INSTITUTE
FOR WORK & HEALTH
INSTITUT DE RECHERCHE
SUR LE TRAVAIL ET
LA SANTÉ

Occupational Health and Safety Management Audit Instruments

A Literature Review

About this report:

Authors:
Philip L. Bigelow[1,2], Lynda S. Robson[1]

Affiliations:
[1] Institute for Work & Health, Toronto, ON Canada
[2] Department of Public Health Sciences, Faculty of Medicine, University of Toronto, Toronto, ON Canada

Acknowledgements
The authors greatly appreciate: the expert guidance of Emma Irvin on the library search strategy; the assistance of Emma, Krista, Quenby and Dan in obtaining the various bibliographic information and materials; the editorial advice of Kathy Knowles Chapeskie, Tony Culyer, Evelyne Michaels and Cam Mustard; and the administrative help from Lyudmila Mansurova.

If you have questions about this or any other of our reports, please contact us at:

Institute for Work & Health
481 University Avenue, Suite 800
Toronto, Ontario, M5G 2E9

E-mail: info@iwh.on.ca

Or you can visit our web site at www.iwh.on.ca

Please cite this report as: Bigelow P, Robson L. Occupational Health and Safety Management Audit Instruments: A Literature Review. Toronto: Institute for Work & Health, 2005.

Table of Contents

Foreword iii

1. Introduction 1
   - 1.1 Background and context for the review 1
   - 1.2 What is an occupational health and safety management audit? 1
   - 1.3 Measurement concepts applied to management audits 2
   - 1.4 Factors which influence the reliability and validity of audits 3

2. Methods 7
   - 2.1 Literature search 7
   - 2.2 Inclusion criteria and publication retrieval 7
   - 2.3 Review process, data extraction and report generation 8

3. Results 9
   - 3.1 Audit instruments for assessing OHS management 9
     - 3.1.1 Diekemper and Spartz (D&S) method 9
     - 3.1.2 MISHA 11
     - 3.1.3 International Safety Rating (ISR) system 11
     - 3.1.4 CHASE audits 13
     - 3.1.5 Adaptation of OSHA's Program Evaluation Profile (PEP) 13
     - 3.1.6 Goodyear Tire and Rubber Company audit 14
     - 3.1.7 AS/NZS 4804-based audit for small- and medium-sized organizations 15
     - 3.1.8 American Industrial Hygiene Association (AIHA) ISO 9001 harmonized OHS management system 15
     - 3.1.9 AIHA Universal OHSMS Assessment Instrument 16
     - 3.1.10 Summary of the research evidence on the reliability and validity of OHS management audits 18
   - 3.2 Audits of safety management systems in high-hazard and high reliability operations 20
     - 3.2.1 Management system auditing of high-hazard and high reliability operations 23

4. Discussion 31
   - 4.1 State of the evidence regarding the reliability and validity of audits 31
   - 4.2 Limitations of the review 32
   - 4.3 Practical implications of review findings 32

5. Conclusions 35

Bibliography 37

Appendix: Organizing structure of OHS management audits 45

Foreword

In recent years, the Institute for Work & Health has been actively engaged in building relationships with Prevention System agencies and organizations in Ontario.

In these encounters, we often hear that potential research users want more evidence about the effectiveness of interventions aimed at protecting workers' health. We are also told that even when research evidence exists, it is often hard to access, difficult to understand and is not always presented in language and formats suitable to non-scientific audiences.

In response to these needs, the Institute for Work & Health has established a dedicated group to conduct systematic reviews of relevant research studies in the area of workplace injury and illness prevention. In instances where there are too few studies to conduct a full Systematic Review we may provide our audiences with a narrative review.

- Our systematic review team monitors developments in the international research literature on workplace health protection and selects timely, relevant topics for evidence review.
- Our scientists then synthesize both established and emerging evidence on each topic through the application of rigorous methods.
- We then present summaries of the research evidence and recommendations following from this evidence in formats which are accessible to non-scientific audiences.

The Institute will consult regularly with workplace parties to identify areas of workplace health protection that might lend themselves to a systematic review of the evidence.

We appreciate the support of the Ontario Workplace Safety & Insurance Board (WSIB) in funding this four-year Prevention Systematic Reviews initiative. As the major funder, the WSIB demonstrates its own commitment to protecting workers' health by supporting consensus-based policy development which incorporates the best available research evidence.

Many members of the Institute's staff participated in conducting this Systematic Review. A number of external reviewers in academic and workplace leadership positions provided valuable comments on earlier versions of the report. On behalf of the Institute, I would like to express gratitude for these contributions.

Dr. Cameron Mustard
President, Institute for Work & Health
December, 2005

1. Introduction

This report summarizes the available research evidence on the reliability and validity of audits of OHS management systems. It reviews literature not only from the occupational health and safety field, but also from the process safety field.

1.1 Background and context for the review

Occupational health and safety (OHS) auditing is a systematic process for assessing compliance and verifying conformance with established guidelines or best practices in occupational health and safety. OHS audits are becoming an important tool as occupational health and safety management systems have matured and become integrated with modern quality initiatives. The growth in the number of international standards and agreements that incorporate environmental, health and safety aspects of production has also facilitated widespread development and use of OHS audits. Increasingly, as regulatory agencies move to performance-based approaches to improving OHS, auditing is seen as an effective method of ensuring compliance and improving the performance of prevention systems. Despite their widespread and expanding use in Canada and internationally, however, there has not been a synthesis of the literature on the effectiveness of OHS auditing or on the reliability and validity of audit instruments.

In March of 2005, preliminary work was begun to determine the feasibility, scope and potential study questions for a full systematic review of the literature on the effectiveness of OHS auditing. In performing a systematic review, clearly formulated questions are developed, and systematic and explicit methods are used to identify, select and critically appraise the relevant literature. This preliminary work found that a full systematic review on the effectiveness, reliability, or validity of OHS auditing is not appropriate given the state of the peer reviewed literature. The evidence base addressing the study questions was scant and many of the studies provided findings that were limited in external validity. Nevertheless, we proceeded with a narrative review to provide potentially useful information to stakeholders. Findings from the feasibility study informed the development of the protocol for the narrative review that was conducted.

1.2 What is an occupational health and safety management audit?

The term "audit" is defined as a detailed examination or analysis, especially to assess strengths and weaknesses[1]. Many definitions of audits include the comparison of findings to specific criteria, guidelines or standards (CCPS, 1993; Clark, 1999; Rainer et al., 2000). Financial audits have been around for a long time, although audits of many other aspects of organizational

[1] *The Canadian Oxford Dictionary,* Barber, K., Editor, Toronto, Oxford University Press, 1998.

functioning are also very common. The growing demand for auditing of specific organizational functions is related to the increasing complexities of modern management systems as well as the need to improve quality and efficiency. Audits of health, environment and safety systems have also been widely utilized as they are seen as valuable tools for continuous improvement of OHS management system performance.

Occupational health and safety (OHS) auditing[2] has been defined a number of ways. Although there is consensus among authors that OHS auditing is a means of assessing OHS system performance, authors differ in the scope of what they consider OHS auditing. Glendon (1995) described six types of OHS audits: (1) OHS audits on specific topics (e.g. human factors or hazardous substances); (2) plant technical audits; (3) site technical audits; (4) compliance or verification audits (compliance with legal or internal standards); (5) validation audits (design of OHS management systems themselves); and (6) management safety audits. Cooper (1998) emphasized that OHS auditing is more than a hazard identification exercise and should involve a comprehensive examination of the whole OHS management system itself. Most authors agree with Cooper and consider OHS auditing as an assessment of an entire OHS management system (Le Coze 2005; Jorgensen, 1998; Kuusisto, 2000).

For this review, an OHS audit is considered to be a systematic assessment of an OHS management system, which is "the integrated set of organizational elements involved in the continuous cycle of planning, implementation, evaluation, and continual improvement, directed toward the abatement of occupational hazards in the workplace" (Robson et al., 2005, p. 16).

1.3 Measurement concepts applied to management audits

Various measurement concepts found in psychology, clinical sciences and program evaluation (Guyatt, 1993; Hinkin, 1995; Lipsey, 1983; McDowell and Newell, 1987; Stewart and Ware, 1992; Streiner and Norman, 1995) can be applied to OHS management system audits. The concepts relevant for this review are the following:

- **Variation in responses** – the variation in audit results in relation to the possible range of results for the population of workplaces; sufficient variation is a prerequisite of construct validity.
- **Interrater reliability** – the consistency of audit results when carried out by different auditors or auditing teams.
- **Test-retest reliability** – the consistency of audit results when carried out at two different times separated by a relatively short time period.
- **Responsiveness** – the extent to which audit scores can show change when there is meaningful change in the OHS management system; like reliability, it depends on a high signal-to-noise ratio.

[2] Many authors use the term "safety auditing" even though they are referring to auditing of an occupational health and safety system.

- **Content validity** – the extent to which the content of the audit is complete relative to the relevant universe defined by a specified definitional standard. For OHS management audits, this would presumably be some form of OHS management system definition.
- **Construct validity** – the extent to which relationships exist between audit measures and measures of other constructs, as specified in theory; i.e. audit scores should be relatively highly correlated with other indicators of OHS management system performance such as injury rate; similarly, changes in audit scores should be correlated with changes in indicators such as injury rate; finally, an intervention to improve and OHS management system should result in a change in audit results. (Some might consider the correlation of audit results and injury rate data to be a form of criterion validity (e.g. Dyjack et al., 2003, p. 790)).

Whenever published data allowed, the audit instruments reviewed here were considered with respect to the above measurement properties. Any absence of mention of these aspects in this report implies that no such information was available for the particular audit tools reviewed.

1.4 Factors which influence the reliability and validity of audits

Although no systematic studies on the topic of factors affecting audit reliability and validity were identified, several authors have commented upon the subject in a prescriptive manner (Beckmerhagen, et al., 2003; Blackmore and Shannon, 1996; Cooper, 1998; Dyjack and Levine, 1996; Dyjack et al., 2003; Gay and New, 1999; Gillette et al., 2004; Glendon, 1995; Glendon and McKenna, 1995; Kuusisto, 2000; Kennedy and Kirwan, 1998; Laitinen et al., 1999) A list of factors that potentially may impact the reliability and validity of OHS audits is presented below. It should be noted that this is not an exhaustive listing of all possible factors.

