Ann. occup. Hyg., Vol. 48, No. 8, pp. 655–661, 2004

Published by Oxford University Press
doi:10.1093/annhyg/meh075

Commentary

Principal Component Analysis is a Powerful Instrument in Occupational Hygiene Inquiries

IGOR BURSTYN

Department of Public Health Sciences, Faculty of Medicine & Dentistry, The University of Alberta, 13-103 Clinical Sciences Building, Edmonton, AB, T6G 2G3, Canada

Received 10 May 2004; in final form 21 August 2004; published online 26 October 2004

Several investigators have successfully used principal component analysis (PCA) in interpreting occupational hygiene data. However, traditional textbooks in occupational hygiene provide no guidance for the application and interpretation of PCA. In this article I briefly review the basics of PCA (for those not statistically inclined), provide some guidelines for performing PCA (and designing studies that use the power of PCA), illustrate its application in understanding exposure to mixtures and the characterization of 'peak exposure', and highlight other benefits that occupational hygienists stand to gain by including PCA in their 'statistical toolkit'. I hope that this article will promote greater use and understanding of a data analysis approach that has long been helping investigators outside the field of occupational hygiene to unravel the structure behind the complex relationships among multiple correlated variables.

Keywords: factor analysis; latent variable; manifest variable; multivariate statistics; tutorial

INTRODUCTION

The greater emphasis on data and empirical evidence has increased the importance of statistics in the profession of occupational hygiene. This has been evident with the increased use, in occupational hygiene studies of recent years, of the analysis of variance (Rappaport, 1991; Kromhout *et al*., 1993) and mixed-effects models (Samuels *et al*., 1985; Symanski *et al*., 1996; Nylander-French *et al*., 1999; Rappaport *et al*., 1999; Burstyn *et al*., 2000; Peretz *et al*., 2002; Vermeulen *et al*., 2004). It is only a matter of time before the use of multivariable statistical techniques becomes the norm in understanding the complexities of workplace exposures. These statistical techniques help investigators uncover patterns within data that arise in situations where many aspects of a phenomenon of interest can be observed and are potentially interdependent. This short review aims at facilitating a broader acceptance of one multivariable technique among occupational hygienists. It focuses on principal component analysis (PCA), which was recently used effectively in advancing the solutions to various problems in occupational hygiene (Burstyn *et al*., 2000, 2002a,b; Burstyn and Kromhout, 2002; Vermeulen *et al*., 2004). This statistical technique is illustrated in papers recently published in the *Annals of Occupational Hygiene* (Preller *et al*., 2004, Meijster *et al*., 2004). It should be noted at the outset that the use of PCA is not new to occupational hygiene and exposure assessment (Pio *et al*., 1989; Sahl *et al*., 1994; Beaton *et al*., 1998; Villeneuve *et al*., 1998; Frenich *et al*., 2002), but only recently has it emerged in mainstream occupational hygiene journals.

Consider the following problem that Meijster *et al*. (2004) faced in characterizing the determinants of exposure in a small Dutch factory that produced plastic tapes. Numerous chemicals from several processes were emitted into the air during tape winding. The goal of the occupational hygienists was to identify the impact of a new experimental process on personal exposures prior to its introduction into full-scale production. How could they design a study that can relate exposures to different mixtures to specific sources and processes in this setting? This problem has a relatively straightforward solution when

Tel: +1-780-492-3240; fax: +1-780-492-9677;
e-mail: iburstyn@ualberta.ca

viewed within the PCA paradigm, which postulates that each type of tape winding (source) emits distinct mixtures of interrelated chemicals. Thus, the problem reduces to identifying sets of chemicals that both characterize each source and discriminate among sources. The details of how this and other occupational hygiene problems were resolved will be presented at the end of the article. However, it is worth noting that Meijster *et al*. (2004) succeeded in attributing different chemical mixtures to specific sources in a timely and cost-effective manner.

