-training scenarios that were assigned to them, providing additional baseline exposure judgments. These were included in the baseline analysis. A 1-hr training session was conducted, explaining each of the three sections in the Checklist and providing instructions on how to apply them. A case study (Scenario 7) was used to illustrate the application of the Excel-based Checklist tool, developed specifically for the study. The Checklist tool is included in the Supplementary Materials. Following training, in addition to reassessing the two baseline scenarios, IHs evaluated two new scenarios. Judgments were expressed probabilistically, with hygienists expressing their beliefs about the true group 95th percentile belonging to each ECC, and assigning the highest probability to the ECC to which the true group 95th percentile most likely belonged. A (hypothetical) example of this probabilistic expression is illustrated in Figure 1. While participants gave their signed consent prior to participating in the study, some

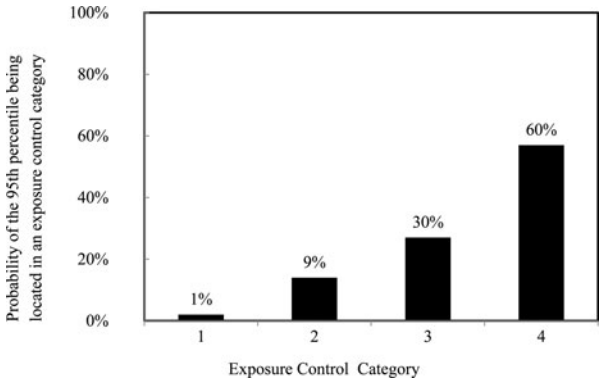

Figure 1. Bayesian Decision Chart, showing the IHs belief that the 95th percentile of the exposure distribution for a given scenario most likely belongs to Exposure Control Category (ECC) 4.

participants were either not comfortable providing all of their judgments or were unable to complete the four assigned scenarios in the time provided. A total of 61 baseline judgments were collected (5 participants provided 1 judgment = 5 judgments; 17 participants provided 2 judgments = 34 judgments; 6 participants provided 3 judgments = 18 judgments; 1 participant provided 4 judgments = 4 judgments; for a total of 61 judgments from 29 participants).

Post-training judgments were provided by 30 participants and totaled 115 participant-judgments. ($1 \times 1 + 1 \times 2 + 1 \times 3 + 26 \times 4 + 1 \times 5$ from 30 participants providing 115 post-training judgments).

Novice IHs were asked to assess three scenarios at the beginning of the study, prior to training, providing 24 baseline judgments. Following Checklist training, they were instructed to re-assess the same three scenarios, and assigned seven more new scenarios. This group was allowed to take the materials home to complete their assessments, submitting their judgments one week later. A total of 80 post-training exposure judgments were submitted.

Evaluating exposure judgments

Exposure judgment accuracy was calculated by comparing the participant's predicted ECC (ECC_{PRED}) to the reference ECC (ECC_{REF}) for each scenario:

$$Accuracy = ECC_{PRED} - ECC_{REF}. \quad (1)$$

For example, if the reference ECC indicated that the exposure most likely belonged to category 2, and the hygienist assigned the highest probability to ECC 2, the judgment was deemed categorically accurate. The difference between the number of accurate baseline judgments and the number of accurate post-training judgments (of all participants) was evaluated using χ^2- analysis.

If the scenarios had been balanced such that there was an equal distribution of scenarios belonging to each of the four ECCs, then the probability of a participant correctly making a judgment by randomly picking a category would have been 25%, the probability of under-predicting or over-predicting by one category would be 18.75%, by two categories would be 12.5%, and by three categories would be 6.25%. If the scenarios are not equally balanced among the four categories, the probabilities of being incorrect by one, two, or three categories would be different (although the probability of being correct would still be 25%). Since the scenarios were not equally distributed among the four categories, a Monte Carlo simulation with 10,000 iterations where an ECC is selected randomly for each of the ten scenarios, representing an exposure judgment for each of those scenarios. The number of times the random selection turns out to be correct across all scenarios, i.e., matches the reference ECCs is calculated, along with the number of times the random selection under- or over-predicts by one, two, or three categories. Thus, the random chance probability of being correct or incorrect by a specific number of categories was calculated.