- **Factors related to the auditor**:
  - **Competence** – differing experience, training, knowledge and skills of auditors was discussed as an important factor.[3]
  - **Auditor bias and independence of auditor** – a bias may exist if an auditor/audit team has conflict(s) of interest or is unbalanced in terms of points of view.[4] Points of view may be related to

[3] Three organizations representing OHS professionals, the American Board of Industrial Hygiene, the American Society of Safety Engineers, and the American Industrial Hygiene Associate, have prepared a position paper to provide recommendations concerning qualifications that should be considered in evaluating the competency of individuals tasked with conducting OHS audits. See: *Position Paper on Auditor Competency for Assessing Occupational Health & Safety Management Systems.* September, 2005 American Board of Industrial Hygiene. http://www.abih.org/about.htm

[4] The US National Academies of Science approach for ensuring review committees are as free from bias as possible and are balanced in points of view is helpful in defining conflict of interest and points of view on scientific matters. http://www.nationalacademies.org/coi/index.html

differences in the auditors' professional background and training. For example, an auditor who is a process operator likely has a different point of view as compared to an auditor who is health and safety professional.

- **Internal versus external auditors** – this may be considered part of "auditor bias and independence of auditor" but it is listed separately as it was considered a major factor by numerous authors. Internal auditors often have a more detailed understanding of the OHS management system being audited but may have higher levels of conflict of interest as compared to external auditors. The independence of external auditors, however, can also vary widely from being totally without conflicts of interest to being very conflicted (an example would be a consultant who may gain additional work that is dependent on the audit findings)

- **Factors related to the audit and audit process**:
  - **Theoretical basis for the audit** – the degree to which an audit is based on previous research and sound theory is a factor that may affect audit validity.
  - **Coherent and comprehensive audit framework** – this refers to clearly defined audit objectives, methods, measures, and reporting systems within a clearly articulated framework.
  - **Clear standards for comparison** – this refers to unclear or ambiguous standards on which the audit is based.
  - **Use of multiple sources of information** – this refers to collection of data from multiple sources in the measurement of an activity. Multiple sources may improve validity but the complexity of integrating multiple sources may influence reliability. A number of authors emphasized the importance of interview and observational data to determine the implementation of written policies and procedures.
  - **Sampling** – this refers to the selection of information/data used in the audit. For example, the method of selection of individuals who will be interviewed during the audit. This also refers to the selection of plants/worksite for auditing.
  - **Detail in procedures and objectivity of audit questions** – the amount of detail and the degree of objectivity in the audit procedures are cited by numerous authors as influencing reliability. Kuusisto (2000) pointed out that the use of a structured audit method may increase reliability, but if the questions do not adequately assess the appropriate OHS activities the validity of the audit can be affected.
  - **Measurement scales and clarity in the delineation between steps** – this refers to the structure of scales used to measure an activity. Clarity between steps of a scale is related to "detail in procedures and objectivity of audit questions" mentioned above.

- **Weighting of various audit components** – this refers to the selection of weighting factors applied to individual component scores to arrive at an overall score for a plant or facility. Some authors felt that weightings may influence audit validity as inappropriate weightings may mask serious problems or overemphasize minor ones.
- **Resources for** auditing – this refers to personnel, time, and financial resources available for auditing.
- **Quality control of auditing** – an external review of the auditing procedures was discussed by some authors as a factor in reliability and validity.

2. Methods

2.1 Literature search

The search strategy targeted management audits. It was developed using MEDLINE, a bibliography of journal articles from the broader medical literature, using a small sample of known relevant articles. MEDLINE classifies each article with relevant keywords from a controlled vocabulary (MeSH terms), allowing researchers to design literature searches with high specificity (i.e. a large percentage of articles will be relevant). The search strategy developed in MEDLINE was also used with four other bibliographic databases (EMBASE, American Business Inform (ABI), Econlit, CCInfoWeb). Together the five databases used in the search cover the fields of medicine, management, economics, and occupational health and safety. The MEDLINE search strategy was customized to each database by: substituting synonymous keywords where necessary, or searching the entire text of the title and abstract with a particular term (free text search) if keywords could not be used.

Two basic search strategies were used. One looked for abstracts classified with the keyword "management audit" and one of the following keywords or free text terms: "wounds and injuries," "accidents, occupational," "accident prevention," or "occupational." The second strategy looked for abstracts classified with the keyword "safety management" and the free text term "audit." No restrictions were placed on the searches regarding date and language of the original publication.

The titles and abstracts arising from the searches were reviewed to identify relevant publications using two inclusion criteria (see below). Forty-four unique titles and abstracts were identified at this initial step. Their source publications were retrieved and reviewed in more detail.

In order to broaden the search beyond the original bibliographic sources, the reference sections of publications deemed relevant were reviewed for relevant titles. Their source publications were also retrieved.

2.2 Inclusion criteria and publication retrieval

The relevance of titles and abstracts was determined by applying two criteria:

- The publication contains information on occupational health and safety management system audit reliability or validity. Publications reporting only on the one time use of an audit were not included.
- The publication is a journal article, book, conference proceeding, or report; it is not a magazine article or newsletter.

If a title and abstract appeared relevant or possibly relevant, the corresponding full publication was retrieved for further review. Additional publications on the topic, which did not meet the criteria, were also retrieved to provide contextual, descriptive or conceptual information related to the study questions.

2.3 Review process, data extraction and report generation

Both authors independently reviewed all retrieved publications. The content of the publications was then discussed and a general, common understanding of the findings was established. Clearly irrelevant publications, including those focusing only on particular hazards, were excluded.

Next, approximately half the publications were allocated to each author for extraction of pertinent information, according to their respective responsibilities for particular subject areas. After extracting this information and summarizing it in the Results section of this report, the authors again discussed the findings and generated the material found in the Discussion and Conclusions sections.

3. Results

The results of the review are presented in two sections. The first focuses on instruments designed for the audit of OHS management systems. The second section is concerned with audit instruments intended for the safety management systems of high-hazard and high reliability operations. It would have been reasonable to exclude the latter group of publications from this review, since their focus on OHS is secondary and they deal with only one category of OHS risk. However, the authors of this review chose to include this group since they overlapped conceptually with the first group. In addition, the high-hazard and high reliability literature seemed to have engineering as its disciplinary base. This literature therefore had the potential to suggest new approaches to OHS management auditing, which has somewhat different disciplinary underpinnings (e.g. organizational theory, psychology, health sciences).

3.1 Audit instruments for assessing OHS management

This section reviews the audit instruments designed for the audit of OHS management systems. Eleven distinct instruments are described[5] and any available evidence about their measurement properties is summarized. The order of presentation of these instruments is roughly in the chronological order of their development. They are also ordered so that the ones most clearly based on system theory concepts (Emery, 1971) appear last. The organizing structures of all instruments are summarized in the Appendix.

3.1.1 Diekemper and Spartz (D&S) method

The OHS management audit method developed by Diekemper and Spartz (1970) appears to be the earliest in the literature. At the time of publication, they had already been using the instrument in their practice. According to Kuusisto (2000), the method has been recommended by well-known safety specialists (Heinrich et al., 1980; Petersen, 1989).

The D&S instrument contained 29 items on OHS activities and was organized into five sections:

- organization and administration;
- industrial hazard control;
- fire control and industrial hygiene;
- supervisory participation, motivation and training; and
- accident investigation, statistics and reporting procedures.

The rater was required to categorize each activity on a four-category scale ranging from poor to excellent. Responses were then subjected to a scoring

[5] The eleven instruments will be described in sections 3.1.1 to 3.1.9.

scheme, which applied weights to the tool's sections, items and even responses.

The Diekemper and Spartz (1970) article was primarily prescriptive, giving no information as to how the audit's content was developed and no explanation for their chosen scoring scheme beyond the statement that it was "determined by in-depth analysis of past occurrences." They claimed "a comparison of the scores from one year to the next is valid to determine progress," but gave no evidence to support this claim.

Uusitalo and Mattila (1989) reported on their use of the D&S method with eleven companies. The companies came from two industrial sectors; some had low accident rates and some had high. The researchers gave evidence of some construct validity of the audit, although that was not their main intent. Mean D&S scores for the five sections of the audit were reported for the low and high companies in both of the sectors. This allowed the reader to make ten comparisons of audit scores for low- and high-accident companies. In seven of the comparisons, as expected, the scores in the low-accident companies were higher than those in the high-accident companies; in three of the comparisons, the scores of the low-accident companies were lower. Scores in two of the five sections of the audit (industrial hazard control; accident investigations, statistics and reporting procedures) clearly discriminated between low and high accident rate companies. However, statistical analyses were not applied.

Kuusisto (2000), from Finland, studied the interrater reliability of a modified version of the D&S method (relatively minor modifications). He found the method had unacceptably poor interrater reliability, ranging from poor to moderate[6] (weighted kappa (κ_W) values from -0.03 to 0.46) when his own ratings were compared with those of local company evaluators for six American workplaces.

Agreement between raters was better when Kuusisto's ratings of three Finnish companies were compared with the ratings of his safety specialist students. Kuusisto concluded that the reliability of the D&S method was highly dependent on the training and local expertise of the auditor. However, even with the greater homogeneity among raters' characteristics, the degree of agreement covered a large range, from fair to near perfect (κ_W from 0.36 to 0.83). The data showed that the degree of agreement varied both by student and by workplace. In the workplace where the least agreement was found, weighted kappa was less than 0.40 for three of the six student-to-Kuusisto comparisons.

[6] The authors classified the weighted kappa values according to the definitions of Landis and Koch (1977).

3.1.2 MISHA

Kuusisto (2000) then developed a new audit tool, called the Method for Industrial Safety and Health Activity Assessment (MISHA). His aim in creating the new tool was to improve upon the D&S method, particularly its inter-reliability. He also wanted to improve its comprehensiveness (i.e. its content validity). In order to do this, he reviewed several organizational assessment instruments and drew on the knowledge of experts in several institutions. The second and final version of MISHA contained 55 items and used a four-category response framework. The organizing framework for the content was taken from Booth and Lee (1995):

- organization and administration
- training and motivation
- work environment
- follow-up

Kuusisto tested the first version of MISHA in one Finnish company, comparing his ratings with those made by four company members (personnel manager, safety director, employee safety representative, safety manager). Interrater reliability ranged only from slight to fair, so the tool was revised. A test at a second Finnish company was carried out in a similar manner. In this case, agreement improved and ranged from fair to moderate (κ_W 0.38 to 0.58). The highest agreement was with the managing director who also acted as the safety manager; the lowest was with the employees' safety representative.

Kuusisto (2000) concluded that if an audit tool were to be used by someone who was untrained, then the MISHA tool would likely show more reliability among users. If trained experts were involved, he thought that the D&S method would be better. The authors of this review think that such a comparison between the two tools should be considered preliminary. One reason is that the testing was carried out under different circumstances. Another is that the new MISHA method was only used in one workplace and the results with the D&S method suggested that the degree of interrater reliability depended in part on the particular workplaces audited.

3.1.3 International Safety Rating (ISR) system

The first International Safety Rating (ISR) system manual was published in 1978 by the International Loss Control Institute (ILCI) in the United States (Eisner and Leger 1988). It was designed to be a generic audit that could be used in most industrial sectors, although it was primarily developed in the steel industry (Eisner and Leger, 1988). The chapters of the fourth edition are shown in the Appendix. This version consisted of 627 questions when used in full; less questions (as low as 100) were used when the standard of achievement (star rating) was less than the maximum (Collison and Booth,

1993). No information has been found in the available publications on the development of the content of the ISR tool.

