SYSTEMATIC REVIEW


Lynda S. Robson, PhD,1 Philip L. Bigelow, PhD1,2

ABSTRACT

Objective: The measurement properties of Occupational Health and Safety (OHS) management audits might be important in some applications, especially when audit scores are treated as performance measures. The review, therefore, sought to identify and summarize the research evidence on the measurement properties (e.g., reliability, validity) of methods of OHS management audit.

Methods: Bibliographic databases in business, medicine and OHS were systematically searched. Evidence from relevant publications was synthesized using traditional narrative review methods.

Synthesis: The literature on the measurement properties of OHS management audit methods is sparse. Seventeen relevant audit methods were identified. Content validity was demonstrated for only five audit methods. Inter-rater reliability was formally tested for only three audit methods and construct validity for only one. There were no studies of test-retest reliability or responsiveness. The investigations of inter-rater reliability (i.e., consistency among auditors) showed that it is often unacceptably low.

Conclusion: There is a research gap concerning the measurement properties of OHS management audit methods. The available research raises questions about the properties of audit methods in current use.

Key words: Accident prevention; management audit; occupational health; reproducibility of results; safety management; validation studies

La traduction du résumé se trouve à la fin de l’article.


An occupational health and safety (OHS) management system is designed to protect the health of workers by the following means: designating roles and responsibilities related to OHS; setting organizational targets and objectives related to OHS; planning and establishing the maintenance of hazard controls; and monitoring, reviewing and improving the system’s implementation and effectiveness. Federal and provincial legislation (e.g., the Occupational Health and Safety Act in Ontario) specifies simple management systems applicable to all workplaces. Exemplary organizations also seek compliance with voluntary standards and guidelines1-3 that are more comprehensive.

Auditing is a means of directly and comprehensively monitoring the implementation and effectiveness of a firm’s OHS management system.4 The auditing process typically involves the following:5 gathering evidence about the management system through interviews, documentation reviews and work site observations, guided by an audit instrument, evaluating the gathered evidence, and providing a summary of the evaluative findings. Auditing is an important component of an OHS management system,1,3 and up to 95% of Fortune 2000 companies perform audits.5 While less prevalent in smaller firms, audits are recognized as nevertheless relevant.7

Depending on the type of application in which a particular OHS management audit is involved, researchers and practitioners might want to consider its measurement properties (i.e., reliability, validity, etc.). In some cases, measurement properties are relatively less important. For example, simple audits are sometimes used to assess the management and programmatic needs of organizations that are just starting to develop their OHS management systems.8 Similarly, there are firms in which management systems are more developed but in which audits are used periodically only to ensure that there are no major gaps in the management system. In these two situations, only a blunt measurement instrument is needed, though one would want some assurance of its content validity: one would want to be confident that the key elements of an OHS management system, appropriate for the organization, are adequately represented in the content of the audit instrument, otherwise it cannot serve its function of identifying gaps. In contrast to the previous examples, additional measurement properties (e.g., inter-auditor reliability, predictive validity, responsiveness) might also be quite important when audits are used in performance measurement applications. These include benchmarking, determining whether a particular standard has been met or monitoring progress over time. Measurement properties may be additionally important in such applications when the results of audits determine organi-

Author Affiliations

1. Institute for Work & Health, Toronto, ON
2. University of Waterloo, Waterloo, ON

Correspondence and reprint requests: Lynda Robson, PhD, Institute for Work & Health, 481 University Ave., Suite 800, Toronto, ON M5G 2E9, Tel: 416-927-2027, ext. 2164, Fax: 416-927-4167, E-mail: lrobson@iwh.on.ca

Acknowledgement: The Institute for Work & Health receives core funding from the Workplace Safety & Insurance Board of Ontario (WSIB). The review was also supported in part by the WSIB Prevention Reviews initiative. The authors appreciate the key support of the library staff, particularly Emma Irvin and Rachel Couban, and the administrative assistance of Diana Pugliese.

Conflict of Interest: None to declare.
zational rewards, e.g., financial,9-12 recognition,13 contract opportunities10,12 or even penalties.14 To the extent that the data provided by the audit of a firm are not reliable, valid or responsive, organizational actions based on the audit report might be misdirected.

Our original intent was to review the research evidence on the measurement properties of OHS management audits using the methods of systematic literature reviews. However, initial screening revealed a lack of published studies, and we therefore only used systematic review methods for searching and screening the literature, and then used traditional review methods to assess and synthesize the evidence.