The rest of this article is structured as follows: first, the theoretical foundation of PCA and some practical guidelines for conducting PCA will be briefly presented, followed by two examples of the application of the technique in occupational hygiene. The review concludes with speculation about potential future applications of PCA in occupational hygiene.

THEORETICAL BACKGROUND

As a nonstatistician, I will now attempt to explain PCA so that other persons with limited training in statistics can understand it. Karl Pearson developed the mathematical foundation of PCA in 1901 (Pearson, 1901). Its application flourished in the social sciences, where it was used to quantify and identify phenomena that could not be measured directly, such as 'health-related quality of life' (Clark *et al*., 1997) or 'sources of stress' (Beaton *et al*., 1998). PCA allows the identification of groups of variables that are interrelated via phenomena that cannot be directly observed. This is accomplished by assuming that any observed (*manifest*) variables are correlated with a small number of underlying phenomena, which cannot be measured directly (*latent* variables). Thus, if variable A is related to B, B is related to C and A is related to C it may well be that A, B and C have something in common. In statistical parlance, the observed correlation matrix is used to make inferences about the identities of any latent variables. Therefore, PCA is merely an automated and systematic examination of correlations among manifest variables, aimed at identifying underlying latent principal components. I am certain that many readers have gone through this procedure with pen and paper without resorting to additional statistical modeling. However, for large correlation matrices the task can be daunting without the aid of PCA.

PCA identifies independent factors that explain the maximum amount of mutual correlation. First, it draws a line that has a minimum possible distance from all the data points from several variables standardized to zero mean (e.g. A, B, C), using the least-sum-of-squares technique. This line represents the first principal component (PC1). Next, PCA draws a line perpendicular to it (ensuring independence), crossing the first line at the point of the greatest concentration of data. This second line is oriented in such a way as to minimize the remaining sum of squares (i.e. explaining the greatest amount of variation not explained by PC1) and becomes the second principal component (PC2). The process is repeated until the number of principal components equals the number of variables. The drawing of sequential perpendicular lines is easy to visualize with three variables (three-dimensional space), but it can still be performed algebraically in *n*-dimensional space.

The extent to which a particular principal component explains multiple correlations is measured by its *eigenvalue*. At the onset of PCA, we assume that there are as many latent variables (principal components) as there are manifest variables, with each being assigned an eigenvalue of 1 (corresponding to the common variability it explains). In PCA, we observe that some manifest variables 'cluster' around particular lines representing principal components, and increase eigenvalues for these components at the expense of the eigenvalues of other principal components. Thus, the variability explained by a given principal component is equal to its eigenvalue divided by the number of manifest variables in the analysis.

A detailed explanation of how PCA is conducted can be found in several statistical textbooks (Harman, 1976; Kleinbaum *et al*., 1988). The graph in Fig. 1 (and explanation) should help the reader to develop some intuition about how principal components are selected and how their eigenvalues are quantified. In the example in this figure, we assume that there are only two variables in PCA (A and B). Thus, variables A and B are each associated with eigenvalues of 1 (total sum of eigenvalues = 2). The data points tend to cluster around the line chosen to represent PC1. The

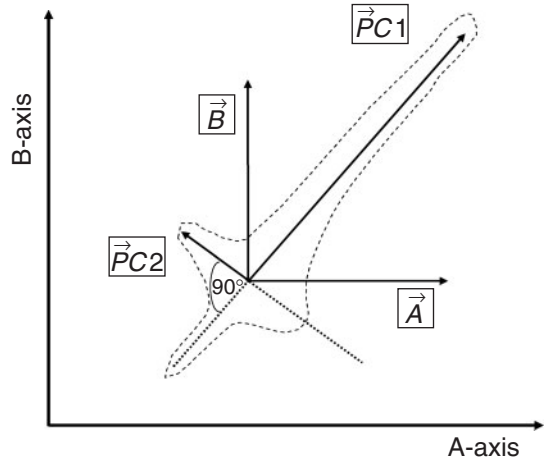

Fig. 1. Graphic representation of PCA with two manifest variables (A and B) that leads to identification of two principal components (PC1 and PC2); the dotted line defines the boundary of data 'cloud'.