Apart from the comparison of content conducted by Collison and Booth (1993) on the generic version of ISR, most published evaluative information on the ISR (Eisner, 1993; Eisner and Leger, 1988; Guastello, 1991) was concerned with a version developed for the mining sector by the President of the International Loss Control Institute. This version of the ISR system had almost 1000 questions, which raised doubts about its practicality (Eisner and Leger, 1988; Eisner, 1993).

Eisner and Leger (1998) also questioned the audit's content validity with respect to mining. Major mining hazards and concerns were completely overlooked, while more minor ones were covered in detail. In addition, they were critical of the weighting scheme in terms of relative weights, thereby raising an issue about its construct validity. While 260 points were awarded for the establishment of standards and for procedures for ensuring standards are maintained, only 40 points were awarded for actual compliance with rules, as observed by the auditors. Apparently, no rationale for the scoring scheme was given in the manual.

Eisner and Leger (1988) took another approach to the ISR tool's construct validity by examining the correlation of the number of stars awarded in the system with each of fatality rate and reportable injury rate. Higher numbers of stars indicated better OHS management systems in the scoring scheme, with a maximum of five stars possible. The data were from 33 work sites and correlations were determined for each of 1985 and 1986. In the case of fatalities, the coefficients were in the expected direction, though quite low (-0.14 and -0.23 for 1985 and 1986, respectively). In the case of injuries, the direction was opposite to that expected (0.07 and 0.02). However, all four correlation coefficients were statistically insignificant. (However, with data from only 33 worksites – and possibly from only half that number of worksites for each of 1985 and 1986 – the analysis would have been underpowered statistically (Cohen, 1977)). At best then, there is weak evidence from fatality rate data, on the construct validity of the IRS audit approach.

One of the problems in the attempt by Eisner and Leger (1988) to look at relationships between the star status and injury or fatality rates, was the apparent "ceiling effect" of the score in stars. In 1986, 58% of mines had five-star status (the maximum); another 27% had four-star status (Eisner, 1993). Eisner and Leger (1988) pointed out that this variation in responses was insufficient, especially given that there was still room for improvement: several mines with five-star status regularly had several fatalities per year. However, this validity problem was not interpreted as being inherent to the instrument. Eisner and Leger (1988) pointed out that five-star status

required a minimum score of 90% in each element. They thought that achievement of this level of mine performance "would be near to perfection as any mine could hope to attain (Eisner and Leger, 1988, p. 154)." As such, the authors harboured the "strong, though unproved, presumption that [the] auditing [was] biased." In this case then, the problem with validity arose through the audit process itself.

3.1.4 CHASE audits

CHASE is the abbreviation for Complete Health and Safety Evaluation, which is a group of related audits developed through a collaboration of academics and private industry in the United Kingdom. This group of audits includes CHASE-I for small employers, CHASE-II for large employers, Construction-CHASE for organizations in the construction sector, and COSHH-CHASE for organizations needing to comply with the Control of Substances Hazardous to Health Regulations (1988). The chapter titles of CHASE-II are shown in the Appendix.

Glendon et al. (1992) give a partial description of how the CHASE audits were developed. Questions were first derived from legislation and "other relevant sources." Pilot testing across "a range of organizations" led to the redraft of questions and the development of accompanying guidance. The weight given to the various elements in the audits was based on their relative importance in regulations and on the risk posed by particular hazards if left uncontrolled. Glendon's description of the development process by Glendon et al. (1992) suggests that the audit likely had good content validity. However, there was little specific information given which would allow confirmation of this.

The authors briefly mentioned audit scores ranging from 18 to 70 per cent among the nine test organizations in one sector, implying good variation in responses. They also said that the scores in all twelve sections of the audit increased following an intervention. This suggests that the instrument is responsive to change and provides evidence of construct validity. An alternative explanation for the increase could have been bias in the audit process. The reader was not given sufficient information to judge between the two alternatives.

3.1.5 Adaptation of OSHA's Program Evaluation Profile (PEP)

The Occupational Safety & Health Administration in the United States developed an audit tool, the Program Evaluation Profile (PEP). The audit was introduced in 1996 as a tool for its inspectorate, but its use for this purpose was discontinued the same year (OSHA, 2005). LaMontagne et al. (2004) reportedly adapted the audit instrument for use in an evaluation of an intervention. While the original PEP tool consisted of 15 items, each with five detailed item-specific response categories (OSHA, 2005), the LaMontagne et al. (2004) tool consisted of 91 items, each requiring a yes/no

response. These items were organized into four sections (see Appendix), each of which was weighted in the scoring so that the final score theoretically ranged from 0 to 100. A score of one hundred indicated that all audited program elements were fully present.

Audits were conducted on seven intervention sites and eight control sites both before and after the intervention. For the intervention group, the focus of the intervention was both OHS and lifestyle risks; for the control group, it was only lifestyle risks. With only seven or eight sites per group, the statistical analysis had low power. Perhaps as a result of this limitation, the changes in the overall audit scores did not differ significantly for the two groups. A significant difference in the change scores of the two groups was found for only one of the four audit sub-sections.

However, consistent with the program theory, and supportive of the audit instrument's construct validity, the audit scores increased in all four sub-sections of the audit for the intervention sites to a greater extent than they did for the control sites. The standard deviations reported for the baseline total program scores (7.4 and 16.3) suggest that the variation in audit scores was reasonable. Furthermore, the size of change in the intervention group (11.1), relative to the baseline standard deviation (16.3), a measure of instrument responsiveness (Beaton et al., 1997), was considered medium-large using Cohen's (1977) effect size standards.

3.1.6 Goodyear Tire and Rubber Company audit

The audits that were used by the Goodyear Tire and Rubber Company were mentioned in two of the publications reviewed. Dyjack et al. (1998) compared the combined content of two Goodyear audits with that of an ISO 9001-aligned OHS management audit instrument developed for the American Industrial Hygiene Association (AIHA). The content of the two instruments was quite similar, even though the Goodyear Tire and Rubber Company audit was structured differently and was considered to be more of a "traditional" audit (i.e. not as influenced by system theory (Emery, 1971)). These findings indirectly demonstrate content validity for the Goodyear tools, since the content validity of the AIHA tool is strong (see below).

A Goodyear Tire and Rubber Company audit instrument was the original source from which another audit tool evolved (Bunn et al., 2001). Bunn and colleagues applied it to monitoring the improvement in a large company's comprehensive health, safety, and productivity intervention. The summary scores over a three-year period indicated continuous improvement: 1997, 63 per cent; 1998, 68 per cent; 1999, 79 per cent. (Audit scores can theoretically range from 0 to 100 per cent). The results reported by Bunn et al. (2001) also suggest that the audit instrument had responsiveness and construct validity. However, little information about the audit process was reported, except that it was cross-plant and conducted by internal auditors.

The authors did not comment on the possibility that auditor bias might explain some of the results.

3.1.7 AS/NZS 4804-based audit for small- and medium-sized organizations

Pearse (2002) developed an audit instrument for small- and medium-sized organizations based on the Australian and New Zealand OHS management system standard (AS/NZS 4804-1997). Baseline audit scores ranged from 12 to 85 per cent (within the theoretical maximum range of 0 to 100 per cent), with a standard deviation of 18.9. The mean change in audit score as a result of the intervention was ten percentage points, which represents a medium-sized effect of 0.5 (Cohen, 1977). This evidence of instrument responsiveness and construct validity must be viewed cautiously since no information on the conduct of the audit was provided, and the potential for bias remains unknown. Pearse also stated that "the audit tool was … trialled independently by different external auditors and was found to deliver almost identical results in their hands," but no further details were given.

3.1.8 American Industrial Hygiene Association (AIHA) ISO 9001 harmonized OHS management system

The remaining two OHS management audit tools reviewed here, in contrast with the preceding tools, show strong evidence of content validity. Both are products of student PhD theses at the University of Michigan under the supervision of Stephen Levine.

The first of these (Dyjack 1996; Dyjack et al., 1998) used ISO 9001 as the organizing framework and incorporated extensive input from a wide range of stakeholders. ISO 9001 is the widely accepted quality management system standard developed by the International Organization for Standardization. As described in Dyjack's thesis (1996), he and Levine developed an initial version of an ISO 9000 harmonized auditable standard and guidance document, after reviewing several OHS and environmental management system or program documents. This was donated to the American Industrial Hygiene Association's OHS Management System Task Force, which included representatives of labour, industry, academia, government, OHS consultants and the insurance industry. The Task Force provided three rounds of input, using a modified Delphi process. Input was also given by AIHA members responding to an invitation in their association's newsletter. Over 50 OHS professionals or their national associations gave feedback on the instrument in the course of its development. The fifth version was adopted by AIHA's Board of Directors in 1996 and it was published by AIHA as a guidance document that same year.

Dyjack et al. (1998) compared their "theoretical" ISO 9001-harmonized audit to two in-house audits being used regularly by the Goodyear Tire

Company. One of the Tire Company's audits focused on industrial health; the other focused on safety. Dyjack et al. (1998) selected the Goodyear audits to represent more "traditional" audits in current use. The organizing structure of the Goodyear industrial health audit did show more vertical structure (i.e. organized around particular risks) than that of the AIHA instrument which had a more horizontal structure (organized around management system elements that apply to all risks). Nevertheless, an analysis of the content of the AIHA document showed it to be comparable to the combined content of the Goodyear Tire Company documents. It should be noted that Goodyear reviewed their audit documents annually, drawing on extensive, diverse corporate expertise. This presumably ensured that the content of their tools remained current.

3.1.9 AIHA Universal OHSMS Assessment Instrument

The second audit instrument developed at the University of Michigan (Redinger, 1998; Redinger and Levine, 1998; Redinger et al., 2002a; Redinger et al., 2002b) aimed to be a "universal" OHS management system audit instrument. This work was associated with the American Industrial Hygiene Association (AIHA, 1999) and a report by the International Occupational Hygiene Association to the International Labour Organization (Dalrymple, 1998). The following description indicates strong content validity for this instrument: Redinger and colleagues defined a standard (the "universe") and then ensured that they fully captured and operationalized the universe by the use of a systematic development process and extensive expert advice.

University of Michigan researchers defined the "universe" of OHS management systems by reviewing 13 publicly available OHS and environmental standards or guidance documents. They then selected four of these, which they thought collectively represented the content in all 13 documents. These were:

- the AIHA ISO 9001-harmonized OHS management system (see 3.1.8)
- the U.S. Occupational Safety and Health Association (OSHA)'s Voluntary Protection Program, which was the most comprehensive management system within OSHA
- ISO 14001, the International Organization for Standardization's environmental management system standard
- BS8800, a voluntary standard from the British Standards Institute, based on both the Health and Safety Executive's HSG65 model (HSE, 1997) and ISO 14001.