METHODS

Literature search
The search strategy was developed in MEDLINE, starting with the Medical Subject Heading (MeSH) terms assigned to known, relevant articles. This search strategy was applied to five other bibliographic databases, adaptations being made as required: American Business Inform (ABI), CCInfoWeb, Econlit, EMBASE and Health and Safety Science Abstracts. ABI and CCInfoWeb catalogue grey literature sources, as well as those that are peer-reviewed. The databases were searched from their point of inception until September 2008. One search strategy looked for abstracts classified with the MeSH term “Management Audit” AND one of the following terms: “Wounds and Injuries”, “Accidents, Occupational”, “Accident Prevention” or “Occupational Health”. The second, more fruitful, strategy looked for abstracts classified with the MeSH term “Safety Management” AND the free text term “Audit.” No restrictions were placed regarding date and language of the original publication. In order to broaden the search, the following sources were also used: reference sections of publications deemed relevant, personal files of the authors, and an in-house bibliographic database assembled for a systematic review of OHS management system effectiveness.16

Relevance screening
The titles and abstracts arising from the searches were reviewed to identify potentially relevant publications using two inclusion criteria:
- The publication is a journal article, book, conference proceeding, dissertation or report,
- The publication contains information on any of the following measurement properties of OHS management audits: content validity, construct validity, inter-rater reliability, test-retest reliability or responsiveness.

Nine references17-25 were eliminated when the following exclusions were applied:
- Publication is a magazine article or newsletter
- Audit focuses on hazards rather than on management17,18
- Audit focuses on the management of a particular type of OHS hazard instead of all OHS hazards19
- Audit is a safety management system audit, which does not focus on OHS20-25

The last exclusion pertains to a distinct stream of research and practice focused on the prevention of out-of-control processes or catastrophic events. Some of this literature was reviewed in an earlier report,26 and the basic conclusion drawn from it is similar to the one we draw for this review.

The review of the titles and abstracts was shared by the authors, each title and abstract being reviewed by a single author. Potentially relevant publications were retrieved and reviewed in more detail.

Evidence extraction and synthesis
Both authors were involved in the data extraction and evidence synthesis. Each author read all retrieved publications, discussed and developed a common understanding of the findings, and agreed upon which publications did not meet the criteria; they shared the extraction and synthesis of the evidence.

RESULTS

Table 1 summarizes the features of 17 distinct audit methods found in the relevant publications. Several well-known audit methods were identified through the review: the Diekemper and Spartz method,27-29 the International Safety Rating System (ISRS)30-35 and the CHASE audits.33,36 Another two were associated with the American Industrial Hygiene Association.37,38,41-45 Several of the audits were intended for multiple sectors of the economy, whereas others target a single sector. The number of items in each audit instrument is in the range of less than 100 to several hundred. Some methods require the auditor to respond to an item by indicating yes or no; other methods allow a greater variety of responses. Less commonly, auditors are asked to assign a number of points out of a maximum possible number. The most common way of summarizing the overall audit results is a percent score (out of 100).

For eight of the methods listed in Table 1, either we considered the content validity to have been evidenced, or a formal test of reliability or validity had been reported. This evidence is summarized in Table 2 and is discussed further below. For the remaining nine audit methods, only information of a more preliminary nature was available, and they are not discussed further.

Content validity
In the present context, content validity56 refers to the comprehensiveness of the audit instrument in its representation of management system concepts. There are five methods for which content validity is demonstrated. The first17,18 had ISO 9001 as the organizing framework and drew its content from several OHS and environmental management system documents. A second method,41-45 also developed at the University of Michigan, defined a “universe” of OHS management system elements. Researchers reviewed 13 OHS and environmental standards or guidance documents and selected four that collectively represented the content of all 13. These four “input models” were deconstructed and then reorganized into an integrative model. A third method was developed in collaboration with the Canadian Pulp and Paper Association6 using a matrix approach to the audit content: 11 OHS activity areas (e.g., emergency responsiveness preparation, health surveillance) were each considered in relation to the presence of key management system components (goals and procedures, assignment of responsibility, performance measurement, review of performance, corrective action). A fourth method, developed in Norway for the mining sector,39,40 was intended for consensus-based, high-level, self-audits by organizational representatives. The final example comes from a Singapore project in the construction sector,34 in which the steps involved included reviewing existing methods, generating items,
### Table 1. Audit Methods Included in the Review