length of the PC1 vector, from the center of the data cloud to its edge along PC1, represents the eigenvalue of PC1. This length is greater (>1) than those of the vectors associated with each variable individually (=1). The 'bulge' at the centre of the data cloud is not adequately accounted for by PC1. Therefore, a line perpendicular to PC1 at the 'bulge' is likely to represent PC2, and can be expected, in this case, to have an eigenvalue (length) <1. This is necessarily so because the total variability represented by two variables is equal to 2 (sum of eigenvalues of variables A and B) and PC1 has eigenvalues >1 (i.e. explains variability common to variables A and B). Thus, only PC1 explains the variability associated with more than one variable in the analysis, and therefore only it should be considered in interpretation of this data. In summary, we have started with two vectors of unit length representing the variability due to the manifest variables in the analysis, and have transformed them into two vectors in such a way that one of them captures most of the common variability associated with the two manifest variables, and thus may represent a latent variable that A and B have in common.

Another interpretation of the example in Fig. 1 is that PCA transformed the coordinate system based on A and B into one based on PC1 and PC2 in such a way that each datum is now characterized by its relationship to the latent variables (PC1 and PC2), rather than the manifest variables (A and B). Therefore, a numerical score can represent every principal component in terms of manifest variables as a distance from the origin along the appropriate principal component axis in the new coordinate system. This score is a weighted sum of the manifest variables, with weights proportional (though this is not strictly correct) to the strength of the relationship between the latent and manifest variables. We calculate the principal component score (sPCj) by multiplying each set of observed variables $X_i \in \{X_1, X_2, \ldots, X_p\}$ with the jth principal component's *eigenvector* (a column of p weights $w_{(j)i}$ that defines the direction of PCj relative to manifest variables X_i) [Note that the vectors PC1 and PC2 in the Fig. 1 represent a product of eigenvectors and their eigenvalues (direction $*$ length)].[1]

$$\text{sPC}j = w_{(j)1} \times X_1 + w_{(j)2} \times X_2 + \cdots + w_{(j)p} \times X_p$$

Principal component scores can be used either as measures of latent variables in statistical analysis in order to identify factors that explain their variability (i.e. as dependent variables) or as a measure of factors that act as explanatory (independent) variables. The association of principal components with manifest variables, measured by *loading* (correlation between principal component scores and manifest variables), can also be examined qualitatively in order to name or interpret a principal component.

An example may help illustrate the transformation of the coordinate system and computation of the principal component scores. In a simplistic situation of PCA with two variables A and B, all variance between A and B could be explained by PC1. In other words, variables A and B could be perfectly linearly correlated with points (A, B) falling on a line with a slope of 1. This would result in (i) loadings of both A and B on PC1 to be equal to 1, the eigenvalue of PC1 being equal to 2; and (ii) loadings of both A and B on PC2 to be equal to 0, the eigenvalue of PC2 being equal to 0. The principal component scores would be computed as sPC1 = 1 $*$ A + 1 $*$ B and sPC2 = 0 $*$ A + 0 $*$ B, transforming the (A, B) coordinate system into the (A + B, 0) coordinate system. A plausible interpretation of this result (if A and B are concentrations in the air of chemicals emitted from an industrial processes) would be to say that the two chemicals come from one source and the best way to estimate activity of this source is by a simple sum of concentrations of chemicals A and B.

PCA can be performed with most statistical packages (e.g. PROC PRINCOMP in SAS).

PCA FOR BEGINNERS

For those readers who are convinced that PCA is a useful technique to adopt, I wish to give an overview of how one may wish to incorporate PCA into the protocol of an occupational hygiene investigation. This overview is very rudimentary, but it should be sufficient to carry out an informative analysis.