Redinger and colleagues deconstructed the four "input models" down to the level of single clauses and then reorganized them into an integrative model.

The final universal OHS management system model had five organizing categories derived from system theory and policy analysis models:

- Initiation (OHS Inputs)
- Formulation (OHS Process)
- Implementation/Operations (OHS Process)
- Evaluation (Feedback)
- Improvement/Integration (Open System Elements)

Distributed across the categories were 27 OHS management system elements (16 primary and 11 secondary; listed in the Appendix). The audit instrument's sections (corresponding to the elements) contained 188 principles (whose level of detail was similar to the input clauses) and 486 corresponding measurement criteria. The criteria operationalized the principles for purposes like auditing. Several OHS experts gave input during the model's development and the penultimate version of the model was reviewed by an experienced multi-stakeholder group (labour, industry, academia, government, and professional trade associations).

The universal OHS management assessment tool was pilot tested in three sites (Redinger et al., 2002a-b), with a particular focus on the Initiation category. This category contained four elements: management commitment and resources; regulatory compliance and system conformance; accountability, responsibility, and authority; and employee participation. Raters assessed the corresponding OHS management system principles using two mutually exclusive scales. The first was intended for workplaces still in the process of developing base conformance to a principle. The second scale classified the degree of conformance after a minimum of base conformance had been achieved. Results of the scale scores were found to be consistent with a qualitative assessment of the three sites, giving preliminary construct validity to the audit instrument.

Dyjack et al. (2003) subsequently looked at the reliability of the universal assessment instrument, using a different scoring system. Four audit sections from three different categories were applied in the testing (employee participation, training, hazard control systems, and communications). In total, these sections contained 102 auditable clauses (called measurement criteria in earlier version). Using these criteria, two auditors conducted one-day audits of four different workplaces. The conformance with each clause was assessed on a scale from zero to five, where zero represented "absence" and five represented "state of the art."

Four analytical approaches were taken to looking at the consistency of the two auditors' ratings. In the first, Pearson product moment correlations were calculated for each section and for each site. Since the Pearson coefficient is insensitive to scale (i.e. could show two auditors having perfect correlation even though they were consistently two points apart in

the rating scale), they also conducted t-tests to see if scores assigned by the auditors were significantly different. Finally, two different intraclass correlation coefficients were calculated: Cohen's weighted kappa and the within-group interrater reliability coefficient.

Pearson correlations were found to range from 0.21 to 0.50, varying by site. The differences between auditor overall mean scores for each site were -0.40, -0.15, 0.07, and 0.47. The two differences at the extreme were statistically significant ($p<0.05$). When looking at the differences at the level of instrument sections and by site, the scores were significantly different for seven of the 16 comparisons and ranged as high as -1.1. Cohen's weighted kappa values ranged from 0.12 to 0.28 among the four sites. Interrater reliability coefficients ranged from 0.28 to 0.66. The authors considered all the coefficients of agreement to be low, relative to 0.70, a commonly used criterion in research on humans (e.g. psychology, clinical sciences).

The original authors had expected higher consistency between auditors because both had been instrumental in the design of the audit tool. They also had PhDs, were certified as industrial hygienists, had attained the status of lead auditor for ISO 14001, and had more than 15 years experience, including VPP site assessments. Some of the factors they thought might have influenced results: sites were not thoroughly prescreened; they had only one day on site; the six point measurement scale was sometimes problematic (they thought the three categories used in ISO audits might be better); and as part of the research protocol and in contrast to normal audit practice, there was no exchange between auditors, nor feedback to hosts.

The authors concluded that their findings raised "potentially disturbing questions regarding the reliability of OHS management program and system audit findings, particularly in light of the emphasis industry has placed on certifications and status achieved secondary to 'passing' an audit." They recommended that OHS management programs and system audit reliability research be expanded to include governmental and commercially available audit instruments in order to "produce confidence in existing tools."

3.1.10 Summary of the research evidence on the reliability and validity of OHS management audits

As indicated at the outset of this report, the literature on the measurement properties of the OHS management audit instruments is sparse. Among the eleven instruments described above, only the AIHA instruments could be considered to have strong content validity. Other instruments might also have this property, but the level of detail in the published information did not allow a judgment about this to be made.

There was some evidence of construct validity for the instruments that measured improvement in audit scores following an intervention on the OHS program (Bunn et al., 2001; Glendon et al., 1992; LaMontagne et al., 2004; Pearse, 2002). However, in all cases except that of LaMontagne et al. (2004), insufficient information was provided to judge whether auditor bias might have influenced the results. Such bias can be a concern when the auditor is responsible for both delivering an intervention and measuring its effect using audit scores. This was the concern of Eisner and Leger (1988) regarding the IRS auditing system used in mining in South Africa. They thought auditor bias was strongly suggested by the high prevalence of maximum five-star ratings.

This "ceiling effect" in the IRS scores, whether attributable to auditor bias or not, might have contributed to the inability of Eisner and Leger (1988) to detect a statistically significant correlation between the star rating and either injury rate or fatality rate. It should be noted that their analysis was conducted with low statistical power. In any case, the size of correlation between fatality rates and star status was small (Hemphill, 2003); between injury rates and star status, it was trivial. The evidence provided by Eisner and Leger's (1988) investigation of the construct validity of the IRS audit must therefore be considered weak at best.

No study validated audits against a quantitative injury rate criterion. Uusitalo and Mattila (1989) reported the audit scores for two small groups of companies in two sectors. One group had high injury rates; the other low. No statistical analysis was conducted, but examination of the raw data suggests that the audit had some ability to discriminate between the two groups of companies. Notably, two sections of the audit were better at doing this than the remaining three sections.

There was sufficient data provided in the Pearse (2002) and LaMontagne et al. (2004) articles to calculate an effect size statistic; i.e. the ratio of the change in mean audit score, relative to the baseline standard deviation (Beaton et al., 1997). In both cases, the instruments demonstrated responsiveness, since the effect sizes were considered medium and medium-large (Cohen, 1977), respectively. It is difficult to know how impressive these findings are, since so little comparative information on audit instrument responsiveness is available. They are a good size in relation to the changes seen in health status measurements of recovery from musculoskeletal disorders reported by Beaton et al. (1997).

Interrater reliability was the measurement property for which the most evidence was found in the literature. Kuusisto (2000) investigated this property in two audit instruments. He found the interrater reliability was low (weighted kappa (κ_W) values from -0.03 to 0.46) when auditors with different cultural and professional backgrounds used a modified version of

the original Diekemper and Spartz (1970) method. The range of interrater reliability was somewhat better when the auditors were more similar (κ_W from 0.36 to 0.83), but still had the potential to be only "fair" (Landis and Koch, 1977) at the low end of the range of agreement. Kuusisto (2000) attempted to improve upon the D&S method by developing the MISHA instrument. Weighted kappa values ranged from 0.38 to 0.58 (considered fair to moderate) when agreement between Kuusisto and selected individuals working in audited workplaces was determined. This range of values is relatively low by the standards for research instruments, for which a reliability coefficient of 0.70 is generally sought. In any case, the results for the MISHA instrument must be considered preliminary since they were investigated in only workplace for the final version of the instrument.

The interrater reliability of the American Industrial Hygiene Association's universal OHS management system audit was the focus of a study by Dyjack et al. (2003). In spite of the expert qualifications of the two auditors in the project, the extreme similarity of their preparation, and the demonstrated content validity of the instrument, the agreement between raters was considered inadequate using several statistical criteria.

3.2 Audits of safety management systems in high-hazard and high reliability operations

Breakdowns in high-hazard processes[7] such as chemical production or nuclear power generation have the potential for devastating consequences. Other processes such as aviation, rail travel, and marine transportation require that critical operations be very high in reliability to prevent sudden catastrophic losses. These high-hazard and high reliability operations[8] require comprehensive safety management systems[9] to ensure the protection of both workers and the public. The evaluation of safety management systems for these high-risk processes requires special consideration, as "the absence of a very unlikely event is not, in itself, a sufficient indicator of good safety performance." (EPSC, 1996) Thus, the typical performance measures that focus on lagging indicators (e.g. accidents) have very limited or no utility.

A number of terms are used to describe the assessment of safety management systems for high-hazard processes and high reliability operations. Terms like "risk-based safety management auditing" as well as "probabilistic safety program auditing" include the auditing of both high-hazard processes and high reliability operations. "Process safety management system auditing," or more simply "process safety auditing," are

[7] Also referred to as "high-hazard sites" and in some jurisdictions as "major hazard sites."
[8] Also referred to as high reliability organizations (HROs). HROs are considered to operate with nearly failure-free performance records.
[9] In the literature pertaining to high-hazard and high reliability operations the term "safety management systems" is commonly used as opposed to OHS management systems.

terms that focus only on high-hazard processes. Although more limited in its definition and applications, process safety management system auditing is well described in the literature since a majority of high-risk processes involve production systems (mostly chemical but including biological and others hazardous energies and materials). Process safety management system auditing refers to the systematic review of process safety management systems (CCPS, 1993)[10].

The assessment of high reliability operations is relatively new when compared with auditing of high-hazard processes. Although differences exist between the two methods, they share many qualities, since both rely on the ability to predict potential future adverse incidents and their consequences.

Before we can discuss the literature in this area, we will briefly review the basic elements and function of safety management systems and auditing for high-hazard processes and high reliability operations.

Modern process safety management systems incorporate a number of elements embedded in a structural model that facilitates feedback and learning (Hale, 2003). Typical program elements include written operating procedures, analysis and identification of hazards, process hazard analysis, risk analysis, employee training, prestart-up reviews, incident investigation, and compliance auditing. Because of the inability to use the frequency of incidents as a measure of performance, a cornerstone of process safety management is the process hazard analysis (Roughton, 1993). Process hazard analysis is the study of potentially hazardous situations associated with a process or specific activity, using qualitative techniques to identify weaknesses in design and operation (CCPS, 1993).

Many regulations governing high-hazard sites insist that risk analysis be included in process safety management systems. Risk analysis, which incorporates process hazard analysis, provides an estimate of risk by integrating information about scenarios, frequencies and consequences of accidents (Wang, 2004). In the risk analysis procedure, each individual process is examined to determine potential hazards and possible scenarios; this is the process hazard analysis. Information about the probability that any hazard scenarios (failures) will occur is obtained from databases. The likely consequences of these hazard scenarios are obtained from historical information or predicted through computer modeling. The probability of occurrence and the consequences for each hazard scenario are used in a risk evaluation of the process. If the risk is acceptable, the risk analysis is

[10] Process safety management is defined as "the application of management systems to the identification, understanding, and control of process hazards to prevent process-related incidents and injuries." (CCPS, 1993)

complete. If the risk is unacceptable, the process safety procedures are modified and the risk analysis is repeated.