Judgment bias was calculated for baseline and Checklist judgments from the following equation:

$$Bias_k = \text{Average Assessed } ECC_k - \text{Reference } ECC_k, \quad (2)$$

where the Average Assessed ECC = average of all predicted ECC judgments for the k^{th} scenario, and Reference ECC_j = Reference ECC for the k^{th} scenario.

For the k^{th} scenario, the standard deviation (SD) is defined as:

$$SD_k = \sqrt{\sum_{i=1}^{N} \frac{\left(ECC_{i,k} - Average\ Assessed\ ECC_k\right)^2}{N-1}}, \quad (3)$$

where $ECC_{i,k}$ is the ith participant's judgment about scenario k and N = number of participants providing judgments, and *Average Assessed* ECC_k is the Average of all Assessed ECC judgments for the kth scenario.Pairwise inter-rater reliability (IRR), a measure of agreement between two assessors making judgments about the same scenario, was calculated for each group, using Cohen's κ.[12] weighted and unweighted κ were calculated, where weighted κ reflect scores generated by assigning differential penalty weights accounting for the magnitude of disagreement between the two judgments; larger weights reflect greater disagreement. Fleiss' κ providing an aggregate κ for novice IHs (n = 8) assessing the same ten scenarios was also calculated.[13] A third IRR metric, G(q,k)[14] evaluating practicing IHs' IRR was calculated, taking into account the non-fully-crossed study design used to assign exposure scenarios. Specifically the design

produced some overlap between raters evaluating a specific scenario, but not every practicing IH assessed every scenario. This alternate measure of IRR explicitly models the variance components (Scenario main effect, Rater main effect, and Scenario-Rater interaction and residuals) and applies a multiplier, q to scale the contribution of the Rater main effect to the observed score variance. The expected value of the observed variance in judgments that have been scaled across *k* raters per scenario is calculated using Brennan's[15] formulation:

$$\sigma_Y^2 = \sigma_T^2 + q\sigma_R^2 + \frac{\sigma_{TR,e}^2}{\hat{k}}, \quad (4)$$

where σ_Y^2 = Expected observed variance, σ_T^2 = Scenario main effects, σ_R^2 = Raters main effects, $\sigma_{TR,e}^2$ = combination of rater x rate interaction and residual effects, $\hat{k}$ = harmonic mean number of rates per scenario, and q = multiplier.

These values were then used to calculate the inter-rater reliability, G(q,k):

$$G(q,k) = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \left(q\hat{\sigma}_R^2 + \frac{\hat{\sigma}_{TR,e}^2}{\hat{k}}\right)}. \quad (5)$$

Statistical analysis was conducted using R, version 3.03. The package lmer was used to calculate the variance components for G(q,k), and for Cohen's kappa, the cohen.kappa (psych) package was used.

Results

Atotal 85 baseline exposure judgments (61 + 24), described in the "Methods" section, were collected and analyzed. Baseline exposure judgment accuracy was low: 32.9% overall; 29.5% for practicing IHs, 41.7% for novice IHs, and not statistically significantly different from random chance (25.1%). Baseline judgments collected from practicing IHs were negatively biased, with 50.8% underestimating the "true" exposure by one (34.4%), two (9.8%), or three (6.6%) ECCs (Figure 2). Additional details are provided in the Supplemental Material, Tables SIIIa and SIIIb.

The post-training evaluations reported here include both re-evaluation of scenarios and evaluation of new scenarios. Since the accuracy rates are similar for the scenarios evaluated twice and the scenarios evaluated only after training/checklist use, we report only the results of the pooled evaluations. Judgment accuracy increased significantly, (χ^2 (1) = 25.36, p < 0.001) when decisions were guided by the Checklist. The percent of accurate judgments increased from pre-training baseline (28/85), to post-training (123/195). Judgments that were categorically accurate are shown in the center columns of the graph, labeled "Accurate". The reduction in the number of exposure judgments underestimating the true ECC for practicing IHs when their decisions were guided by the Checklist can also be seen in Figure 2. Judgment accuracy based on random chance, for baseline and Checklist judgments are presented in Figure 3. A detailed breakdown of judgment accuracy for novice and practicing IHs is provided in the Supplemental Materials (Tables SIa and SIb).