<table>
<thead>
<tr>
<th>Name of Audit Method*</th>
<th>Target Population</th>
<th>Conceptual Basis</th>
<th>Number of Items</th>
<th>Nature of Response Options</th>
<th>Form of Final Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diekemper &amp; Spartz method27-29</td>
<td>Manufacturing</td>
<td>Unknown</td>
<td>29</td>
<td>1 to 4 (poor to excellent)</td>
<td>Percent (of total possible points)</td>
</tr>
<tr>
<td>International Safety Rating System (ISRS) – Mining30-32</td>
<td>Mining operations in South Africa</td>
<td>Unknown</td>
<td>“almost 1000”31</td>
<td>Mainly yes/no; some 0-20 scales</td>
<td>Numerical score and star rating. Star rating has five standard levels (no. of stars 1-5) and five advanced (no. of stars 1-5). Each level involves a different number of elements.</td>
</tr>
<tr>
<td>ISRS – Generic, 4th ed.32-35</td>
<td>General industry, with first version developed in the steel industry (Eisner &amp; Leger46)</td>
<td>Unknown</td>
<td>From about 100 to 627, depending on the star level attempted35</td>
<td>Mainly yes/no; some 0-20 scales</td>
<td>See ISRS – Mining</td>
</tr>
<tr>
<td>CHASE series33,36</td>
<td>CHASE-I and -II: general industry (&lt;100 and 100+ employees); Construction-CHASE: all sizes</td>
<td>Based on review of regulations (UK) and other sources, professional judgment and input from pilot testing.</td>
<td>CHASE-I: 200; CHASE-II: 400; Construction-CHASE: 350</td>
<td>Yes/no (but software can also accept a graded response such as %)</td>
<td>Percent (of total possible points) overall and by section</td>
</tr>
<tr>
<td>Safety Element Method39,40</td>
<td>Norwegian mining</td>
<td>Principles of safety management and quality assurance. Action research process used to develop content, with experts referring to ISRS and other existing tools.</td>
<td>6 elements + 12 subelements</td>
<td>Each subelement and element scored from stage 1 to stage 5 (stage 1 = no formal system; stage 2 = minimum regulated standards; stage 5 = fully developed)</td>
<td>Categorical assessments for each subelement and element. Also, can convert results to score by summing items, with item score corresponding to the stage number.</td>
</tr>
<tr>
<td>AIHA ISO 9001 harmonized37,38</td>
<td>All sectors</td>
<td>ISO 9001</td>
<td>135 clauses (determined by review authors based on Dyjack37)</td>
<td>In conformance, minor non-conformance, major non-conformance</td>
<td>Qualitative: Non-conformance reports that identify deviation from standard and its root cause</td>
</tr>
<tr>
<td>Goodyear Tire and Rubber Co. – Industrial Health Review and Corporate Safety Systems Review; 1995 version37,38</td>
<td>Manufacturing</td>
<td>Unknown</td>
<td>At least 191 (based on the combined number of subelements in the two reviews)</td>
<td>Each subelement worth a number of points. Auditor awards full or partial points.</td>
<td>Percent (of total possible points). Also, list of significant items requiring action.</td>
</tr>
<tr>
<td>AIHA Universal OHSMS Assessment Instrument41-45</td>
<td>All sectors</td>
<td>Reconstructed from four input models, using system theory and policy analysis models for organizing framework.</td>
<td>486</td>
<td>Ordinal score for each measurement criterion (0-5) in most recent version41</td>
<td>Average score (0-5)</td>
</tr>
<tr>
<td>Safety Management Audit for Construction (SMA-CON)46</td>
<td>Construction in Hong Kong</td>
<td>Based on British Standard 8800:1996 and elements specified by Labour Department.</td>
<td>456 (less for firms with &lt;10 employees)</td>
<td>Yes/no and assigning points out of 20 or 30.</td>
<td>Percent (of total possible points)</td>
</tr>
<tr>
<td>Canadian Pulp and Paper Association47</td>
<td>Canadian pulp and paper</td>
<td>General OHS practices/principles and management system principles. 11 OHS activities assessed with respect to 5 management system elements.</td>
<td>55</td>
<td>Score 0 or 1</td>
<td>Scores summed by activity or by management system component</td>
</tr>
<tr>
<td>Water utilities benchmarking audit48</td>
<td>All sectors</td>
<td>Principles from British Standards Institute and the Health and Safety Executive</td>
<td>56</td>
<td>Scores from 1 (low) to 4 (high)</td>
<td>Percent (of total possible points) overall and by section</td>
</tr>
<tr>
<td>Method for Industrial Safety and Health Activity Assessment (MISHA)48</td>
<td>Manufacturing</td>
<td>Framework based on Booth and Lee's49 key elements of safety management</td>
<td>55</td>
<td>Scores from 0 (not acceptable) to 3 (no weakness; strong improvement process in place)</td>
<td>Percent (of total possible points) overall and by section</td>
</tr>
<tr>
<td>AS/NZS 4804-based50,51</td>
<td>Small and medium enterprises in metal prefabrication</td>
<td>Based on Australian/New Zealand standard, AS/NZS 4804</td>
<td>80</td>
<td>Yes=1; no=0; partial=0.5</td>
<td>Percent (of total possible points) overall and by section</td>
</tr>
</tbody>
</table>