At this point, we feel the need to differentiate between safety management audits and methods of risk analysis and assessment. Our definition of safety management audits of high-hazard processes and high reliability operations focuses on management functions. However, most risk analyses and assessments focus on the technical elements of process safety.

A brief review of the common risk analysis methods and their potential application in safety management system audits is important. More detailed reviews of risk assessment and its roles in process safety management strategy and evaluation can be found in the literature (Le Coze, 2005; Frick, in press).

Most specific risk analysis methods focus on equipment failure and procedures closely associated with potential accidents. For example, HAZOP entails an examination of process piping and instrumentation to identify possible failures and the need for safeguards. Some of the methods, such as Event Tree Analysis and Failure Modes and Effects Analysis include equipment failure and human errors as risk contributors. Of these, Human Reliability Analysis and HAZOP[11] have been adapted to include not only human errors at the operator level, but also management factors that may affect human performance and error.

An important feature of these specific risk analysis methods is that their validity and reliability can be studied because of the availability of historical data for equipment failures and human error. For example, Kirwan (1997) summarized 22 validation studies of nine methods of conducting Human Reliability Assessments. The goal was to determine which methods were supported by empirical evidence, which were not, and which were in need of validation studies. Although critical analysis of each of these validation studies is beyond the scope of this report, findings and recommendations from Kirwan's (1997) review may be useful in discussing the reliability and validity of safety management audits for high-risk operations.

Another important component of process safety management is quantitative risk assessment (QRA), also termed quantified risk assessment (Hurst et al., 1994). QRA involves a numerical evaluation of incident frequencies and consequences for an entire facility and relies upon specific hazard and risk analysis methods. QRA, along with most methods of risk analysis, uses historical data and modeling as inputs to quantify the consequences of specific accidents (e.g. dispersion modeling of vapour plumes and predicted

[11] An adaptation of HZOP that includes safety management process flows and safety management errors is termed "SCHAZOP" (Safety Culture Hazard and Operability) (Kennedy and Kirwan, 1998).

exposure concentrations). It also allows for the determination of the major components of risk for a given plant along with their susceptibly to failure due to human factors and equipment failure. QRA studies use information on failure rates and error rates obtained from databases. Using these generic rates without consideration of the management systems is a limitation of most QRAs (Williams and Hurst, 1992).

3.2.1 Management system auditing of high-hazard and high reliability operations

There are a number of audit methodologies that have been used in high-hazard industries and high reliability operations. Fourteen of the articles retrieved for this review were identified as relevant for management systems auditing of high-hazard or high reliability operations. A more detailed review of the 14 studies resulted in the exclusion of six because they did not meet the inclusion criteria. Five of the six excluded (Hurst and Ratcliff, 1994; Hurst et al., 1994; Ratcliff, 1993a; Ratcliff, 1993b; Ratcliff, 1993c) concerned the initial development of STATAS (Structured Audit Technique for the Assessment of Safety Management Systems) and little information on reliability and validity was provided. Each of the studies that met the inclusion criteria are grouped below by audit name and discussed in chronological order.

3.2.1.1 MANAGER

A process safety management system evaluation technique called MANAGER was developed in 1986 for use in QRA in the chemical industry (Pitblado et al., 1990). Subsequent versions of MANAGER reduced the number of questions to 114, revised the scoring system and included non-linear effects on failure frequencies for plants that deviated from average. The revised questionnaire was structured into 12 broad topic areas that paralleled the major elements of the Center for Chemical Process Safety "Guidelines for Technical Management of Chemical Process Safety." In 1990 the updated MANAGER audit included questions to more fully explore underlying principles of effective management (organizational implementation, problem definition, control and auditing). The revised version also permitted both a qualitative and quantitative assessment of safety management system factors (Pitblado et al., 1990).

Pitblado et al. (1990) provided a discussion of some of the applications of MANAGER as well as overall findings from about 30 facility assessments. The scoring system produces an overall score called the Management Factor (MF) and a MF of 1.0 is representative of an average plant. The possible range of scores is from 0.1 for the best possible plant (10 times better than average) to 100 for a completely unsafe plant (100 times worse than average).

Among the 30 or so audits that have been conducted with MANAGER, the range of scores was from 0.5 to 8.0 (Pitblado et al., 1990). The authors pointed out that the scores obtained for each plant, in virtually all cases, reflected the findings of Chemical Process Quantitative Risk Assessments. The sites that scored well were those operated by companies with a well-known safety management culture, supported by OHS resources and information from large central safety organizations. Smaller companies and those known for largely ignoring safety scored poorly. The authors suggested that MANAGER might be an appropriate way to measure the quality of safety management systems and that findings could be combined with traditional quantitative risk assessment studies. No data were provided that could be used to quantitatively assess the reliability of MANAGER.

Williams and Hurst (1992) selected two similar major hazards sites to participate in a study to compare MANAGER audit scores to quantitative risk assessment results and other measures of safety performance. The two sites were practically identical in terms of process and hazards (chlorine and sulfur dioxide hazards at both). Both plants were parts of larger organizations (numerous plants and operations) and for each the day-to-day management operations were organized at the site. Assessment of the safety management systems using MANAGER indicated that one plant was performing slightly better than the industrial average while one was slightly underperforming (MFs of 0.9 and 1.7, respectively).

Quantified risk assessments for each of the two plants were as follows: for the better managed plant, an individual risk of 1×10^{-5} per year at a distance of 290 m; for the somewhat less well-managed plant a risk of 1×10^{-5} at 400 m. Injury rates over a three-year period were also lower in the better-managed plant. The authors suggest that the findings using the MANAGER assessment were in the direction predicted and the difference in MFs was compatible with observed safety performance (Williams and Hurst, 1992).

Although the MANAGER audit was quite widely used in the early 1990s, only the two studies on validity were identified in the review. Both studies were largely descriptive and it is difficult to draw definitive conclusions about MANAGER's validity from the findings. The range of scores reported by Pitblado et al. (1990) indicted that the audit had sufficient variation in responses.

The MANAGER technique was based on an extensive review of accident causation and best practices in process safety (Pitblado et al., 1990), thus it has substantial content validity. The case studies presented in the two articles provide evidence of construct validity. It was reported that sites that scored well were those operated by companies known to have good process safety management systems and those who scored poorly were more likely to be less proactive in OHS safety (Pitblado et al., 1990). In the two studies,

MANAGER audit scores compared well to QRA findings; although QRA findings are not an ideal outcome measure, their consistency with audit scores does provide evidence of validity for the MANAGER technique.

3.2.1.2 PRIMA

In the early 1990s, investigators from the UK Health & Safety Executive, Four Elements Ltd., VROM Int., and Norks Hydro began developing an auditing tool for the quantitative assessment of process safety management systems (PSMS) (Hurst et al., 1996). Further development of this audit method produced an audit tool called PRIMA (Process Risk Management Audit) which had the following characteristics (Hurst et al., 1996):

- eight key audit areas
  - Hazard review of design
  - Human factors review of maintenance
  - Checking/supervision of maintenance tasks
  - Routine inspection and maintenance
  - Human factors review of operations
  - Checking/supervision of construction/installation
  - Hazard review of operations
  - Checking/supervision of operations
- a model of an ideal PSMS defined by the control and monitoring loop, which covers both the PSMS design, implementation, monitoring, and revision
- a set of four key themes within each audit area
- a question set which provides detailed questions to guide the auditor
- an audit manual which describes the audit methodology and practical aspects of auditing
- a calculation method to generate the modification factor

The PRIMA system produces a range of outputs that include practical recommendations, individual assessments of each of the audit areas, and one quantitative output measure – the modification factor (Hurst et al., 1996). The audit system was developed so that the modification factor was compatible with and could be used to modify generic failure rate data used in QRA.

Hurst et al. (1996) conducted a study at six major hazards sites in four European countries to test the hypothesis that positive safety attitudes and PSMS performance lead to low accident rates and low loss-of-containment incidence rates. The Safety Attitude Questionnaire, containing attitude scales such as “workforce satisfaction”, “safety information” and “safe working procedures”, was distributed to participants at the six sites that were later audited using the PRIMA system. Included in the questionnaire were items asking about the respondent’s involvement in incidents of any kind in the last 12 months. This information was used to calculate each sites self-reported accident rate.

Findings from correlation analyses provided no evidence of relationships between PRIMA results and lost-time injury rates or loss-of-containment rates. A strong correlation between self-reported accident rates and PRIMA results was reported using data from six sites, although no findings from statistical tests were provided. Self-reported accident rates were also correlated with the attitude scale scores, although the strength of the correlations was not as strong as those found in previous investigations (Hurst et al., 1996).

The authors suggested that the self-reported accident rates were the most reliable outcome data available. Lost-time injury and loss-of-containment rates were based on data not consistently defined across the six sites (internal company data). The authors suggested that the strength of the PRIMA methods relates to its sound theoretical and statistical basis. However, they cautioned on its application for purposes of site comparisons across jurisdictions (i.e. cross-national) (Hurst et al., 1996).

Nivolianitou and Papazoglou (1998) reported on the use of a revised version of PRIMA in assessing process safety management systems in two major hazard plants in Greece. The audit team consisted of two engineering researchers with varied experience (between "little" and eight years) in risk assessment. Both attended a two-week workshop on the PRIMA methodology. The authors provided detailed findings for each of the eight audit areas and highlighted the deficiencies found. They concluded that the methodology provided feedback that would substantially reduce process safety risks in the two plants. No quantitative analyses were presented in this paper.

The two articles on the PRIMA method do not provide information on interrater or test-retest reliability. Modification factors using the PRIMA method did vary across six hazard sites providing an indication that the measure has adequate variation in response. Although the authors caution on the use of the audit across jurisdictions, Nivolianitou and Papazoglou (1998) argue that this is not a limiting factor in its use, as the method is based on fundamental process safety management principles. Both articles discussed the usefulness of the written assessments and practical recommendations provided through the PRIMA audit and this provides an indication of the method's content validity. Construct validity was assessed by Hurst et al. (1996) and the findings were generally very positive.

3.2.1.3 Assessment of management of maintenance (not named)

Approximately 30 per cent of fatal accidents in the chemical industry have been linked to maintenance activities (UK Health and Safety Executive as cited in Hale et al., 1998). Hale et al. (1998) studied the management of safety in maintenance activities in the chemical process industry in the

Netherlands and developed a theoretical safety maintenance model. Their model was subjected to peer review by consulting with maintenance experts in five highly reputable organizations known for their outstanding process and maintenance safety management. The theoretical model was then used to derive three assessment instruments for the quality of maintenance management in the process industry: (1) a template to assess available data from reported accidents and incidents; (2) a set of audit questions for an in-depth evaluation of maintenance in eight case study companies; and (3) a questionnaire to be sent to all major hazard plant in the Netherlands to gain an insight into management systems across the country.