The negative bias observed in the baseline judgments of practicing IHs was attenuated in Checklist-guided judgments such that the absolute magnitude of bias was reduced. Precision, measured using the standard deviation, also improved for both groups, although not in all cases. The values for bias and precision are presented in Table 7 Baseline and Checklist Judgment Bias and Precision: novices and Table 8 practicing IHs.

Fleiss' κ, measuring interrater agreement of novice assessors, evaluating the same 10 scenarios was $\kappa = 0.39$, $p < 0.001$. Fleiss' κ represents an aggregate value for interrater agreement indicating in this case, that the intra-novice IH group judgment agreement was far greater than

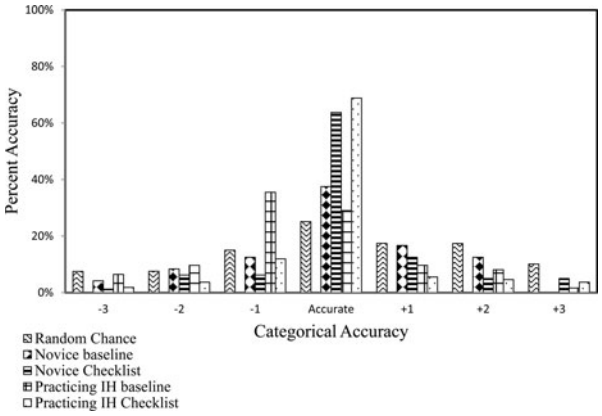

Figure 2. Categorical Judgment Accuracy, showing accuracy attributable to random chance pre-training (Baseline), post-training Checklist-guided judgment accuracy for Novices and practicing IHs.

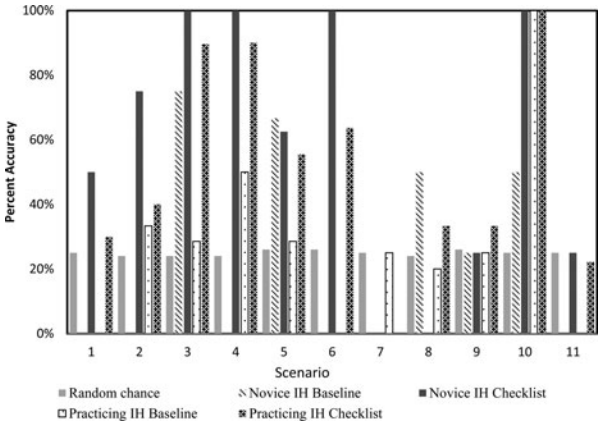

Figure 3. Baseline and Checklist based judgment accuracy for Novice IHs and practicing IHs, respectively, broken out by scenario.

Table 7a. Results from novice IHs' exposure judgments, showing bias (the difference between the average predicted ECC and reference ECC), and precision (standard deviation) for each scenario (n = 8).

	Bias		Precision	
Scenario	Baseline	Checklist	Baseline	Checklist
1	0.5	−.13	0.84	0.98
2	0.3	−.25	0.64	0.46
3	−0.5	0	0.76	0
4	−3	0	0	0
5	0.3	0.8	0.64	1.04
6	−1	0	0	0
7	0.5	1.6	0.84	0.92
8	1.3	1.3	1.2	1.16
9	−1	0	1.19	0
10	−1	−1.6	1.19	1.06

Table 7b. Results from practicing IH exposure judgments, showing bias (the difference between the average predicted ECC and reference ECC), and precision (standard deviation) for each scenario.

