



# Agreement of Experts and Non-Experts in a Desktop Exercise Evaluating Exposure to Asthmagens in the Cotton and Textile, and Other Industries

Christine Robinson<sup>1</sup>, Annemarie Money<sup>1</sup>, Raymond Agius<sup>1</sup> and Frank de Vocht<sup>1,2,\*</sup>

1.Centre for Occupational and Environmental Health, Institute of Population Health, Manchester Academic Health Sciences Centre, University of Manchester, Ellen Wilkinson Building, Oxford Road, Manchester, UK;

2.School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, UK

\*Author to whom correspondence should be addressed. Tel: +44- (0)117-928-7239; email: [frank.devocht@bristol.ac.uk](mailto:frank.devocht@bristol.ac.uk)

Submitted 10 July 2014; revised 1 September 2014; revised version accepted 1 September 2014.

## ABSTRACT

In the absence of personal exposure measurements, expert assessment, generally on a case-by-case basis, is often used to estimate exposures. However, the decision processes of individual experts when making assessments are unknown, making it difficult to assess the quality of these assessments or to compare different assessments to each other. We conducted a study in primarily the textile and cotton industries, but also in baking, metal work, and agriculture industries in which we assessed agreement between experts assessing intensity and probability of exposure in the absence of exposure measurements to compare how well their performance compares to agreement of non-desktop-based exercises reported in literature. In addition, agreement was compared with that of non-experts undertaking the same exercise, and results were further stratified to assess the impact of factors expected of affected assessments. Intraclass correlation coefficients of absolute agreement (ICC1) and consistency (ICC3) between raters were calculated. Sensitivity and specificity were estimated using a probabilistic simulation methodology developed previously. Fourteen occupational hygienists and exposure assessors with complete data for all 48 job descriptions and 8 non-experts participated. Although confidence intervals about correlation-coefficient differences are not reported, the individual limits were found to be so broad as to suggest that no statistically significant comparisons can be made. Nevertheless, preliminary observations are presented here as suggested by the computed means. Absolute agreement between expert raters was fair-good, but was somewhat better for intensity (ICC1 = 0.61) than for probability (ICC1 = 0.44) of exposure and was better for experts than non-experts. Estimated sensitivity was 0.95 and specificity 0.82 for intensity, and 0.91 and 0.78 for probability of exposure, respectively. Stratification for factors hypothesized to affect agreement did not show statistically significant differences, but consistent patterns of point estimates indicated that agreement between raters (both expert on non-experts) dropped for medium levels of information compared with little or extensive information. Inclusion of a photo or video generally improved agreement between experts but not between non-experts, whereas the year of the job description had no influence on the assessments. These data

indicate that the desktop exposure assessment exercise was of similar quality to previously reported levels of agreement. Agreements between experts' assessments were independent of the time period of the job and can be improved by inclusion of visual material. Agreement between experts as well as the non-experts does not increase with the detail of provided job information.

**KEYWORDS:** agreement; asthmagens; cotton industry; expert assessment; exposure assessment; textile industry

## INTRODUCTION

In the absence of personal exposure measurements, for example in case-control studies or retrospective cohorts, expert assessment is often used to evaluate (historical) exposure to a pre-determined set of agents. Generally, assessment is done on a case-by-case basis by an occupational hygienist, exposure assessor, occupational physician or engineer (an expert) based on information about the job history supplemented with additional detail about the work conducted or exposures encountered, if available (Teschke *et al.*, 2002). Although this method is thought to be less subject to misclassification than other subjective methods such as job-exposure matrices and self-reporting because of the experts' training and their overview of the industry, range of jobs, and important exposures (Teschke, 2003), it is much more costly, while importantly, it is also not a transparent methodology. The decision processes employed by individual experts when making assessments is a 'blackbox' of unknown and unstructured opinions (Cherrie *et al.*, 1996; Teschke *et al.*, 2002; Burstyn, 2011), and as such it is difficult to assess the quality and validity of these assessments or to compare different assessments to each other. It is argued that this is of particular concern in multicenter studies, where the quality of the exposure assessment does not only depend on the ability of each local expert but also on the feasibility of standardization of the work of the experts from the various distant settings (Mannetje *et al.*, 2003). Experts are expected to have a better vantage point than others, which should be ascribed to the idea that, they understand the mechanisms of occupational exposures (Teschke *et al.*, 2002), but in general, it remains unclear what exactly classifies an expert. It has been shown that experience results in organizing knowledge in (sophisticated) patterns and pattern recognition superior to novices or non-experts (Ericsson, 2008; Vadali *et al.*, 2009). A 2002 review of comparison of expert assessment in occupational epidemiology indicated that

agreements between experts varied widely depending on various factors including the specific exposure and their specific knowledge of particular industries and jobs, but with a median agreement of ~0.6 (Teschke *et al.*, 2002), indicating that moderate agreement between experts in general exists but also that with better knowledge of the decision patterns of experts, it should be possible to improve on this.

We conducted a study to combine quantitative estimation with qualitative methods to investigate exposure assessment as performed by experts to explore their internal decision making processes and heuristics (e.g. 'the black box'). However, as a prerequisite for interpretation of the qualitative results we assessed agreement between the experts to evaluate how well, in the absence of exposure data their judgments compare and how well their performance compares to agreement of non-desktop based exercises reported in literature. More specifically, this study focuses on exposure to asthmagens with exposure scenarios primarily in the textile and cotton industries, but we also included comparison scenarios in baking, metal work and agriculture industries. To evaluate whether experts are more consistent (better) in their assessments, even in the absence of measurement data (Logan *et al.*, 2011), than non-experts we further compared the experts' assessments with those of a group of non-experts. Finally, we used a probabilistic methodology to estimate sensitivity and specificity of the assessments in the absence of quantitative measurements data.