Hale et al. (1998) presented the findings from an investigation that applied the three instruments described above. Only findings from the audit component are discussed here. The audit instrument contained questions that covered each of the 16 blocks in the theoretical safety maintenance model. The blocks could be categorized into the following management system levels: policy, planning and procedures, and execution and feedback. Audits of each of the eight companies were carried out by two auditors independently; the duration of audits ranged from one to three days depending on the size of the company (Hale et al., 1998). Audit reports were quite detailed and averaged 12 pages in length. The reports were used as a basis for scoring each of the eight aspects of maintenance safety management. For each of the eight aspects, auditors assigned a score of 0 to 3 in relation to the aspect's application in practice and the company's systematic approach to that aspect. The authors provided mean scores obtained from the auditors for each aspect and noted little discordance between the auditors' scores. No differences in scores greater than one point were observed (Hale et al., 1998).

Hale et al's. (1998) audit of management of safety in maintenance appears to have preliminary findings that its interrater reliability is not unreasonable. No findings or data were provided that would allow reviewers to conduct any quantitative tests of reliability. The content validity of the audit also appears to be high as the audit was derived directly from a newly developed model of management of safety in maintenance that was well supported by past research (this included analysis of data on maintenance accidents).

3.2.1.4 Safety Management Assessment System (SMAS)

As discussed previously, high reliability organizations such as those in the marine sector (e.g. marine terminals, offshore platforms, shipping) have unique challenges in ensuring sustainable, safe operations. Hee et al. (1999), through a review of the literature, identified five characteristics of high reliability organizations that are crucial for long-term avoidance of accidents. These are: 1) process auditing, 2) appropriate reward systems, 3) high standards of quality, 4) appropriate risk perception, and 5) command and control functions. SMAS was developed as a screening method to assess

safety management systems in marine organizations by comparing them to these characteristics of high performing, high reliability organizations.

In their preliminary review prior to the development of SMAS, Hee et al. (1999) noted that the attribution of human and organizational errors to accidents in high reliability organizations has increased as equipment reliability has improved. For example, in marine systems such as offshore platforms, organizational errors are said to cause or contribute to over 80 per cent of accidents (Hee et al., 1999). They suggest that when both human and organizational factors are considered, they account for up to 99 per cent of reported accidents in the marine industry. The SMAS audit focuses on identifying and providing feedback on human and organizational factors that differentiates it from other assessment instruments. Another important feature of the audit is that it incorporates system operators as assessors and thus is, in large part, a self-assessment instrument.

A field test using SMAS consisted of two independent assessment teams completing audits of one marine terminal in California (Hee et al., 1999). The five-day audit assessed 140 "attributes" grouped into seven modules: structure, procedures, organization, operating team, interfaces, equipment/hardware, and environmental. The authors examined differences in scores between the two teams for each of the audit's seven modules. On a scale of one to seven, differences in scores between the auditors were less than 0.3 for six of the seven modules. The score difference for the environmental module was 1.3; a closer examination of the scores for this module, which was based on only four attributes, showed that the assessment teams differed only in their evaluation of how social influences (e.g. the media) may affect operator performance.

The reliability of SMAS was determined by comparing a frequency distribution of score differences (between assessment teams) to two frequency distributions of score differences generated using a Monte Carlo Simulation (Hee et al., 1999). A comparison of the actual and random difference distributions showed that the actual distribution had an over 50 per cent larger (65 to 42) number of zero differences than the randomly generated distributions. Although the p-value was not provided, the authors report "greater consistency than randomness."

From the data provided by Hee et al. (1999) a preliminary, minimal level of interrater reliability of SMAS was established. The authors did note that assessment teams attended the same training course and there were some overlaps in the times that both teams were visiting the marine terminal; these circumstances may have helped ensure the observed level of consistency between raters. Content validity of SMAS was not discussed by Hee et al. (1999) but is expected to be high based on their review of the literature in the area and their conceptual framework. Comments from assessment team

members (all who were well qualified) were used to modify attributes thus increasing content validity of the final audit.

3.2.1.5 I-Risk

Both the MANAGER and PRIMA systems attempt to link QRA to safety management assessment. In contrast, Papazoglou et al. (2003) present an evaluation methodology that integrates a QRA model and a process safety management system model. The I-Risk audit methodology advances the state-of-the art by employing detailed technical models for estimating the frequencies of releases in terms of parameters that characterize the stochastic[12] aspects of performance hardware and humans as well as human systems (i.e. management systems).

The objective of the I-Risk methodology is to "quantify the effect of the safety management system of a hazardous installation on the risk" (Papazoglou et al., 2003). The method integrates three well-supported theoretical models: 1) a technical model containing a probabilistic safety assessment, 2) a detailed management model covering all aspects of a process safety management system, and 3) a management-technical interface model which uses quantitative information from the technical and management models to provide an integrated assessment of risk. A safety management system audit is the instrument used to obtain data for the management model.

Papazoglou et al. (2003) presented the findings from a case study of the application of the I-Risk for assessing the integrated risk of loss containment for an ammonia storage facility. Two auditors visited the facility, reviewed company documents, and interviewed 22 people. The strategy for conducting the audit was refined after meetings between the technical and management teams of I-Risk along with information provided by the company. Each auditor assigned scores from zero to ten for each of 51 topics. Results from the audit were used to develop modification factors that were then applied to the technical parameters derived from the technical model component of I-Risk.

A detailed assessment of the reliability or validity of the I-Risk safety management system audit was not part of the case study presented by Papazoglou et al. (2003). However, the authors reported reasonable concordance in the auditors' ratings. Eightyfive per cent of the scores assigned by the two independent auditors were identical or within plus or minus one point on an 11-point scale (0 to 10) and only three per cent were more that two points apart (Papazoglou et al., 2003).

The I-Risk safety management system audit appears to have reasonable interrater reliability based on the findings presented by Papazoglou et al.

[12] Random or probabilistic.

(2003). The case study was not designed to be an extensive reliability or validation investigation and is weak in terms of providing evidence for this review. Content validity of the I-Risk audit appears to be reasonable although the development of the instrument was not described in detail. Criterion or construct validity was not addressed in the case study (Papazoglou et al., 2003).

3.2.1.6 Summary – Reliability and validity of safety management system auditing of high-hazard and high reliability operations

Surprisingly little research has been published on the validity or reliability of auditing of safety management systems in high-risk operations. Despite the lack of evidence, audits are widely used, and, in fact, are required in numerous regulations governing high-hazard processes. For example, European Union Directive 96/82/EC for the control of major-accident hazards, the so called Seveso II directive, requires that major hazard companies implement auditable safety management systems (Papazoglou et al., 2003). Other countries have similar regulations and best practices in both high reliability organizations and high-hazard industries highly recommend auditable safety management systems.

The eight studies reviewed indicate the low overall quality of evidence on the reliability of auditing of safety management systems for high-risk operations. Most of the studies focused on the development and initial field-testing of the audit methodology. Therefore, it is not surprising the study designs were generally weak. Although a number of the studies reported results of audits conducted by independent assessors (Hale et al., 1998; Hee et al., 1999; Papazoglou et al., 2003), no formal tests of reliability were reported. None of the studies reported on the consistency of audit results over time so no evidence is available on the test-retest reliability of any of the audit instruments.

The studies did provide some evidence for the content validity of the auditing methods. All the audit methods were well supported by past reviews of accident causation and most authors presented theoretical frameworks for the derivation of audit measurements. The comparison of audit findings to other OHS performance indicators was performed in studies of MANAGER (Pitblado et al., 1990) and PRIMA (Hurst et al., 1996). Findings from these comparisons were positive and this provided some evidence of construct validity for these audit instruments. The findings for construct validity should be interpreted with caution because typical indicators of safety performance (e.g. rates of accidents) have limited relevance[13] in the evaluation of safety management systems in high-hazard operations or high reliability organizations.

[13] Relevance is limited because process safety management focuses on high risk aspects of process operations not general OHS.

4. Discussion

4.1 State of the evidence regarding the reliability and validity of audits

The authors found very little literature that examined the reliability and validity of audits. Among the literature discussed here, few had the primary intent to specifically look at the measurement properties of the instruments.

Being a narrative review, the authors did not set out to systematically assess the methodological quality of the literature. However, it can be stated that the quality was not strong. The highest quality work appeared to be associated with graduate theses.

This paucity of literature might result from the literature search not being exhaustive. However, other researchers have had similar experiences. In 1988, Eisner and Leger remarked that, "A thorough search of the scientific literature on occupational safety and health failed to discover any publication evaluating the [ISR] scheme" by academic authorities (p. 143). Dyjack and Levine (1996) said, "The authors have been unable to identify published studies evaluating the accuracy and repeatability of either publicly or privately held occupational health and safety assessment instruments." Two years later, the same research group had a similar statement about audit reliability when their paper on the subject was published (Dyjack et al., 1998, p. 790).

Certainly there are obstacles to conducting validity studies that compare audit scores against a criterion like injury rate. Resource availability is one challenge, since audits often require several days on site. Availability and comparability of criterion data across work sites can sometimes be an issue, especially for the high-hazard processes and high reliability organizations. There are likely fewer obstacles to conducting studies of other measurement properties, such as content validity, interrater reliability and responsiveness.

The review found some reports of audit tools demonstrating their content validity. Other reports were surprisingly lacking in this information, even when the audit tool was their focus. It seems that this issue is not well appreciated in some circles. Perhaps the literature from other fields could offer some guidance (e.g. Ware, 1987).

Interrater reliability was studied in the literature concerned with OHS management audits. Agreement among raters was often surprisingly low. Interrater reliability was studied in only a preliminary manner in the literature concerned with audits of high-hazard processes and high reliability organizations. It raises the question of whether this concept is little known

among experts studying these types of organizations. None of the reviewed articles in either stream of literature considered test-retest reliability.

Audit instrument responsiveness to changes in the OHS program was never studied directly, but some studies provided data that allowed the calculation of effect sizes, which ranged from medium to medium-large (Cohen, 1977).

Construct validity was demonstrated in a couple of studies through a comparison or correlation of audit scores and outcome criteria like injury rates. In others, consistent with prediction, audit scores were shown to increase in response to an OHS intervention. There is room for further attempts at construct validation in the literature. The relationship between audit scores and other measures of organizational OHS performance (e.g. safety climate) could be investigated.

Given the common use of audit instruments, there is ample room in the literature for more information about their measurement properties.

4.2 Limitations of the review

A limitation of this review is that the literature search was not exhaustive and relevant information may have been missed. However, as noted above, others have found the research literature sparse in the area of OHS management audit reliability and validity. On the other hand, the authors are not confident that the evidence available in the research literature on audits on high-hazard processes and high reliability organizations is thoroughly represented here. If these types of audits had been of primary interest to us, the bibliographic databases would have been selected to include more engineering sources.

There might be relevant information situated outside of the research literature, but no attempt was made to access this information.