	n		Bias		Precision	
Scenario	Baseline	Checklist	Baseline	Checklist	Baseline	Checklist
1	2	10	0	−0.7	NA	0.48
2	3	10	1	−0.6	1.00	0.52
3	30	29	−1.1	−0.1	0.94	0.33
4	2	10	−0.5	−0.3	0.71	.95
5	7	9	−0.7	0.3	0.57	0.87
6	4	11	−1.5	−0.5	1.0	1.00
7	4	3	1.2	1.2	0.5	1.17
8	6	6	1.5	1.7	1.0	1.32
9	2	9	0	0	0	0
10	2	9	−2	1.4	NA	1.01

would be observed by chance alone ($\kappa = 0$). The pair-wise evaluation is shown in Table SIIa While there is no one widely accepted interpretation of values for Fleiss' κ, Landis and Koch (1977) suggest values of 0.2 to 0.4 represent fair agreement and values of 0.4 to 0.6 represent moderate agreement. Cohen's (1960) weighted and unweighted κ, scores were calculated for the novice IH (0.77 and 0.81) and practicing IHs (0.93 and 0.89). These values represent good to excellent agreement.[15] G(q,k), calculated for practicing IHs only, was 0.76 and would similarly indicate good agreement. The pair-wise evaluation is shown in Table SIIb.

Discussion

In disciplines where increasing complexity has led to specialization, expertise alone may not guarantee acceptable performance. Many fields, including exposure assessment are complex, with the amount of information exceeding the capacity of the pre-frontal cortex (PFC), the decision-making area of the brain. This overload makes the brain vulnerable to flaws of memory, distraction and thoroughness, inviting bias and over-confidence in our decisions.[7,16] It also leads to inconsistent summary judgments: given the same scenario, experts, rarely come to the same conclusion twice, and two experts may not come to the same conclusion.[16]

Simple algorithms typically perform better than "expert professional judgment." Expressed as checklists, they ensure that the critical steps in a process are followed in order consistently. Effective checklists contain only the essential inputs or steps, so they are not forgotten when the mind is occupied by multiple tasks. "Under conditions of complexity, not only are checklists a help, they are required for success. There must always be room for judgment, but judgment aided – and even enhanced – by procedure".(7) Checklists free the practitioner from having to focus cognitive energy on the mundane but critically important tasks, maximizing the energy available for innovation and for dealing with non-routine events.[3] First adopted by the aviation safety industry, the use of checklists led to an impressive safety track record that continues till today.[7] Checklists are also impacting medical performance measures, reducing errors and health complications associated with its inherently complex and uncertain environment.[8,9,17]

Qualitative exposure judgments, like many of the decisions in aviation safety and medicine, are inherently complex and frequently require decision-making under pressure, with minimal data. The low baseline judgment accuracy observed in this study, consistent with previous research[1,2] reflect these complexities and uncertainties, and the inability of our cognitive systems to handle them. It is interesting that the novice IHs were more accurate and less biased in their baseline judgments. Without prior experience in conducting exposure assessments or collecting exposure measurements, this cohort could not draw upon "professional judgment" and instead, applied their academic training in IH, as determined by follow-up conversations. The students had attended several lectures on IH statistics, including characteristics of the (skewed) lognormal distribution that typifies exposure measurement data, and two lectures on exposure modeling during the semester. In contrast, ~ 70% of the practicing IH group reported limited expertise in IH statistics and modeling, but >50% had extensive experience conducting exposure assessments and had considerable expertise collecting exposure measurement data. Of this group, 70% reported having conducted ≤10 exposure assessments following the AIHA Strategy. This suggests, based on our understanding of the conventional approach to exposure assessment, that the vast majority of data sets upon which the practicing IHs based their decisions and calibrated their professional judgment were small, most likely n = 0, 1, or 2. Moreover, the data were probably not

Table 8. Exposure scenario using the OEL as the benchmark, OEL = 10 ppm, GSD = 2.5.