Ideally, the use of PCA should be considered in the design of the study. Because PCA exploits observed relationships among many manifest variables to make inferences about latent variables, the choice of which manifest variables to record is crucial. The manifest variables should be selected in such a way that they reflect different aspects of the underlying phenomenon of interest. Good examples of this are the study of exposure to 4–6-ring polycyclic aromatic hydrocarbons (PAHs) among asphalt workers, which used measurements to 27 PAHs from the same 414 samples (Burstyn *et al*., 2000), and the study of exposure to hydrocarbons in shoe factories, which measured 16 different hydrocarbons in all 60 personal samples from one factory (Vermeulen *et al*., 2004). In terms of the number of observations per manifest variable, there are no easy power calculations that can help determine how many manifest variables to have, since PCA will produce a result even if the number of manifest variables (e.g. individual chemicals measured in a sample) exceeds the number of observations (e.g. particle filters collected in the course of monitoring exposure). However, it is

generally recommended to have three to five observations per manifest variable in PCA. Thus, the study of Vermeulen *et al*. (2004) appears to have adequate power to produce reliable results in PCA (60/16 = 3.75, >3). In PCA, the number of manifest variables (and their choice) is probably more important than the number of observations. Nonetheless, standard considerations of sampling strategy (e.g. random versus selective) and the stability of estimates (statisticians always prefer large numbers of observations) apply to a study that intends to employ PCA as a tool to explore the data and test hypotheses. According to various rules of thumb, 20–100 observations can be sufficient for PCA.

Leaving the mechanics of PCA to software, for the purpose of this review I will still mention that (i) the data set should be set up in the same way as it would be for computing a correlation matrix: one observation per row and multiple columns for each manifest variable, and (ii) the assumptions that apply to estimating correlations must be met by the manifest variables (e.g. normality, linearity).

PCA should be first carried out without restricting the number of retained principal components. This allows empirical selection of the number of principal components that should be extracted when PCA is performed the second time. There are two common ways to select principal components worthy of interpretation (and confirmatory PCA). First, one can select all the principal components that explain more common variability in correlation than a single variable. This would lead to a selection of all principal components with eigenvalues >1. However, some principal components will explain very little common correlation, while others (if manifest variables were correctly selected) will account for a lion's share of it. Second, a scree plot (i.e. principal component number versus eigenvalue) can help determine how many of the principal components with eigenvalues >1 should be retained. As the name of the plot implies, the place where the slope of the 'scree' begins to flatten out marks the location of the last principal component that makes a significant contribution to explaining multiple correlation (i.e. the next principal component will only make a marginal contribution to the cumulative multiple correlation explained). Of course, each investigator should also be guided in this choice by a specific hypothesis about the possible number of latent variables that principal components represent.

Once confirmatory PCA has been performed (restricting the number of principal components extracted), one can start naming the resulting principal components. Recall that the relationship between manifest variables and principal components (and thus latent variables) can be represented by the correlation of principal component scores and manifest variables. For example, these loadings can indicate that a particular aspect of peak exposure is strongly correlated with the time-weighted average (TWA) exposure (loading 0.97), providing a strong suggestion that this aspect of peak exposure is related to 'peak intensity' (Preller *et al*., 2004) (Note that the factor analysis used by Preller *et al*. is very similar in its basis and interpretation to PCA, a special case of factor analysis.) This is the creative part of statistical modeling, which calls upon each investigator's imagination and general knowledge to determine what each principal component represents. If the hypothesis was clearly stated at the outset, this task should be relatively simple (Preller *et al*., 2004). In complex situations, plots of principal components against explanatory variables or modeling of principal components' scores can be necessary for interpretation (Burstyn and Kromhout, 2002; Meijster *et al*., 2004; Vermeulen *et al*., 2004).

In some cases, the grouping of manifest variables (i.e. the identity of principal components) will change within subsets of the data. It is always 'safe' to stratify data in order to see if the same groupings of manifest variables emerge in each stratum.