Another limitation of the review is that it was not a systematic review. The literature search was nevertheless quite thorough on the subject of OHS management audits. The same search strategy was systematically applied to all databases. In contrast to systematic reviews, no attempt was made to systematically assess the quality of the literature or separate the "best evidence" from the pool of limited evidence. However, the authors have flagged a number of the quality issues for particular publications throughout this report.

4.3 Practical implications of review findings

In the case of OHS management system—based audits, the findings raise questions about instruments in common use. It appears that a good deal of effort goes into developing the content for many of the audit tools reviewed. Unfortunately, a lot of this effort is not documented well, so content validity

often remains uncertain. It would be helpful if authors went into more detail about the conceptual models and definitions that guided their work, as well as the process used to draw upon expert opinion.

Given the available findings, it is conceivable that some of the audit instruments in common use have low interrater reliability. This is not a large concern when instruments are used to make a baseline assessment or initial diagnosis of an organization. It is a concern, however, when audits are used to determine whether an organization has met a particular standard, since it could result in the inappropriate withholding or awarding of an accreditation. Low reliability would also be a concern when audits are used to monitor an organization on an ongoing basis, especially since they are carried out infrequently. Poor agreement between the auditors used over time could generate a false picture of the progress being made in an organization.

There has been little study of audit results in conjunction with outcome criteria. A database with both quantitative audit scores and OHS outcomes would provide the basis for the weighting used in scoring different sections of a quantitative audit instrument.

5. Conclusions

There is little published research information on the measurement properties of OHS management audits. The evidence that is available is often weak in quality.

The issue of audit content validity did not seem to be appreciated by many authors. There were cases where the focus of the publication was the audit tool itself, but little information on the basis for the audit tool's content was provided.

Reports of audits being validated using outcome measures like injury rates are rare in the literature. Yet, this is an important approach to audit tool validation. While there are real difficulties in carrying this out in the high-hazard and high reliability processes, the challenges are less formidable in other industries. Analyses of audit results and injury rates could not only help the process of validation, but also assist in audit development. It could provide an empirical basis for the weighting of particular sections of the audit.

Construct validity has also been demonstrated in studies where the audit score was found to increase following an intervention on the OHS management system. There is still the scope for further studies of construct validity, such as a comparison of audit scores and results from employee safety climate surveys.

In the few cases where interrater reliability has been systematically examined in OHS management audits, it has often been found to be surprisingly low. This occurred even when the tool had superior content validity. Low interrater reliability is not a large concern for audits used only for initial diagnostic purposes. It *is* a concern when audits are used to measure ongoing progress in the development of an OHS management system and when they are used to certify a certain level of OHS management system quality. There should therefore be a greater expectation for reports on the reliability of audits used for such purposes − both in research studies and in the "real world."

Bibliography

1. American Industrial Hygiene Association (AIHA): Occupational Health and Safety Management System: An AIHA Guidance Document. Fairfax, VA: AIHA Press; 1996.

2. American Industrial Hygiene Association (AIHA): Occupational health and safety management system performance measurement: A universal assessment instrument. Fairfax, VA: AIHA Press; 1999.

3. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. Journal of Clinical Epidemiology. 1997; 50(1):79-93.

4. Beckmerhagen IA, Berg HP, Karapetrovic SV, Willborn WO. Auditing in support of the integration of management systems: A case from the nuclear industry. Managerial Auditing Journal. 2003; 18(6/7):560-8.

5. Bennett D. Health and safety management systems: liability or asset? Journal of Public Health Policy. 2002; 23(2):153-171.

6. Blackmore GA, Shannon HD. Risk-based safety management auditing. Process Safety and Environment Protection. 1996; 74(B1):38-44.

7. Bunn WB, Pikelny DB, Slavin TJ, Parlkar S. Health, safety, and productivity in a manufacturing environment. Journal of Occupational and Environmental Medicine. 2001; 43(1):47-55.

8. Byrom NT. The assessment of safety management systems using an auditing approach. In: Cacciabue PC, Gerbaulet I, Mitchison N, editors. Safety management systems in the process industry. Proceedings CEC Seminar; 1993 Oct 7-8; Ravello, Italy. Joint Research Centre, Institute for Systems Engineering and Informatics; 1994. p. 150-6.

9. Center for Chemical Process Safety (CCPS). Guidelines for auditing process safety management systems. New York: American Institute of Chemical Engineers; 1993.

10. Clark A. Principles of safety auditing. Fire Engineering. 1999; 152(7):77-63.

11. Cohen, J. Statistical power analysis for behavioural sciences, revised edition. New York: Academic Press; 1977.

12. Collison JE, Booth RT. An evaluation of two proprietary health and safety auditing systems. Journal of Health and Safety 1993; 9:31-38.

13. Cooper D. Safety management system auditing. In: Cooper,D. Improving safety culture – a practical guide. Chichester, UK: John Wiley & Sons Ltd; 1998. p. 144-176.

14. Dalrymple H, Redinger C, Dyjack D, Levine S, Mansdorf Z. Occupational health and safety management systems: Review and analysis of international, national, and regional systems and proposals for a new international document. Geneva, International Labour Organization; 1998.

15. Denault WB. The role of the audit. Occupational Health and Safety Canada. 1998; 14(7):42-46.

16. Diekemper RF, Spartz DA. A quantitative and qualitative measurement of industrial safety activities. ASSE Journal. 1970; Dec:12-19.

17. Dyjack DT. Development and evaluation of an ISO 9000-harmonized occupational health and safety management system [dissertation]. Ann Arbor: University of Michigan; 1996.

18. Dyjack DT, Levine SP. Critical features of an ISO 9001/14001 harmonized health and safety assessment instrument. American Industrial Hygiene Association Journal. 1996; 57(10):929-935.

19. Dyjack DT, Levine SP, Holtshouser JL, Schork MA. Comparison of AIHA ISO 9001-based occupational health and safety management system guidance document with a manufacturer's occupational health and safety assessment instrument. American Industrial Hygiene Association Journal. 1998; 59(6):419-429.

20. Dyjack DT, Redinger CF, Ridge RS. Health and safety management system audit reliability pilot project. American Industrial Hygiene Association Journal. 2003; 64(6):785-791.

21. Eisner, HS. Safety rating systems in South African mines. Journal of Health and Safety. 1993; 9:25-30.

22. Eisner, HS, Leger, JP. The international safety rating system in South African mining. Journal of Occupational Accidents. 1988; 10:141-160.

23. Emery FE, editor. Systems thinking: selected readings. Harmondsworth (England): Penguin; 1971.

24. European Process Safety Centre (EPSC). Safety Performance Measurement, Institution of Chemical Engineers (IChemE), 1996; Rugby, UK.

25. Falconer L, Hoel H. Occupational safety and health: A method to test the collection of 'grey data' by line managers. Occupational Medicine. 1997; 47(2):81-89.

26. Frick, K. European Union Legal Standards on Risk Assessment. In: Karwowski W, Rodrick D, Rau P, Horino S, Horino S editors. Handbook of Standards and Guidelines in Ergonomics and Human Factors. In press: New Jersey: Lawrence Erlbaum Publisher.

27. Gay AS, New NH. Auditing health and safety management systems: A regulator's view. Occupational Medicine. 1999; 49(7):471-473.

28. Gillette D, Campbell P, Busby B. The evolution of a radiation safety audit program for a research institution. Radiation Safety Journal. 2004; 86(2)S80-S84.

29. Glendon I. Safety auditing. Journal of Occupational Health and Safety, Australia and New Zealand. 1995; 11(6):569-75.

30. Glendon AI, McKenna EF. Human safety and risk management. London: Chapman and Hall; 1995.

31. Glendon AI, Boyle AJ, Hewitt DM. Computerized health and safety audit systems. In: Matilla M, Karwowski W, editors. Computer applications in ergonomics, occupational safety and health. Amsterdam: Elsevier; 1992. p. 241-8.

32. Grote G, Künzler C. Diagnosis of safety culture in safety management audits. Safety Science. 2000; 34(1-3):131-150.

33. Guastello, SJ. Some further evaluations of the International Safety Rating System. Safety Science. 1991; 14:253-259.

34. Guastello, SJ. Do we really know how well our occupational accident prevention programs work? Safety Science. 1993; 16:445-63.

35. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Annals of Internal Medicine. 1983; 118(8):622-9.

36. Hale AR. Safety management in production. Human Factors and Ergonomics in Manufacturing. 2003; 13(3):185-201.

37. Hale AR, Heming BHJ, Smit K, Rodenburg FGTh, Van Leeuwen ND. Evaluating safety in the management of maintenance activities in the chemical process industry. Safety Science. 1998; 28(1):21-44.

38. Health and Safety Executive. Successful health and safety management. Sudbury (UK): Health and Safety Executive; 1991.

39. Health and Safety Executive. Successful health and safety management: HSG65. Sudbury (UK): Health and Safety Executive; 1997.

40. Hee DD, Pickrell BD, Bea RG, Roberts KH, Williamson RB. Safety management assessment system (SMAS): a process for identifying and evaluating human and organization factors in marine system operations with field test results. Reliability Engineering and System Safety. 1999; 65:125-140.

41. Henriksson L. Looking for gold. Occupational Health and Safety Canada. 1998; 14(7):48-51.

42. Heinrich HW, Petersen D, Roos N. Industrial accident prevention. A safety management approach. 5th ed. New York: McGraw Hill Book Company; 1980.

43. Hemphill JF. Interpreting the magnitudes of correlation coefficients. American Psychologist. 2003; 58(1):78-80.

44. Heron RJL. Audit and 'Responsible Care' in the chemical industry. Occupational Medicine (Oxford). 1999; 49(6):407-410.

45. Hinkin TR. A review of scale development practices in the study of organizations. Journal of Management. 1995; 21:967-988.

46. Hurst NW, Hankin R, Bellamy LJ, Wright MJJ. Auditing – a European perspective. Journal of Loss Prevention in the Process Industries. 1994; 7(2):197-200.

47. Hurst NW, Ratcliffe K. Development and application of a structured audit technique for the assessment of safety management systems (STATAS). Hazards XII: European advances in process safety; 1994 April 19-21; Manchester, UK. Manchester: UMIST; 1994. p. 315-339.

48. Hurst NW, Young S, Donald I, Gibson H, Muysellar A. Measures of safety management performance and attitudes to safety at major hazard sites. Journal of Loss Prevention in the Process Industries. 1996; 9(2):161-172.

49. Jorgensen, EB. Safety and health auditing: A misunderstood process. Professional Safety. 1998; 43(4):29-32.

50. Karapetrovic S, Willborn W. Quality assurance and effectiveness of audit systems. The International Journal of Quality & Reliability. 2000; 17(6):679.

51. Karapetrovic S, Willborn W. Generic audit of management systems: Fundamentals. Managerial Auditing Journal. 2000; 15(6):279-294.