				Probability (%) that All Indicated Measurements of Dataset Size N (N = 1,2,3, 4, or 5) Will Fall Below the OEL				
EF	GM	95th% (ppm)	Distribution < AL	N = 1	N = 2	N = 3	N = 4	N = 5
0.5	10	45.15	0.5	50	25	12.5	6.25	3.13
0.25	5.39	24.32	0.75	75	56.3	42.2	31.6	23.7
0.1	3.09	13.95	0.9	90	81	72.9	65.6	59.1
0.05	2.22	10	0.95	95	90.3	85.7	81.5	77.4
0.02	1.52	6.87	0.98	98	96	94.1	92.2	90.4

subjected to any statistical treatment, resulting in erroneous feedback.

To illustrate why these behaviors might lead hygienists to underestimate the true exposure, consider the following examples. In example 1 (Table 8), three exposure measurements collected for a scenario in which the true exceedance fraction is 25% will all fall below the OEL. With N = 3 samples collected, there is a 42% that the three samples will all be below the OEL, meaning that the hygienist may not realize the exposure is unacceptable 42% of the time. Consequently, the hygienist's professional judgment may be miscalibrated by this erroneous feedback and further reinforced each time only a few measurements are collected. In example 2, the benchmark is the Action Limit (AL) (Table SIV in the Supplemental Materials). Using the same parameters (exceedance fraction = 25%, N = 3 samples) there is a 10% probability that these three samples would all be below the AL. The feedback loop, even when using the more conservative AL may be faulty, reinforcing biased decision making.

Checklist based judgments improved qualitative judgment accuracy significantly, increasing them by a factor of 2. Novice IHs produced judgments that were just as accurate as their more experienced colleagues, suggesting that this objective approach is equally helpful for hygienists of all experience levels. Further, the level of accuracy observed in this study using the Qualitative Exposure Assessment Checklist tool was comparable to the results observed in quantitative studies.[1,2] For scenarios involving specialized engineering controls, such as the glutaraldehyde in scenario 1 (using general mechanical and local exhaust ventilation), and mannitol in scenario 2 (a clean room environment using primary and secondary containment), IHs underestimated the true exposure by 1 ECC, possibly because most of them were not familiar with this specialized work environment and engineering controls, leading to a misclassification of the level of controls used ("ObsLC"). The effectiveness of the local exhaust ventilation (LEV) was highly dependent upon work practices in the methylene chloride scenario (scenario 11), a fact that many practicing and almost all novice IHs did not take into account (nor did the training provided by the investigators suggest it). Most IHs overestimated the true exposure for bystander asbestos (scenario 7) and acetone (scenario 8) by 1 ECC, which could reflect the fact that the agents were present as mixtures, not pure or relatively pure chemicals. The algorithms in the current version of the tool do not account for the lesser magnitude of exposure, and therefore tend to overestimate in these cases. The next version of the Checklist tool will take mixtures into account.

Judgments made by practicing IHs using the Checklist were also less biased compared to when they did not use the Checklist. In fact, maximal accuracy for practicing IHs and novice IHs was achieved when the exposure judgments were based on the Checklist algorithms. This suggests that judgments guided by objective methods and based on only the critical inputs produce exposure judgments that are superior (because they are more accurate) to decisions based solely on subjective professional judgment. Intuition adds value if, and only if, it follows disciplined collection of objective information with robust scoring and analysis, i.e., if the judgment has been carefully calibrated with immediate, clear, and accurate feedback. Judgments reflect true expertise when the environment is sufficiently regular to be predictable AND the expert has had time and the opportunity to learn these regularities through practice AND the expert can express a judgment accurately in probabilistic terms. Algorithms outperform experts because experts try to think "outside the box", considering complex combinations of inputs.[3,18,19]

This may help explain why baseline exposure judgments made by novice IHs, lacking professional experience, did not exhibit the same bias towards underestimating the true exposure as was observed by "expert" practicing hygienists. It may also explain why the novice IH group's Checklist judgments were more precise, as they were less likely to try to outsmart the algorithms.