## METHODS

We conducted a desktop exposure assessment exercise in which a group of independent occupational hygienists and exposure assessors were asked to assess intensity and probability of exposure to an asthmagen for a series of job descriptions without the help of additional quantitative measurements. Occupational hygienists and occupational exposure assessors, identified

from the researchers' networks, British Occupational Hygiene annual conference attendance and publications and considered a random selection of the respective professional communities, were contacted. A total of 48 job descriptions were generated from available transcripts of interviews, testimonials, and observations with three different levels of detail provided (little, medium, and high). Little detail just included a job title, industry, location, and year; medium level of information additionally included a description of the tasks conducted by a particular worker; and high level of information included more task information, information about the specific environment in which the job was conducted, and details regarding protective equipment. A number of job descriptions were further supplemented by a photo or short video (for an example of each, see online [supplementary material](#) available at *Annals of Occupational Hygiene* online). Thirty job descriptions were situated in the cotton and textile industries in the UK and abroad and 18 in other industries with known exposure to asthmagens (baking, metal work, and agriculture), covering a time period from the year 1940–2012. An online tool was developed in which exposure assessors and occupational hygienists could access each job description in turn and, on the same page, could rate the job and describe their 'mental assessment procedure'. Ratings of intensity (low/none to high) and probability (very unlikely to very likely) were completed using a five-point ordinal scale (none to high). Categories 1–5 were not anchored to specific exposure levels or probabilities, nor were the assessors advised on whether ratings corresponded to any specific exposure intensity or probability distribution cut-offs; as was for example done by [Friesen et al. \(2011\)](#). Assessors could switch back and forth between job descriptions and also did not necessarily need to complete all assessments in one session. In addition to the experts, a small number of non-experts (e.g. non-occupational hygienists or exposure assessors), identified through the researchers' social networks, also completed the ratings section of the desktop exposure assessment exercise to provide a comparison.

Agreement between assessors was analyzed using random effects (assuming the participating raters were a random sample of the exposure assessment and occupational hygiene community) intraclass correlation coefficients (ICC) ([Shrout and Fleiss, 1979](#)).

These are similar to weighted  $\kappa$  with the weights at 1 minus squared distance from the diagonal (often considered the default). We calculated two different ICC values; ICC1 indicating absolute agreement between assessors, or in other words, assessors had exact same judgment or not, and ICC3 indicating consistency in that the mean differences between raters were removed and as such indicated relative ranking of jobs. Similar to previous work, we used the following cut points to signify poor agreement (ICC < 0.40), fair to good agreement (0.40–0.75), and excellent agreement (>0.75) ([Mannetje et al., 2003](#)). Stratified analyses were conducted to investigate the impact of various factors that may influence agreement. We only compared ratings of assessors who had no missing data to ensure that direct comparison were always between the same set of raters.

Furthermore, [Burstyn et al. \(2013\)](#) described a probabilistic methodology to estimate the quality of exposure assessment, measured as sensitivity and specificity, in the absence of 'the truth' (e.g. personal measurement data, for example). Sensitivity and specificity are a function of agreement (K or ICC) and the prevalence of exposure and can be estimated using a Markov Chain Monte Carlo (MCMC) approach. The specific assumption underpinning this approach is that expert assessment is better than chance, and the methodology and the R code to use this method are described in detail in [Burstyn et al. \(2013\)](#). The prevalence of exposure for all 48 job descriptions was estimated from 100% minus '% jobs without exposure'; the latter estimated by multiplying the average intensity and probability for each job description and using an a priori, but arbitrary cut-off of 5 (of a maximum of 25) to define no, or very low exposure. Additionally, a score of <10 as the cut-off to define none/low exposure was defined to lower band of the prevalence distribution [defined as a  $\beta$  distribution ([Burstyn et al., 2013](#))]. This resulted in a  $\beta$  distribution with a mode prevalence of 90% and a 95% certainty that the prevalence of exposure was >71% ( $a = 16.2$  and  $b = 2.7$ , respectively). 10,000 MCMC samples were generated and the mean and 95% credible intervals are calculated.

## RESULTS

Fifty occupational hygienists and exposure assessors from across the globe were approached, and of these,

25 (50%) agreed to participate. Of the 25 participants, 14 returned complete data. The remaining 11 providing assessments with missing values and were excluded from these comparisons. The 14 experts were from the UK, USA, Canada, Ireland, and Thailand. Additionally, eight non-experts were contacted and all agreed to participate. The occupations of the non-experts were an environmental consultant, three managers, one academic support officer, two PhD students in unrelated fields, and a retired emergency planner, seven of the non-experts provided assessments

without missing values. As such, in total, we had 1344 assessments from experts (672 [14 experts  $\times$  48 job descriptions] for intensity and 672 for probability of exposure) and 672 from non-experts (similarly: 336 for each).