Negative correlations among variables and negative loadings do not cause any specific concerns in PCA. In the interpretation of PCA, a negative loading simply means that a certain characteristic is lacking in a latent variable associated with the given principal component. For example in the study of hydrocarbon exposure among commercial painters, PC2 had negative loadings for ethylbenzene and xylene, but positive loadings for toluene, *n*-hexane and *n*-decane (Burstyn and Kromhout, 2002). A more detailed analysis revealed that paints rich in toluene, *n*-hexane and *n*-decane, but poor in ethylbenzene and xylene, were associated with spray- and house-painting with water-based paints (Burstyn and Kromhout, 2002).

ILLUSTRATIVE APPLICATIONS

I will now show how PCA has been used in characterizing exposure to mixtures and the identification of distinct features of peak exposures, two important problems in occupational hygiene.

Understanding mixtures

One of the fundamental problems in occupational hygiene is that exposures to pure substances rarely occur. More commonly, a workplace yields a complex mixture of chemicals that are potentially hazardous to workers. For example, numerous compounds are emitted during tape-winding operations studied by Meijster *et al*., (2004). It is impossible to measure all the individual compounds in such emissions for both practical and financial reasons. Two solutions exist to this problem. One is to measure representative constituents of the mixture that are deemed to be

toxicologically relevant. The other is to assume that the composition of the emissions is constant and to use a chemically nonspecific method to characterize them (e.g. total volatile organic compounds). The first method can be wasteful if the environmental concentrations of all individual constituents are highly correlated. However, the second method can be based on incorrect assumptions and can thus miss important changes in the composition of the mixture, which do not influence total amount of emissions. An occupational hygienist familiar with PCA must realize that it is not necessary to measure all components of mixtures all the time in order to study the determinants of exposure to mixtures (Burstyn *et al*., 2000, 2002b; Burstyn and Kromhout, 2002; Meijster *et al*., 2004; Vermeulen *et al*., 2004).

Thus, Meijster *et al*. (2004) measured 23 constituents of a mixture emitted during the tape-winding processes (118 samples). For each sample they determined multiple chemicals representative of the range of compounds likely to be present in the mixture. In a PCA that explained 67% of the multiple correlation by three principal components, Meijster *et al*. (2004) observed that the hydrocarbon emissions consisted of three 'independent' mixtures:

(i) short-chained aliphatic hydrocarbons (C_2–C_6),
(ii) larger hydrocarbons (C_9–C_{11}) and some cyclic hydrocarbons and
(iii) larger aromatic hydrocarbons (xylene and toluene) and hexanes (normal and cyclic).

Regression analyses of principal component scores showed that emission of short-chained aliphatic hydrocarbons was associated with cleaning activities and the use of epoxy resins. The other two mixtures showed a strong association with the type of tape used in the new tape-winding process that was being evaluated. As a result, the investigators were able to detect a new type of mixture, which was emitted because of this modification of the workplace. The study was successful in meeting the challenge of disentangling the contribution of different simultaneously present processes to the overall exposure. The keys to this success lay in the study design (a careful selection of chemicals to monitor in such a way as to represent potential sources and source-oriented sampling) and the use of appropriate statistical tools (PCA and the use of principal component scores as dependent variables in regression analyses aimed at identification of determinants of exposure). One of the most encouraging features of the study by Meijster *et al*. (2004) is that it was carried out to resolve a factory-specific occupational hygiene problem, thereby emphasizing that PCA (and its underlying philosophy) can be of great benefit to occupational hygienists who are trying to get the most information possible from their exposure monitoring budgets.