IRR helps discriminate between variance in observed judgment accuracy due to variance in the true accuracy (scores) after the variance due to measurement error between raters has been removed.[20] Since each metric is based on different assumptions, using several measures

is recommended.[21] Cohen's κ[12] tends to give lower estimates of reliability although in this study, unweighted values for practicing IHs were relatively high. This may be due to the non-fully crossed study design, resulting in rater pairs sometimes overlapping for only one or two scenarios; if the two raters agreed in their judgments, $\kappa = 1$ so Cohen's κ for the practicing IHs may be somewhat artificially inflated. The value for G(q, k) = 0.80 is very similar to the Cohen's κ (unweighted) score observed with the novice IHs' judgments that were produced using a fully-crossed design. G (q, k) explicitly accounts for the rater main effect component of the variance (σ_R^2). G (q, k) uses the q multiplier to scale the contribution of σ_R^2 to the observed judgment variance, based on the amount of overlap between the sets of raters evaluating each scenario. Inter-rater agreement was consistently good to excellent[16] and while the results should be interpreted conservatively given the study size, they suggest that the Checklist contributes to greater inter consistency in qualitative exposure judgments.

Checklist judgments may prove useful in a broader context. The Checklist provides one approach to developing accurate, informative priors in Bayesian exposure assessments which, used in conjunction with maximum likelihood estimates (calculated from exposure measurements, for as few as N = 1), produce more confident or precise posterior judgments, making Bayesian Decision Analysis more powerful. In other words, by facilitating accurate, informative priors, exposure measurements can serve a validation role, producing highly confident and accurate judgments with fewer measurements. This could and should motivate a major shift in exposure assessment practice.

There are several important limitations to this study. One is that personal exposure data were used to characterize the reference ECC, thereby suggesting quantitative measurement data is the gold standard. We defined a minimum of six personal samples to ensure a reasonably high level of confidence in these ECCs. However, when insufficient samples are collected or the data are not analyzed appropriately, using relevant statistical metrics, quantitative measurements, and conclusions drawn from such data can be highly misleading. Second, a systematic approach was used in conducting the basic characterization for each scenario, and the information collected was presented to participants logically and consistently. This may have impacted the degree to which Checklist guided judgments agreed with the reference ECCs. Lacking this kind of systematic and thorough characterization, the IH may have come to different, less accurate conclusions. Finally, as with any small study, selection bias may occur. The decision by some participants to potentially refrain from submitting their judgments which were likely less accurate, may have favorably biased the results.

Conclusions

Qualitative exposure judgments form the foundation upon which most comprehensive exposure assessments are based. Their accuracy is critical to ensuring appropriate exposure and risk assessment and risk management outcomes. The widely prevalent practice of conducting qualitative assessments based on subjective professional judgment not only fails to meet this imperative, it often leads to negatively biased exposure judgments in which the true exposure and risk is underestimated.

Judgments aided by the Checklist, on the other hand, significantly improved judgment accuracy, producing ~ 60% judgments categorically accurate, and ~ 70–74% accurate or overestimating by one ECC. This approach, applying algorithms consistently through the use of a checklist and other objective methods, offers a pathway to more accurate decisions.

To maximize the Checklist's value and impact, further evaluation against additional scenarios is recommended. Scenarios should be developed for specific industry types and task environments, and include sufficient personal sampling data to generate reasonably confident reference ECCs. These additional studies, conducted across a broader spectrum of exposure scenarios will further illuminate the bounds within which the tool contributes to accurate exposure judgments, and the limits beyond which it will not. Additional studies should also be conducted with novice assessors, to determine the generalizability of the results reported here. Lastly, continuous feedback, provided through additional research and from those using the tool is necessary to improve the Checklist and identify other useful objective approaches to improving qualitative exposure judgment accuracy.

Acknowledgments

We thank the many individuals at several organizations who advocated for and collected the information necessary to develop the scenarios and, of course, the IHs who participated in the study by assessing the scenarios.

Funding

This research was made possible by funding under NIOSH 1R01OH010093-01A2.