Overall (Table 1) absolute agreement (ICC1) between expert assessors was fair-good for probability and intensity, although it was somewhat better for intensity (ICC1 = 0.61) than for probability (ICC1 = 0.44) of exposure. Absolute agreement of non-experts was overall lower than that of the experts

**Table 1. Intraclass correlation coefficients and 95% CIs indicating absolute agreement (ICC1) of 14 experts and for 8 non-experts assessing intensity and probability of exposure**

|                       | Absolute agreement (ICC1) |        |      |             |        |      |             |        |      |             |        |      |
|-----------------------|---------------------------|--------|------|-------------|--------|------|-------------|--------|------|-------------|--------|------|
|                       | Intensity                 |        |      |             |        |      | Probability |        |      |             |        |      |
|                       | Experts                   |        |      | Non-experts |        |      | Experts     |        |      | Non-experts |        |      |
|                       | ICC                       | 95% CI |      | ICC         | 95% CI |      | ICC         | 95% CI |      | ICC         | 95% CI |      |
|                       |                           | LL     | UL   |             | LL     | UL   |             | LL     | UL   |             | LL     | UL   |
| Overall               | 0.61                      | 0.51   | 0.72 | 0.43        | 0.27   | 0.59 | 0.44        | 0.32   | 0.57 | 0.39        | 0.25   | 0.55 |
| Level of information  |                           |        |      |             |        |      |             |        |      |             |        |      |
| Low                   | 0.68                      | 0.51   | 0.84 | 0.40        | 0.20   | 0.65 | 0.44        | 0.27   | 0.67 | 0.33        | 0.16   | 0.59 |
| Medium                | 0.43                      | 0.27   | 0.66 | 0.19        | 0.06   | 0.43 | 0.27        | 0.13   | 0.50 | 0.24        | 0.07   | 0.48 |
| High                  | 0.66                      | 0.50   | 0.83 | 0.64        | 0.43   | 0.83 | 0.55        | 0.37   | 0.76 | 0.59        | 0.36   | 0.80 |
| Inclusion video/photo |                           |        |      |             |        |      |             |        |      |             |        |      |
| Yes                   | 0.70                      | 0.53   | 0.86 | 0.46        | 0.24   | 0.72 | 0.48        | 0.29   | 0.72 | 0.38        | 0.18   | 0.65 |
| No                    | 0.58                      | 0.46   | 0.71 | 0.42        | 0.25   | 0.60 | 0.48        | 0.30   | 0.58 | 0.41        | 0.25   | 0.58 |
| Industry              |                           |        |      |             |        |      |             |        |      |             |        |      |
| Cotton/textile        | 0.72                      | 0.61   | 0.83 | 0.50        | 0.33   | 0.68 | 0.52        | 0.37   | 0.68 | 0.40        | 0.23   | 0.59 |
| Other                 | 0.41                      | 0.26   | 0.63 | 0.32        | 0.13   | 0.57 | 0.31        | 0.17   | 0.53 | 0.39        | 0.21   | 0.64 |
| Time period           |                           |        |      |             |        |      |             |        |      |             |        |      |
| 1940–1980             | 0.61                      | 0.47   | 0.76 | 0.48        | 0.28   | 0.68 | 0.43        | 0.28   | 0.62 | 0.48        | 0.3    | 0.67 |
| 1980–2012             | 0.57                      | 0.43   | 0.73 | 0.32        | 0.16   | 0.54 | 0.42        | 0.27   | 0.61 | 0.30        | 0.14   | 0.51 |
| Location              |                           |        |      |             |        |      |             |        |      |             |        |      |
| UK                    | 0.63                      | 0.52   | 0.75 | 0.47        | 0.3    | 0.64 | 0.42        | 0.29   | 0.58 | 0.44        | 0.29   | 0.61 |
| Other                 | 0.54                      | 0.33   | 0.81 | 0.26        | 0.08   | 0.59 | 0.45        | 0.24   | 0.75 | 0.26        | 0.08   | 0.59 |

LL, lower limit; UL, upper limit.

and sometimes was below the 0.40 threshold indicating fair agreement, although confidence intervals (CIs) largely overlapped. Consistency (ICC3) was, as expected, somewhat higher than absolute agreement, especially for probability, which was just above the threshold for fair agreement (0.40–0.75, Table 2). Again, as a general tendency, consistency was somewhat better for the experts than for the non-experts although the CIs largely overlap.

Further stratification by level of provided information, inclusion of a photo or video or not, time period

of the job description, and whether the job was originally in the UK or not showed some interesting patterns, although generally the CIs overlapped so firm conclusions on the importance of specific assessment factors cannot be made. Nonetheless, stratification by level of provided information indicated that, for the experts, additional information did not improve agreement; both for low and high levels of information agreement was good for intensity (ICC3 are 0.69 and 0.68, respectively) and fair-good for probability (ICC3 0.53 and 0.65, respectively). However, interestingly, there is

**Table 2. Intraclass correlation coefficients and 95% CIs indicating consistency (ICC3) of 14 experts and for 8 non-experts assessing intensity and probability of exposure**