In another example, it was demonstrated that benzo(a)pyrene can be used as a proxy for exposure to 4–6-ring PAHs among asphalt pavers (Burstyn *et al*., 2000). This enabled the investigators to use data more efficiently, developing statistical models for quantitative assessment of exposure to 4–6-ring PAHs, which are proving to be essential tools in discovering associations between exposure to PAHs and adverse health effects among asphalt pavers (Burstyn *et al*., 2003a,b). Furthermore, the results of PCA in this study suggested that in future studies of exposure to bitumen fumes during asphalt paving it is sufficient to restrict PAHs determined from personal samples to (i) one PAH with molecular weight <228 g/mol, (ii) benzo(a)pyrene and (iii) naphthalene (Burstyn *et al*., 2002b) in order to fully characterize the mixture [Unless asphalt is modified with components that have a chemistry distinct from that of bitumen and coal tar, such as sulfur-containing crumb rubber, in which case one may wish to repeat the analysis of the composition of the mixture in a pilot study that incorporates sulfur-substituted hydrocarbons among manifest variables (i.e. monitored compounds)]. Such knowledge can result in more efficient allocation of the resources available to an occupational hygienist, for example by monitoring more different people on different days instead of analyzing samples for a large number of chemicals.

Characterizing peak exposures

Characterization of peak exposures in occupational hygiene is a problem that has long eluded an empirical solution. Numerous theoretical arguments were advanced in favor of different measures of 'peaks' (Blair and Stewart, 1990; Kennedy *et al*., 1991; Morrow *et al*., 1991; Salisbury *et al*., 1991; Wegman and Einsen, 1992; Chan-Yeung *et al*., 1994; Kumagai and Matsunaga, 1995; Nieuwenhuijsen *et al*., 1995; Ott *et al*., 2002), but Preller *et al*. (2004) were the first to propose and use a method that empirically identified distinguishable (and independent) aspects of peak exposure that were different from task-specific TWA exposures. The notion of the independence of different aspects of peak exposure is crucial here because if peak exposures produce the same ranking of subjects in a health-effects study as the use of TWAs, it would be impossible to decide which aspect of the exposure was responsible for the health effects. Preller *et al*. (2004) were able to measure 13 different aspects of peak exposure (representing intensity, frequency and peak duration) in 27 personal real-time continuous samples of exposure to solvents. There were three statistically independent sources of correlation among the metrics of the peak exposure, explaining 87% of the multiple correlation. The first factor reflected the intensity of peak exposure; it was also associated with the TWA exposure. The second

and third factors were related to measures of variability (in frequency and intensity) and duration of peaks, respectively. It was also shown that averaging time in the definition of the peaks was immaterial in the characterization of peaks in the studied situations. Although the particular findings of Preller *et al*. (2004) cannot be generalized to other workplaces and exposures, the approach adopted by the investigators is the only way toward empirically characterizing 'exposure to peaks'. Thus, the advent of real-time exposure monitoring equipment and the application of PCA have contributed equally to setting the stage for a rational resolution of the debate around the importance and characterization of peak exposures in workplaces.

FUTURE APPLICATIONS

Many problems in occupational hygiene require the application of PCA and related multivariable statistical techniques (e.g. factor, discriminant and cluster analyses; structural equation modeling). The identification of task patterns in studies of the determinants of exposure is one such application: it may well be that certain groupings of tasks can be used to devise more uniformly exposed groups (Frenich *et al*., 2002). If one wishes to know what type of exposure control approach is most effective (e.g. engineering or administrative controls, or some combination of them), the types (i.e. broad classes) of exposure controls can be considered as latent variables in a study design. This is important because occupational hygienists have little *documented* empirical evidence of the general types of exposure controls that are effective. Thus, the recent analysis of exposure trends in the Dutch rubber manufacturing industry uncovered some surprises, indicating that the relationship between the input of occupational hygienists and the reduction of exposures may follow a more complex mechanism than the direct modification of workplaces (Vermeulen *et al*., 2000). PCA has a natural application in the identification of psychosocial exposures (Beaton *et al*., 1998) and psychosocial determinants of exposure. The latter topic is largely unexplored, but it may yield important clues as to sources of large differences in exposures among people apparently doing the same job (Kromhout *et al*., 1993; Kromhout and Vermeulen, 2001). The list of potential applications, quite naturally, cannot be exhaustive. Multivariable statistical methods, like PCA, can be of great help in increasing the extent to which the practice of occupational hygiene is based on the solid foundation of rigorously documented empirical evidence.