---

...continues...
<table>
<thead>
<tr>
<th>Audit Method</th>
<th>Content Validity*</th>
<th>Inter-auditor Reliability†</th>
<th>Construct Validity‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>D&amp;S method†</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRS – Mining‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIHA ISO 9001 harmonized‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Safety Element Method‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIHA Universal OHSMS Assessment Instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canadian Pulp and Paper Association‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MISHA§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction Safety Index lighten</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Classification of the evidence on content validity: +, content validity is considered adequate by the review authors because the method of developing the content for the audit instrument is well described and has involved appropriate stakeholders; –, formal test of content validity in the research literature concludes that the audit method is not valid. Cell is left blank when the available information is not sufficient for reviewers to assign either + or –.

† Classification of the evidence on inter-rater reliability: +, test(s) of inter-rater reliability yield a Kappa statistic that is 0.41 or more (categorical data) or an intraclass coefficient that is 0.75 or more (continuous data); –, test(s) of inter-rater reliability yield a Kappa statistic that is less than 0.41 or an intraclass coefficient that is less than 0.75. Cell is left blank when no test has been conducted.

‡ Classification of the evidence on construct validity: +, the majority of hypotheses constructed about the expected relations between audit data and other OHS measure(s) in a test of construct validity are supported; –, the majority of hypotheses constructed about the expected relations between audit data and other OHS measure(s) in a test of construct validity are not supported. Cell is left blank when no tests of construct validity have been conducted.

Inter-rater reliability

Inter-rater reliability is the consistency of assessments of the same workplace(s) by different auditors. For categorical data, agreement is best expressed using kappa or weighted kappa (κw), which should minimally be 0.41.58,59 For continuous data, intra-class coefficient statistics are used, with values of 0.75 considered minimal.59 Inter-rater reliability was investigated for three audit methods as part of thesis dissertations.

Kuusisto28 studied the reliability of a version of the Diekemper and Spartz method.27 He found poor reliability when his own ratings were compared with those of local company evaluators (κw of -0.03 to 0.46). Agreement between raters was better when Kuusisto’s ratings were compared with those of his safety specialist students (κw of 0.36 to 0.83). Kuusisto28 also examined the reliability of his own Method for Industrial Safety and Health Activity Assessment by comparing his ratings with those made by members of the company being audited. Weighted kappas were all less than 0.41, so the tool was revised. A test of the new version, which had modified guidance for decision making, showed improvement (κw of 0.38 to 0.58).

Dyjack et al.41 looked at the reliability of a portion of the Michigan Universal Assessment Instrument. The agreement between raters was considered inadequate by several statistical criteria, in spite of the expert qualifications of the two auditors in the project, the similarity of their substantial preparation to use the audit and the content validity of the instrument.

Construct validity

The construct validity of an audit method is established by making hypotheses about the expected relations between audit scores and other measures of OHS, testing the relations using appropriate methods and then confirming the majority of the expectations.56,57 There is only one study of construct validity in the literature on OHS audits that uses statistical analyses. Eisner and Leger30 examined the correlation of the number of stars awarded by the ISRS audit with each of fatality rate and reportable injury rate. Correlations were small, not all in the
expected direction, and none were statistically significant. However, the study had significant methodological limitations.

Other measurement properties
There were no formal investigations found of test-retest reliability (consistency of results upon repeat administration of the audit) or responsiveness (ability to accurately detect change over time).36,57

DISCUSSION
Principal findings
The research literature on the measurement properties of OHS management audits is sparse. There has been some investigation of inter-rater reliability, which showed that it was often unacceptably low according to usual standards.

Strengths and limitations of the review
This is the first literature review about the measurement properties of OHS management audits. An extensive, systematic search of the literature was conducted. We are confident that our portrayal of the research literature as sparse in this area is valid. We note, however, that there may be relevant information, particularly on proprietary methods, in sources outside of the research literature. Although this additional information would expand upon what is known from the research literature, it is not feasible to collect it in a systematic way.