|                       | Consistency (ICC3) |        |      |             |        |      |             |        |      |             |        |      |
|-----------------------|--------------------|--------|------|-------------|--------|------|-------------|--------|------|-------------|--------|------|
|                       | Intensity          |        |      |             |        |      | Probability |        |      |             |        |      |
|                       | Experts            |        |      | Non-experts |        |      | Experts     |        |      | Non-experts |        |      |
|                       | ICC                | 95% CI |      | ICC         | 95% CI |      | ICC         | 95% CI |      | ICC         | 95% CI |      |
|                       |                    | LL     | UL   |             | LL     | UL   |             | LL     | UL   |             | LL     | UL   |
| Overall               | 0.63               | 0.54   | 0.74 | 0.55        | 0.45   | 0.67 | 0.56        | 0.44   | 0.68 | 0.49        | 0.38   | 0.63 |
| Level of information  |                    |        |      |             |        |      |             |        |      |             |        |      |
| Low                   | 0.69               | 0.53   | 0.84 | 0.53        | 0.36   | 0.74 | 0.53        | 0.33   | 0.75 | 0.39        | 0.20   | 0.65 |
| Medium                | 0.49               | 0.32   | 0.71 | 0.44        | 0.29   | 0.67 | 0.32        | 0.14   | 0.58 | 0.36        | 0.17   | 0.61 |
| High                  | 0.68               | 0.52   | 0.84 | 0.65        | 0.48   | 0.82 | 0.71        | 0.53   | 0.87 | 0.69        | 0.51   | 0.86 |
| Inclusion video/photo |                    |        |      |             |        |      |             |        |      |             |        |      |
| Yes                   | 0.74               | 0.58   | 0.88 | 0.62        | 0.44   | 0.81 | 0.58        | 0.37   | 0.80 | 0.48        | 0.27   | 0.74 |
| No                    | 0.59               | 0.47   | 0.72 | 0.53        | 0.41   | 0.67 | 0.55        | 0.41   | 0.70 | 0.51        | 0.37   | 0.67 |
| Industry              |                    |        |      |             |        |      |             |        |      |             |        |      |
| Cotton/textile        | 0.75               | 0.65   | 0.85 | 0.63        | 0.50   | 0.76 | 0.61        | 0.47   | 0.75 | 0.54        | 0.39   | 0.69 |
| Other                 | 0.45               | 0.29   | 0.66 | 0.46        | 0.30   | 0.67 | 0.51        | 0.32   | 0.73 | 0.48        | 0.29   | 0.71 |
| Time period           |                    |        |      |             |        |      |             |        |      |             |        |      |
| 1940–1980             | 0.63               | 0.49   | 0.77 | 0.54        | 0.39   | 0.71 | 0.60        | 0.44   | 0.76 | 0.56        | 0.39   | 0.73 |
| 1980–2012             | 0.61               | 0.47   | 0.76 | 0.56        | 0.42   | 0.73 | 0.47        | 0.31   | 0.66 | 0.43        | 0.27   | 0.62 |
| Location              |                    |        |      |             |        |      |             |        |      |             |        |      |
| UK                    | 0.65               | 0.54   | 0.76 | 0.54        | 0.41   | 0.68 | 0.59        | 0.47   | 0.73 | 0.53        | 0.40   | 0.67 |
| Other                 | 0.60               | 0.39   | 0.84 | 0.60        | 0.39   | 0.84 | 0.41        | 0.18   | 0.74 | 0.41        | 0.17   | 0.73 |

LL, lower limit; UL, upper limit.

a consistent pattern for probability and intensity, also observed in the non-experts, that agreement dropped for medium levels of information. Inclusion of a photo or video generally improved agreement between experts (intensity ICC3 0.59–0.74 and probability ICC3 0.53–0.62), but had no effect for the non-experts (intensity ICC3 0.55–0.58 and probability ICC3 0.51–0.49). We also observed that in general both experts and non-experts had a better agreement for occupations in the cotton and textile industry than for the comparison industries (baking, metal work, and agriculture).

Interestingly, there was more agreement between non-experts for job descriptions in the more distant past (prior to 1980) than for those in the previous three decades (post-1980, intensity ICC3 1940–1980, 0.60 and post-1980, 0.47 and probability ICC3 1940–1980, 0.56 and post-1980, 0.43), although similarly the CIs largely overlapped. For the experts, the year of the job description had no influence on the assessments (experts ICC3 0.63, 0.61 and 0.54, 0.56, respectively). The influence of whether the job was in the UK seemed to have minimal impact on agreement (intensity, experts ICC3 UK 0.65 international 0.60, non-experts ICC3 UK 0.59 international 0.41; probability, experts ICC3 UK 0.54 international 0.60, non-experts ICC3 UK 0.53 international 0.41).

Sensitivity and specificity were estimated using probabilistic modeling [intensity (Fig. 1a) and probability (Fig. 1b)]. Lower bands of ICC1 were taken from medium level of information from Table 1 and where 0.43 ( $\pm 10\%$ ) and 0.27 ( $\pm 10\%$ ), respectively, and similarly for the upper ICC1 bands [0.69 ( $\pm 10\%$ ) and 0.65 ( $\pm 10\%$ )]. For intensity, sensitivity is 0.95 and specificity is 0.82. Similarly, for probability, the sensitivity is 0.91 and the specificity is 0.78.