Acknowledgement—Drs Trevor Ogden and Hans Kromhout encouraged me to produce this manuscript. My uncle, Dr Mikhail Kulikov introduced me to PCA during his brief visit from Moscow to western Canada in 1996. Ms Bonny Lambert prepared the figure. Ms Nasrin Dhanani, occupational hygienist at the University of Alberta, made valuable suggestions on making the manuscript more accessible to practicing occupational hygienists. Ms Pamela Cruise proofread and edited the final manuscript.

REFERENCES

Beaton R, Murphy S, Johnson C *et al*. (1998) Exposure to duty-related incident stressors in urban firefighters and paramedics. J Trauma Stress; 11: 821–8.

Blair A, Stewart PA. (1990) Correlation between different measures of occupational exposure to formaldehyde. Am J Epidemiol; 131: 510–16.

Burstyn I, Boffetta P, Heederik D *et al*. (2003a) Mortality from obstructive lung diseases and exposure to polycyclic aromatic hydrocarbons among asphalt workers. Am J Epidemiol; 158: 468–78.

Burstyn I, Boffetta P, Kauppinen T *et al*. (2003b) Performance of different exposure assessment approaches in a study of bitumen fume exposure and lung cancer mortality. Am J Ind Med; 43: 40–8.

Burstyn I, Ferrari P, Wegh H *et al*. (2002a) Characterizing worker exposure to bitumen during hot mix paving and asphalt mixing operations. Am Ind Hyg Assoc J; 63: 293–9.

Burstyn I, Kromhout H. (2002) Trends in inhalation exposure to hydrocarbons among commercial painters in the Netherlands. Scand J Work Environ Health; 28: 429–38.

Burstyn I, Kromhout H, Kauppinen T *et al*. (2000) Statistical modeling of the determinants of historical exposure to bitumen and polycyclic aromatic hydrocarbons among paving workers. Ann Occup Hyg; 44: 43–56.

Burstyn I, Randem B, Lien JE *et al*. (2002b) Bitumen, polycyclic aromatic hydrocarbons and vehicle exhaust: Exposures levels and controls among Norwegian asphalt workers. Ann Occup Hyg; 46: 79–87.

Chan-Yeung M, Lam S, Kennedy S *et al*. (1994) Persistent asthma after repeated exposure to high concentrations of gases in pulpmills. Am J Respir Crit Care Med; 149: 1676–80.

Clark JA, Wray N, Brody B *et al*. (1997) Dimensions of quality of life expressed by men treated for metastatic prostate cancer. Soc Sci Med; 45: 1299–309.

Frenich AG, Aguilera PA, Gonzalez FE *et al*. (2002) Dermal exposure to pesticides in greenhouses workers: Discrimination and selection of variables for the design of monitoring programs. Environ Monit Assess; 80: 51–63.

Harman HH. (1976) Modern factor analysis. Chicago: The University of Chicago Press.

Kennedy S, Enarson DA, Janssen RG *et al*. (1991) Lung health consequences of reported accidental chlorine gas exposures among pulpmill workers. Am Rev Respir Dis; 143: 74–9.

Kleinbaum DG, Kupper L, Muller KE. (1988) Variable reduction and factor analysis. In Payne M, editor Applied regression analysis and other multivariable methods. CA: Belmont, Duxbury Press, an imprint of Wadsworth Publishing Company. pp. 595–641.

Kromhout H, Symanski E, Rappaport SM. (1993) A comprehensive evaluation of within and between-worker components of occupational exposure to chemical agents. Ann Occup Hyg; 37: 253–70.

Kromhout H, Vermeulen R. (2001) Temporal, personal and spatial variability in dermal exposure. Ann Occup Hyg; 45: 257–73.