References

[1] **Logan, P.W.**, **G. Ramachandran**, **J.R. Mulhausen**, **and P. Hewett**: Occupational exposure decisions: Can limited

data interpretation training help improve accuracy? *Am. Occup. Hyg*. 1–14 (**2009**).

[2] **Vadali, M.**, **G. Ramachandran**, **J.R. Mulhausen**, **and S. Banerjee**: Effect of training on exposure judgment accuracy of industrial hygienist. *JOEH* 9(*4*):242–256 (**2012**).

[3] **Meehl, P.E.** *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press, **1954**.

[4] **Billings, C.E.**, **and W.D. Reynard**: Human factors in aircraft incidents: results of a 7-year study. *Aviat. Space Environ. Med*. 55(*10*): 960–965 (**1984**).

[5] **Magnusson, B.M.**, **W.J. Pugh**, **and M.S. Roberts**: Simple rules defining the potential of compounds for transdermal delivery or toxicity. *Pharmaceut. Res*. 1047–1054 (**2004**).

[6] **Lipinski, C.**, **F. Lombardo**, **B. Dominy**, **and P. Feeney**: Experimental and computational approaches to estimate solubility and permeabiliyt in drug delivery and development settings. *Adv. Drug Del. Rev*. 46:3–26 (**2001**).

[7] **Gawande, A.**: *The Checklist Manifesto: How to Get Things Right*. Vol. 13. New York: Metropolitan Books, **2010**.

[8] **Apgar, V.**, **D.A. Holaday**, **L. James**, **I.M. Weisbrot**, **and C. Berrien**: Evaluation of the newborn infant-second report. *JAMA* 168(*15*): 1985–1988 (**1958**).

[9] **Pronovost, P., D. Needham, S. Berenholtz et al.:** An intervention to decrease catheter-related bloodstream infections in the ICU. *N. Engl. J. Med*, 355:2725–2732 (**2006**).

[10] **Hewett, P.**, **P.W. Logan**, **J.R. Mulhausen**, **G. Ramachandran**, **and S. Banerjee**: Rating exposure control using Bayesian decision analysis. *J. Occup. Environ. Hyg.* 3(10):568–581 (**2006**).

[11] **Stenzel, M.**: Rules and Guidelines to Facilitate Professional Judgment. In *A Strategy for Assessing and Managing Occupational Exposures, 4th ed*., **S. Jahn**, **W.H. Bullock**, **and J.S. Ignacio** (eds.). Fairfax,VA: AIHA, **2015**.

[12] **Cohen, J.**: A coefficient of agreement for nominal scales. *Educat. Psychol. Measure.* 20(*1*): 37–46 (**1960**).

[13] **Fleiss, J.L.**: Measuring nominal scale agreement among many raters. *Psychol. Bull*. 76(*5*): 378–382 (**1971**).

[14] **Putka, D.**, **L. Huy**, **R.A. McCloy**, **and T. Diaz**: Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *J. Appl. Psychol.* 91(*5*):959–981 (**2008**).

[15] **Brennan, R.L.**: *Generalizability Theory. An Instruction Module*. New York: Springer-Verlag, **1992**.

[16] **Landis, J.R.**, **and G.G. Koch**: The measurement of observer agreement for categorical data. *Biometrics* 33(*1*):159–174 (**1977**).

[17] **Luby, S.P.**, **M. Agboatwalla**, **D.R. Feikin** et al. : Effect of handwashing on child health: a randomised controlled trial. *The Lancet* 366(*9481*):225–233 (**2005**).

[18] **Kahneman, D.**: (**2011**). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

[19] **Ashenfelter, O.**: Predicting the quality and prices of bordeaux wine. *Econ. J.* 118(*529*): F174–F184 (**2008**).

[20] **Taylor, J.**, **and D. Watkins**: Indexing reliability for condition survey data. *Conservator* 30: 49–61 (**2007**).

[21] **Vadali, M.**, **G. Ramachandran and J.R. Mulhausen**: Exposure modelling in occupational hygiene decision making. *JOEH* 6(6): 353–362 (**2009**).