52. Karapetrovic S, Willborn W. Audit system: Concepts and practices. Total Quality Management. 2001; 12(1):13-28.

53. Kennedy R, Kirwan B. Development of a hazard and operability-based method for identifying safety management vulnerabilities in high-risk systems. Safety Science. 1998; 30(3):249-274.

54. Kirwan B. Validation of human reliability assessment techniques: Part 2 – Validation results. Safety Science. 1997; 27(1):43-75.

55. Kuusisto A. Safety management systems: Audit tools and reliability of auditing. [dissertation] Tampere (Finland): Tampere University of Technology; 2000.

56. LaMontagne AD, Barbeau E, Youngstrom RA, Lewiton M, Stoddard AM, McLellan D, et al. Assessing and intervening on OSH programmes: Effectiveness evaluation of the Wellworks-2 interventions in 15 manufacturing worksites. Occupational & Environmental Medicine. 2004; 61(8):651-660.

57. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 159-74.

58. Laitinen H, Marjamaki M, Paivarinta K. The validity of the TR safety observation method on building construction. Accident Analysis and Prevention. 1999; 31:463-472.

59. Le Coze J-C. Are organisations too complex to be integrated in technical risk assessment and current safety auditing? Safety Science. 2005; 43(8): 613-638.

60. Leighton C. Auditing and complacency. The Safety & Health Practitioner 1998; Nov:65-66.

61. Levine SP, Dyjack DT. Critical features of an auditable management system for an ISO 9000 - compatible occupational health and safety standard. American Industrial Hygiene Association Journal. 1997; 58(4):291-298.

62. Lindsay FD. Successful health and safety management. The contribution of the management audit. Safety Science. 1992;15:387-402.

63. Lipsey MW. A scheme for assessing measurement sensitivity in program evaluation and other applied research. Psychological Bulletin. 1983; 94(1):152-65.

64. McDowell I, Newell C. Measuring health. A guide to rating scales and questionnaires. New York: Oxford University Press; 1987.

65. Nivolianitou ZS, Papazoglou IA. An auditing methodology for safety management of the Greek process industry. Reliability Engineering and System Safety 1998; 60(3):185-197.

66. Occupational Safety & Health Administration, U.S. Department of Labor [homepage on the Internet]. Washington: U.S. Department of Labor. Program evaluation profile (PEP) [cited 7 Dec 2005]. Available from: http://www.osha.gov/SLTC/safetyhealth/pep.html.

67. Papazoglou IA, Bellamy LJ, Hale AR, Aneziris ON, Ale BJM, Post JG, et al. I-risk: development of an integrated technical and management risk methodology for chemical installations. Journal of Loss Prevention in the Process Industries. 2003; 16(6):575-591.

68. Pearse W. Club Zero: Implementing OHSMS in small to medium fabricated metal product companies. Journal of Occupational Health and Safety – Australia New Zealand. 2002; 18(4):347-356.

69. Petersen D. Techniques of safety management. A system approach. 3rd ed. New York: Aloray Inc; 1989.

70. Pitblado R, Williams JC, Slater DH. Quantitative assessment of process safety programs. Plan/Operations Progress. 1990; 9(3):169-175.

71. Rainer D., Kretchman K, Cox J. The power of environmental health and safety audits. Chemical Health and Safety. 2000; May/June: 20-25.

72. Ratcliffe KB. (1993a). STATAS: development of an HSE audit scheme for loss of containment incidents. Part 1: A loss of containment model. Loss Prevention Bulletin. 1993; 112:1-6.

73. Ratcliffe KB. (1993b). STATAS: development of an HSE audit scheme for loss of containment incidents. Part 2: Management and organizational factors: A socio-technical model of accident causation. Loss Prevention Bulletin. 1993; 113:15-24.

74. Ratcliffe KB. (1993c). STATAS: Development of an HSE audit scheme for loss of containment incidents. Part 3: Constructing an audit scheme. Loss Prevention Bulletin. 1993; 114:21-27.

75. Reason J. A systems approach to organizational error. Ergonomics. 1995; 38(8):1708-1721.

76. Redinger CF. Occupational health and safety management system conformity assessment: Development and evaluation of a universal assessment instrument. [dissertation]. Ann Arbor: University of Michigan; 1998.

77. Redinger CF, Levine SP. Development and evaluation of the Michigan occupational health and safety management system assessment instrument: a universal OHSMS performance measurement tool. American Industrial Hygiene Association Journal. 1998; 59:572-581.

78. Redinger CF, Levine SP, Blotzer MJ, Majewski MP. Evaluation of an occupational health and safety management system performance measurement tool-II: scoring methods and field study sites. American Industrial Hygiene Association Journal. 2002a; 63(1):34-40.

79. Redinger CF, Levine SP, Blotzer MJ, Majewski MP. Evaluation of an occupational health and safety management system performance measurement tool-III: Measurement of initiation elements. American Industrial Hygiene Association Journal. 2002b; 63(1):41-6.

80. Robson L, Clarke J, Cullen K, Bielecky A, Severin C, Bigelow P, el al. The Effectiveness of occupational health and safety management systems: A systematic review. Toronto: Institute for Work & Health; 2005.

81. Roughton J. Process safety management: An implementation overview. Professional Safety. 1993; 38(8):28-33.

82. Schweigert MK, House RA, Holness DL. Occupational health and safety management systems in the Canadian Pulp and Paper Industry: Methods of auditing. Journal of Occupational & Environmental Medicine. 1999; 41(10):857-862.

83. Stewart, AL, Ware, JE, Jr., editors, Measuring functioning and well-being: The medical outcomes study approach. Durham: Duke University Press; 1992.

84. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. New York: Oxford University Press; 1995.

85. Tinmannsvik RK, Hovden J. Safety diagnosis criteria - Development and testing. Safety Science. 2003; 41(7):575-590.

86. Uusitalo T, Mattila M. Evaluation of industrial safety practices in five industries. In: Mital A, editor. Advances in industrial ergonomics and safety, vol.1. London: Taylor & Francis; 1989. p. 353-358.

87. Wang, Y. Development of a computer-aided fault tree synthesis methodology for quantitative risk analysis in the chemical process industry. [dissertation] USA: Texas A&M University; 2004.

88. Ware, JE Jr. Standards for validating health measures: definition and content. Journal of Chronic Diseases. 1987; 40(6):473-80.

89. Williams JC, Hurst NW. A comparative study of the management effectiveness of two technically similar major hazard sites. In: Major hazards onshore and offshore: A three day symposium; 1992 Oct 20-22; Manchester, UK. Manchester: UMIST; 1992. p. 73-83.

Appendix: Organizing structure of OHS management audits

Diekemper and Spartz (D&S) method (Diekemper and Spartz, 1970; Kuusisto, 2000)	**MISHA (**Kuusisto, 2000)	**IRS system, 4th edition** (Collison and Booth, 1993)
Organization and administration Industrial hazard control Fire control and industrial hygiene Supervisory participation, motivation and training Accident investigation, statistics and reporting procedures	Organization and administration Safety policy Safety activities in practice Personnel management Training and motivation Safety training of the personnel Work instructions Incentives to safe work practices Communication Work environment Physical work environment Psychological work environment Analysis of the work environment Follow-up Occupational illnesses Occupational accidents Occupational diseases Work ability of the personnel Social work environment	Leadership and administration Management and training Planned inspections Job/task analysis and procedures Accident/incident investigation Job/task observations Organizational rules Emergency preparedness Accident/incident analysis Employee training Personal protective equipment Health control and services Program evaluation system Purchasing & engineering controls Personal communications Group meetings General promotion Hiring and placement Records and reports Off-the-job safety

IRS system for mining (Eisner and Leger, 1988)	**CHASE-II, 4th version** (Collison and Booth, 1993)	**Adaptation of OSHA's Program Evaluation Profile** (LaMontagne et al., 2004)
Leadership and administration	Management of legal requirements and resources	Management commitment and employee participation
Management training	Management of tools, equipment, fixtures and fittings	Workplace analysis
Planned inspections	Management of machinery and plant	Hazard prevention and control
Rules and regulations	Management of chemicals and substances	OSH training and education
Accident/incident investigation	Management of vehicles	
Accident/incident analysis	Management of energy	
Emergency preparedness	Management of health	
Care of the injured and ill	Management of tasks	
Task analysis and procedures	Management of people	
Skill training	Monitoring and feedback for health and safety	
Planned task observations	Management of change	
Protective equipment	Management of emergencies and special cases	
Program monitoring system		
Group meetings		
Off-the-job safety		
Purchasing and engineering controls		
General promotion		
Physical capability screening and monitoring		
Physical conditions		
Compliance with recommended safe practices		

Goodyear Tire and Rubber Company audits (Dyjack et al., 1998)	**Modified version of Goodyear Tire and Rubber Company audit** (Bunn et al., 2001)	**AS/NZS 4804-based audit for small- and medium-sized organizations** (Pearse, 2002)
Safety Systems Management organization, participation, and administration Record keeping/statistics/files Incident investigation Operator work rules and procedures Operator safety training Safety compliance/control of hazardous energy Safety management development Communications Ergonomics Approval of machines, equipment, materials, and processes General safety promotion Off-the-job safety Proactive safety systems/processes *Industrial Systems* Leadership and administration Industrial hygiene activities Hazardous communications Noise Personal protective equipment Confined space Respiratory protection Special procedures Medical services Ventilation Radiation Heat stress	Management organization, participation, and administration Training, compensation Operation procedures and programs Compliance programs Risk reduction, continuous improvement Medical/first-aid services Health promotion and wellness General walk-around observations	Management commitment and policy Responsibility and accountability Risk management Purchasing and contractors OHS training and education Emergency planning Performance indicators and records Workplace injury management

AIHA ISO 9001 harmonized OHS management system (Dyjack et al., 1998)	**AIHA Universal OHSMS performance assessment tool** (Redinger and Levine, 1998)
OHS management responsibility	Management commitment and resources
OHS management systems	Regulatory compliance and system conformance
OHS compliance and conformance review	Accountability, responsibility, and authority
OHS design and control	Employee participation
OHS document and data control	Occupational health and safety policy
Purchasing	Goals and objectives
OHS communication systems	Performance measures
OHS hazard identification and traceability	System planning and development
Process control for OHS	Baseline evaluation and hazard/risk assessment
OHS inspection and evaluation	OHSMS manual and procedures
Control of OHS inspection, measuring and testing equipment	Training system
OHS inspection and evaluation status	Technical expertise and personnel qualifications
Control of nonconforming OHS processes and devices	Hazard control system
OHS corrective and preventive action	Process design
Handling, storage, and packaging of hazardous materials	Emergency preparedness and response system
Control of OHS records	Hazardous agent management system
Internal OHS management system audits	Preventive and corrective action system
OHS training	Procurement and contracting
Operations and maintenance services	Communication system
Statistical services	Document and record management system
	Evaluation system
	Auditing and self-inspection
	Incident investigation and root cause analysis
	Medical program and surveillance
	Continual improvement
	Integration
	Management review