Kumagai S, Matsunaga I. (1995) Changes in the distribution of short-term exposure concentration with different averaging times. Am Ind Hyg Assoc J; 56: 24–31.

Kumagai S, Matsunaga I. (1999) Within-shift variability of short-term exposure to organic solvent in indoor workplaces. Am Ind Hyg Assoc J; 60: 16–21.

Meijster T, Burstyn I, Van Wendel de Joode B *et al*. (2004) Evaluating exposures to complex mixtures of chemicals during a new production process in the plastics industry. Ann Occup Hyg; 48: 499–507.

Morrow LA, Ryan CM, Hodgson MJ *et al*. (1991) Risk factors associated with persistence of neuropsychological deficits in persons with organic solvent exposure. J Nerv Ment Dis; 179: 540–5.

Nieuwenhuijsen MJ, Lowson D, Venables KM *et al*. (1995) Correlation between different measures of exposure in a cohort of bakery workers and flour millers. Ann Occup Hyg; 39: 291–8.

Nylander-French LA, Kupper L, Rappaport SM. (1999) An investigation of factors contributing to styrene and styrene-7,8-oxide exposures in the reinforced-plastics industry. Ann Occup Hyg; 43: 99–109.

Ott MG, Klees JE, Poche SL. (2002) Respiratory health surveillance in a toluene di-isocyanate production unit, 1967–97: Clinical observations and lung function analyses. Occup Environ Med; 57: 43–52.

Pearson K. (1901) On lines and planes of closest fit to systems of points in space. Phil Mag; 2: 559–72.

Peretz C, Goren A, Smid T *et al*. (2002) Application of mixed-effects models for exposure assessment. Ann Occup Hyg; 46: 69–77.

Pio CA, Nunes TV, Borrego CS *et al*. (1989) Assessment of air pollution sources in an industrial atmosphere using principal component and multilinear regression analysis. Sci Total Environ; 80: 279–92.

Preller L, Burstyn I, de Pater N *et al*. (2004) Characteristics of peaks of inhalation exposure to solvents. Ann Occ Hyg; 48: 643–692.

Rappaport SM. (1991) Assessment of long-term exposures to toxic substances in air. Ann Occup Hyg; 35: 61–121.

Rappaport SM, Weaver M, Taylor D *et al*. (1999) Application of mixed models to assess exposures monitored by construction workers during hot processes. Ann Occup Hyg; 43: 457–69.

Sahl JD, Kelsh MA, Smith RW *et al*. (1994) Exposure to 60 Hz magnetic fields in the electric utility work environment. Bioelectromagnetics; 15: 21–32.

Salisbury DA, Enarson DA, Chan-Yeung M *et al*. (1991) First-aid reports of acute chlorine gassing among pulpmill workers as predictors of lung health consequences. Am J Ind Med; 20: 71–81.

Samuels SJ, Lemasters GK, Carson A. (1985) Statistical methods for describing occupational exposure measurements. Am Ind Hyg Assoc J; 46: 427–33.

Symanski E, Kupper LL, Kromhout H *et al*. (1996) An investigation of systematic changes in occupational exposure. Am Ind Hyg Assoc J; 57: 724–35.

Vermeulen R, de Hartog J, Swuste P *et al*. (2000) Trends in exposure to inhalable particulate and dermal contamination in the rubber manufacturing industry: Effectivenes of control measures implemented over a 9-year period. Ann Occup Hyg; 44: 343–54.

Vermeulen R, Li G, Lan Q *et al*. (2004) Detailed exposure assessment for a molecular epidemiology study of benzene in two shoe factories in China. Ann Occup Hyg; 48: 105–16.

Villeneuve PJ, Agnew DA, Corey PN *et al*. (1998) Alternate indices of electric and magnetic field exposures among Ontario electrical utility workers. Bioelectromagnetics; 19: 140–51.

Wegman DH, Einsen EA. (1992) Measuring exposure for the epidemiologic study of acute effects. Am J Ind Med; 21: 77–89.