Review findings in relation to other research
There are no other reviews of this nature in the research literature, preventing comparison with a similar study. However, others have commented on the paucity of available research. In 1988, Eisner and Leger10,11 remarked that, “A thorough search … failed to discover any publication evaluating the [ISRS] scheme” by academic authorities, even though the audit method was internationally recognized at the time. One decade later, Dyjack12,13,14 said he was “unable to identify published studies evaluating the accuracy and repeatability of either publicly or privately held occupational health and safety assessment instruments.”

There is an additional study of inter-auditor reliability known to the authors but not included in the review because it has only been reported as a conference abstract.56 It found acceptable levels of agreement between two qualified auditors for only one of three organizational units audited.

Practical implications of review findings
On the basis of our findings, we expect that some audit instruments in common use have low inter-rater reliability. This could be a concern when various auditors conduct audits within an auditing program and the audit scores are used as a performance measure. Low inter-rater reliability under these circumstances might lead, for example, to inconsistent classification of whether firms meet the performance benchmark, erroneous ranking of firms in intra- and inter-firm comparisons, and erroneous trend information. Our findings indicate a lack of information on the construct validity and responsiveness of audit scores. Researchers, therefore, still need to establish the extent to which decision makers should rely on audit scores as an indicator of firm OHS performance. Our recent study of audit methods8 indicates that practitioners use them for performance measurement purposes, even in the absence of formal knowledge about their measurement properties. We encourage a practitioner who relies on an audit method for such purposes to seek more information about its measurement properties and in the absence of that information be cautious in their decisions based on audit results.

Future research
This review points to a large research gap regarding the measurement properties of OHS management audits. There is also little empirical information on the extent to which various factors affect these properties, since there are only single case reports58-61 and prescriptive information from experts.62-67 According to these and the research literature on financial audits,68-71 we would expect determining factors to fall into the following categories: 1) auditors (e.g., their independence, their expertise), 2) auditing programs (e.g., training processes, quality control processes) and 3) the auditing method (e.g., content, usability of the audit instrument, number of response categories in the audit instrument, explicitness of decision aids for auditors). Further, it would be interesting to know, through an intervention study in the field, the degree to which measurement properties can be improved. On the other hand, the degree to which such change would affect decision making and at what cost also remains to be determined.

There are many practical challenges involved in research on management audits: they are labour intensive for the auditing organization and the workplace, the methods involved in studying audits (e.g., presence of a researcher or a second auditor) could have an impact on the audit results that is difficult to measure, sample size may be constrained by the volume of an audit program, and other data sources available for construct validity studies (e.g., injury reports) may be of poor quality. However, the studies involved in this review provide some examples of what is possible. OHS researchers could also look to the clinimetrics field for further guidance in and illustrations of high-quality measurement research.56,77,59,78-80

REFERENCES
RÉSUMÉ

Objectif : Les propriétés de mesure des audits de gestion en santé et sécurité du travail (SST) pourraient être importantes pour certaines applications, surtout lorsque les scores d’audit sont considérées comme des mesures de rendement. Nous avons donc voulu répertorier et résumer les résultats de recherche sur les propriétés de mesure (p. ex., fiabilité, validité) des méthodes utilisées pour les audits de gestion en SST.

Méthode : Les bases de données bibliographiques des domaines des affaires, de la médecine et de la SST ont été systématiquement interrogées. Les données probantes de publications pertinentes ont été résumées à l’aide de méthodes classiques de recension narrative des écrits.

Synthèse : Les travaux publiés sur les propriétés de mesure des méthodes utilisées pour les audits de gestion en SST sont rares. Dix-sept méthodes d’audit pertinentes ont été recensées. Nous n’avons pu démontrer la validité de contenu que pour cinq de ces méthodes. La fiabilité inter-évaluateurs n’a été véritablement testée que pour trois méthodes d’audit, et la validité de construit, pour une seule méthode. Il n’y avait aucune étude de fiabilité de test-retest, ni de sensibilité. Les études de fiabilité inter-évaluateurs (cohérence d’un évaluateur à l’autre) ont montré que cette fiabilité est souvent trop faible pour être acceptable.

Conclusion : Il y a des lacunes dans la recherche sur les propriétés de mesure des méthodes de gestion en SST. Les travaux publiés soulèvent des questions quant aux propriétés des méthodes d’audit utilisées actuellement.

Mots clés : prévention des accidents; audit de gestion; santé au travail; reproductibilité des résultats; gestion de la sécurité; études de validation