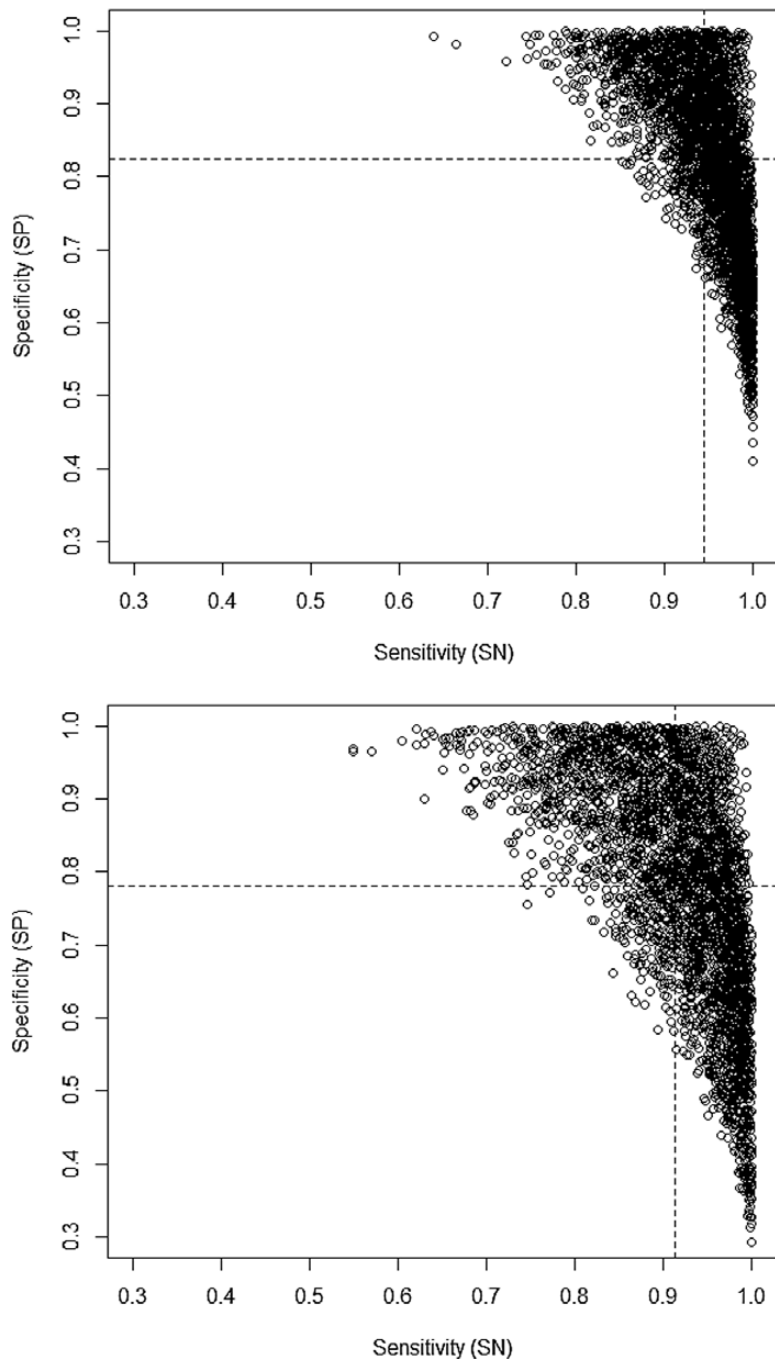
## DISCUSSION AND CONCLUSION

This manuscript describes the agreement between exposure assessment experts in a desktop exposure assessment exercise of asthmagen exposure in the cotton and textile industries and compare this to similar exercises reported in the literature. In addition, we aimed to explore the influence of selected factors *a priori* assumed to impact the assessment as well as, in the absence of quantitative measurements, whether experts still had better agreement than non-experts doing the same assessments. Although CIs about correlation-coefficient differences are not reported, the

individual limits were found to be so broad as to suggest that no statistically significant comparisons can be made. Nevertheless, preliminary observations are presented here as suggested by the computed means. Overall, agreement (ICC1 and ICC3) between expert assessors was fair to good and was generally better than the agreement of non-experts, although CIs largely overlapped. This is, in general, comparable with that found in other studies (Goldberg *et al.*, 1986; Kromhout *et al.*, 1987; Macaluso *et al.*, 1993; de Cock *et al.*, 1996; Stewart *et al.*, 2000; Mannetje *et al.*, 2003; Walker *et al.*, 2003; Steinsvag *et al.*, 2007, 2008; Chen *et al.*, 2014), although these also showed that agreement differs between agents, industries, and individual experts, which is something we could take into account to a small degree (e.g. cotton industry versus others). ICCs are lower than those found in a study on dermal exposures, but this was likely caused by the opportunity for calibration from the use of side-by-side observations (van Wendel de Joode *et al.*, 2005). Estimated sensitivity and specificity (because we had no quantitative exposure measurements available as ‘the truth’) were both high in this study. This was comparable with those previously reported for expert assessments with excellent agreement reported in a multicenter study (Mannetje *et al.*, 2003). However, simulated sensitivity and specificity in our study are likely inflated because of the high estimated prevalence of exposure in our study. Other occupational studies have reported lower sensitivity but varying specificity (Steinsvag *et al.*, 2008; Chen *et al.*, 2014), but community-based studies, with a much lower prevalence of specific exposures, indicated similar specificity but much lower ( $\sim 0.3$ ) sensitivity than reported in our study (Benke *et al.*, 1997; Tielemans *et al.*, 1999).

It is interesting to see that the pattern of agreement statistics, although not statistically significant, indicates that intensity of exposure seems more consistently assessed by the experts than probability, whereas this difference is less prominent for the non-experts. This may well indicate an effect of training that, especially for the occupational hygienists, tends to focus on recognizing (potential) high exposure situations in workplaces where exposure is evident.

Although as described in the Results, the CIs largely overlap and the effects are within 10–20% of ICCs, the patterns of the stratified analyses intuitively make sense. There was better agreement for jobs in the cotton and



1 (a) Estimated sensitivity and specificity for intensity of exposure based on 10 000 MCMC samples. (b) Estimated sensitivity and specificity for probability of exposure based on 10 000 MCMC samples

textile industries than for the other industries, which can likely be ascribed to the fact that the desktop exercise focused on this industry and experts who may have not felt comfortable assessing exposures in this particular

industry may have decided not to participate. Differences in agreement depending on the industry or exposure has been observed previously (Friesen *et al*, 2011; Schinkel *et al*, 2014). Agreement between experts was slightly

better when a video or photo was provided while this made no difference for the non-experts; although this was only observed for intensity and not probability, further indicating that training of exposure assessors is aimed at recognizing high exposure situations.

These results further indicate that agreement between assessors does not increase with the amount of information available to the rater. This has been observed previously in a number of industries (Post *et al.*, 1991; de Cock *et al.*, 1996; Stewart *et al.*, 2000; Friesen *et al.*, 2011; Chen *et al.*, 2014), but is not universally true as for general population studies, small improvements have been observed when more information became available to the raters (Tielemans *et al.*, 1999; Mannetje *et al.*, 2003). What is of specific interest in this study, however, is the clear dip in agreement for medium information (compared with little and much detail) that can be observed especially for the experts but also for the non-experts, which has not been observed before. Although it remains unknown why this occurs, this could imply that while with little information (job title, industry, location, and year), most assessors use similar heuristics to assess exposure, and with very detailed information (especially supplemented by a video or photo) available to them, they have enough information available to make good assessments, for medium information, it depends on the specific experts whether certain information is taken into account. It has also been argued that the additional information on potential determinants given to the experts may not influence the ratings because it generally confirms what they already know (de Cock *et al.*, 1996), but this would not explain the dip for medium level of information. Studies evaluating the repeatability of ratings by the same experts have shown that agreement is comparable with that between raters (median  $\kappa \sim 0.6$ , range 0.3–0.8) indicating that any internal decision making is likely not stable over time either (McGuire *et al.*, 1997; Rybicki *et al.*, 1998). Although speculation at this stage, this may imply, rather counter-intuitively, that if not all information is available (such as often happens in retrospective exposure assessment) and in the absence of available exposure measurement data, the quality of the assessment may improve if only basic information is provided and snippets of additional information withheld. Preferably, it is important to further study what partial information is required to avoid

increasing errors in assessments leading to reduced agreement between assessors.

Importantly, however, and in agreement with recommendations from others, these results suggest that differences between experts in their assessments, not surprisingly, do exist. It would therefore be advisable to use panels of assessors rather than just a single expert while also some pre-study training or calibration has been shown to be beneficial (Mannetje *et al.*, 2003; Walker *et al.*, 2003; Scholz and Hansmann, 2007; Logan *et al.*, 2009, 2011; Friesen *et al.*, 2011; Schinkel *et al.*, 2014). Specific training of heuristics has also shown to be beneficial as detailed in a study aimed at providing specific 'rules of thumb' and data interpretation training (Vadali *et al.*, 2012). Furthermore, although this and previous studies have indicated that experts can provide relative ranks of occupations reasonably well in the absence of quantitative exposure data, in order for them to provide absolute ranking, it has been shown that anchoring using exposure data is required (Tversky and Kahneman, 1974; Post *et al.*, 1991; Teschke *et al.*, 2002; Fritschi *et al.*, 2003; Logan *et al.*, 2011) to improve ranking of jobs, tasks, and individual works to improve the assessments (Kromhout *et al.*, 1987) or use triaging of different exposure assessment methods to improve ratings (Fritschi *et al.*, 2012).

Estimated sensitivity of this desktop exercise was very good (>90%) and better than that shown previously (Steinsvag *et al.*, 2008; Chen *et al.*, 2014). However, this can be ascribed to the fact that sensitivity is a function of agreement (ICC) and the prevalence of exposure, with high prevalence of exposure in these selected job descriptions. Along these lines, specificity was still good, but lower (~80%) because of the low prevalence of un-exposed jobs. It should be noted however, that there is a certain circular argument here in that un-exposed was defined based on the average of the assessments and not independently from measurements. Furthermore, the methodology is based on the assumption that the validity of the assessments is better than chance, which has been shown to be the case for relative ranking in our study and others (Post *et al.*, 1991; Friesen *et al.*, 2011). However, measurement data were not available so accuracy could not be directly assessed, whereas previous studies have indicated that the assumption of 'better than chance' may not necessarily be correct (Logan *et al.*, 2009; Vadali *et al.*, 2012; Chen *et al.*, 2014).

As mentioned previously, a main limitation of this study is that exposure measurements are not available. As such, 'the truth' against which to compare the assessments is not known. Although agreement between assessors as such is desirable, and in this study has been interpreted as 'better assessments', this is not the same as the validity of the assessments; the latter ultimately being what the assessment was aimed to do. Just to illustrate this, if all non-experts would have all consistently rated all jobs as having a high probability of exposure this would have resulted in excellent agreement; even though we know this not to be correct in terms of validity.

Furthermore, confidence in each assessment, such as was done previously (Mannetje *et al.*, 2003; Scholz and Hansmann, 2007) was not included in this study, but in retrospect would have been useful and may have provided some further insight in the observed differences.

A strength of this study is that although exposure measurements were not available, all job descriptions were based on information available from other studies, transcripts and interviews with (ex-) workers, and as such described actual exposure circumstances in the respective job/time period combinations. Moreover, these analyses have been based on a large number of assessments that enable relatively robust statistics. Also the fact that these assessments were done independently without assessors discussing any of the job descriptions or even engaging in some form of benchmarking, are further advantages of this study design.

In conclusion, these data indicate that the desktop exposure assessment exercise was of similar quality to retrospective exposure assessment by experts in previous epidemiological studies in the occupational environment, and further that experts' agreements were somewhat better than that of non-experts. Further work should be undertaken to determine what information, as a minimum, is required to avoid the observed dip in agreement between assessors for future occupational epidemiology studies. Ultimately, this will enable development of accurate, real-life exposure assessment rules and decision trees for automated assessment, preferably within a Bayesian framework (Vadali *et al.*, 2009) and may be very useful within the REACH framework (Schinkel *et al.*, 2014). Further identification of heuristics that lead to biases will be useful for continuing education.

## SUPPLEMENTARY DATA

Supplementary data can be found at <http://annhyg.oxfordjournals.org/>.

## FUNDING

British Cotton Growing Association: Work People's Collection Fund.

## ACKNOWLEDGEMENTS

The authors would like to thank all occupational hygienists and exposure assessors from across the globe who agreed to participate in this study. The authors would further like to thank Dr Michele Abendstern (The University of Manchester) for providing the testimonials of and interviews with (ex) cotton and textile workers that were used to generate some of the job descriptions.

## REFERENCES

- Benke G, Sim M, Forbes A *et al.* (1997) Retrospective assessment of occupational exposure to chemicals in community-based studies: validity and repeatability of industrial hygiene panel ratings. *Int J Epidemiol*; 26: 635–42.
- Burstyn I. (2011) The ghost of methods past: exposure assessment versus job-exposure matrix studies. *Occup Environ Med*; 68: 2–3.
- Burstyn I, de Vocht F, Gustafson P. (2013) What do measures of agreement ( $\kappa$ ) tell us about quality of exposure assessment? Theoretical analysis and numerical simulation. *BMJ Open*; 3: e003952.
- Chen YC, Coble JB, Deziel NC *et al.* (2014) Reliability and validity of expert assessment based on airborne and urinary measures of nickel and chromium exposure in the electroplating industry. *J Expo Sci Environ Epidemiol*. Epub ahead of print. doi:10.1038/jes.2014.22.
- Cherrie JW, Schneider T, Spankie S *et al.* (1996) A new method for structured, subjective assessments of past concentrations. *Occup Hyg*; 3: 75–83.
- de Cock J, Kromhout H, Heederik D *et al.* (1996) Experts' subjective assessment of pesticide exposure in fruit growing. *Scand J Work Environ Health*; 22: 425–32.
- Ericsson KA. (2008) Deliberate practice and acquisition of expert performance: a general overview. *Acad Emerg Med*; 15: 988–94.
- Friesen MC, Coble JB, Katki HA *et al.* (2011) Validity and reliability of exposure assessors' ratings of exposure intensity by type of occupational questionnaire and type of rater. *Ann Occup Hyg*; 55: 601–11.
- Fritschi L, Nadon L, Benke G *et al.* (2003) Validation of expert assessment of occupational exposures. *Am J Ind Med*; 43: 519–22.

- Fritschi L, Sadkowsky T, Benke GP *et al.* (2012) Triaging jobs in a community-based case-control study to increase efficiency of the expert occupational assessment method. *Ann Occup Hyg*; 56: 458–65.
- Goldberg MS, Siemiatycki J, Gérin M. (1986) Inter-rater agreement in assessing occupational exposure in a case-control study. *Br J Ind Med*; 43: 667–76.
- Kromhout H, Oostendorp Y, Heederik D *et al.* (1987) Agreement between qualitative exposure estimates and quantitative exposure measurements. *Am J Ind Med*; 12: 551–62.
- Logan P, Ramachandran G, Mulhausen J *et al.* (2009) Occupational exposure decisions: can limited data interpretation training help improve accuracy? *Ann Occup Hyg*; 53: 311–24.
- Logan PW, Ramachandran G, Mulhausen JR *et al.* (2011) Desktop study of occupational exposure judgments: do education and experience influence accuracy? *J Occup Environ Hyg*; 8: 746–58.
- Macaluso M, Delzell E, Rose V *et al.* (1993) Inter-rater agreement in the assessment of solvent exposure at a car assembly plant. *Am Ind Hyg Assoc J*; 54: 351–9.
- Mannetje A, Fevotte J, Fletcher T *et al.* (2003) Assessing exposure misclassification by expert assessment in multicenter occupational studies. *Epidemiology*; 14: 585–92.
- McGuire V, Longstreth WT Jr, Nelson LM *et al.* (1997) Occupational exposures and amyotrophic lateral sclerosis. A population-based case-control study. *Am J Epidemiol*; 145: 1076–88.
- Post W, Kromhout H, Heederik D. (1991) Semiquantitative estimates of exposure to methylene chloride and styrene: the influence of quantitative exposure data. *Appl Occup Environ Hyg*; 6: 197–204.
- Rybicki BA, Peterson EL, Johnson CC *et al.* (1998) Intra- and inter-rater agreement in the assessment of occupational exposure to metals. *Int J Epidemiol*; 27: 269–73.
- Schinkel J, Fransman W, McDonnell PE *et al.* (2014) Reliability of the advanced REACH tool (ART). *Ann Occup Hyg*; 58: 450–68.
- Scholz RW, Hansmann R. (2007) Combining experts' risk judgments on technology performance of phytoremediation: self-confidence ratings, averaging procedures, and formative consensus building. *Risk Anal*; 27: 225–40.
- Shrout PE, Fleiss JL. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*; 86: 420–8.
- Steinsvåg K, Bråtveit M, Moen B *et al.* (2008) Expert assessment of exposure to carcinogens in Norway's offshore petroleum industry. *J Expo Sci Environ Epidemiol*; 18: 175–82.
- Steinsvåg K, Bråtveit M, Moen BE *et al.* (2007) Inter-rater agreement in the assessment of exposure to carcinogens in the offshore petroleum industry. *Occup Environ Med*; 64: 582–8.
- Stewart PA, Carel R, Schairer C *et al.* (2000) Comparison of industrial hygienists' exposure evaluations for an epidemiologic study. *Scand J Work Environ Health*; 26: 44–51.
- Teschke K. (2003) Expert assessment of occupational and environmental exposures. In Nieuwenhuijsen MJ, editor. *Exposure assessment in occupational and environmental epidemiology*. Oxford, UK: Oxford University Press. pp. 283.
- Teschke K, Olshan AF, Daniels JL *et al.* (2002) Occupational exposure assessment in case-control studies: opportunities for improvement. *Occup Environ Med*; 59: 575–93; discussion 594.
- Tielemans E, Heederik D, Burdorf A *et al.* (1999) Assessment of occupational exposures in a general population: comparison of different methods. *Occup Environ Med*; 56: 145–51.
- Tversky A, Kahneman D. (1974) Judgment under uncertainty: heuristics and biases. *Science*; 185: 1124–31.
- Vadali M, Ramachandran G, Mulhausen J. (2009) Exposure modeling in occupational hygiene decision making. *J Occup Environ Hyg*; 6: 353–62.
- Vadali M, Ramachandran G, Mulhausen JR *et al.* (2012) Effect of training on exposure judgment accuracy of industrial hygienists. *J Occup Environ Hyg*; 9: 242–56.
- van Wendel de Joode B, van Hemmen JJ, Meijster T *et al.* (2005) Reliability of a semi-quantitative method for dermal exposure assessment (DREAM). *J Expo Anal Environ Epidemiol*; 15: 111–20.
- Walker KD, Catalano P, Hammitt JK *et al.* (2003) Use of expert judgment in exposure assessment: part 2. Calibration of expert judgments about personal exposures to benzene. *J Expo Anal Environ Epidemiol*; 13: 1–16.

**Agreement of experts and non-experts in a desktop exercise evaluating exposure to asthmagens in the cotton and textile, and other industries**

Christine Robinson<sup>1</sup>, Annemarie Money<sup>1</sup>, Raymond Agius<sup>1</sup>, Frank de Vocht<sup>1,2</sup>

<sup>1</sup> Centre for Occupational and Environmental Health, Institute of Population Health, Manchester Academic Health Sciences Centre, The University of Manchester, Manchester, UK.

<sup>2</sup> School of Social and Community Medicine, University of Bristol, Bristol, UK.

Corresponding author: Dr Frank de Vocht. School of Social and Community Medicine, University of Bristol. Canynge Hall, 39 Whatley Road, Bristol, UK, BS8 2PS. email: frank.devocht@bristol.ac.uk

***Online Supplementary Material***

## Example – Low level of information

### **Job description 1.**

**Cotton industry (Assessment of exposure to allergen: cotton dust)**

*In 2002, I worked in the tailoring industry as a seamstress in central London.*

### **Job description 36.**

**Railroad industry (assessment of exposure to: welding fumes)**

*In 1995, I was a CO<sub>2</sub> welder in an auto repair garage in Birmingham, West Midlands.*

## Example – Medium level of information

### **Job description 6.**

#### **Cotton industry (Assessment of exposure to allergen: cotton dust)**

*In 1990 I was working as a spinner in the ring room of Manor Mill, Chadderton in Oldham. I had to look after the bobbins on the frame, make sure that when the bobbins filled up they were replaced by empty ones and piece up. We had to keep the frames clean. On Saturday mornings we'd take all the wheels out of the frame and clean them. You had to do the rollers and steel bars when it was moving. We used to stop for half an hour on Friday while we did the frame ends, the rest was done with them running.*

## Example – Higher level of information

### **Job description 31.**

#### **Baking industry (Assessment of exposure to: wheat flour)**

*(2011) I started my own cupcake business last year. It's been amazingly successful and I now have a shop and employ 6 people. I work 16 hours a day every day of the week. I am always thinking and talking about my business. I can count on one hand the amount of times I've done something 'social' in the last six months that didn't pertain to my business.*



*I am forever scouting the internet for new trends in baking to anticipate. I send Twitter and Facebook updates at all hours of the night. I answer around 100 emails every single day. I am forever looking for new products to bring in and new directions to take the business in.*

*I manage the finances, the menu, the recipes, the stock, the payroll, the sales, the events etc. This is just the tip of the iceberg I haven't even mentioned the hours and hours my team spends actually baking.*

*You don't have time to pursue new markets, shops or ways to retail your cakes, because you are too busy making cakes. So, I had to hire help. The passion, love and skill only makes you a good baker. I have to be a good business person too, but I still invent every recipe we bake and I still control which of those recipes we bake every day.*

*We wear headscarves and aprons but don't have any other personal protective equipment. The machinery we use is very new and the premises recently refurbished to suit our needs and so the ventilation and environmental controls, heating etc. are excellent.*