Safety Auditing

Applying research methodology to validate a safety audit tool

By Yueng-Hsiang Huang and Stanford A. Brubaker

SH&E MANAGERS ARE OFTEN ASKED—and even ask themselves—"How well does the organization's risk reduction and safety plan work? How can we be certain that [any specific safety activity] will contribute to overall program improvement and reduction in losses?"

Traditional approaches to benchmarking risk management and safe work performance with incident rate reduction have failed to provide a strong connection between program causes and direct factors influencing risk and loss frequency. Benchmarking, by definition, provides a comparative baseline to serve as a "stake in the ground" for basing various measurements or systems of evaluation. For example, many businesses illustrate comparative progress by measuring reduction in workplace injuries (a frequency metric) or lost-time per case (a severity metric), often judging their performance against other businesses within the same industry or SIC.

Workplace audits are a common form of measuring an organization's safety performance. Many traditional forms of audits emphasize compliance activity and are often designed to provide a relative measure of comparison to federal OSHA or related requirements. But how effective is that process in determining the ultimate value of an overall safety program? Are the activities examined those that will have the greatest impact? While benchmarking can be a useful tool, it demonstrates a historic trend that often does not reflect the specific methods or processes a program will need to effectively implement and sustain long-term improvement. A measurement system with direct, proven links to causation is an essential tool to stimulate comprehensive organizational change.

Workplace audits and assessments can have various forms and purposes. Determining the appropriate dimensions or topics can render a wide variety of conclusions, each of which may or may not achieve the best result, or measure those dimensions that have the greatest positive effect. Although many off-the-shelf audits are available, in some cases the best choice is an audit designed specifically for a given application, based on what the users want to measure, what they ultimately wish to learn and how they will use what they learn.

When selecting an off-the-shelf audit tool or working to develop a customized tool, how can one know whether it will be a good choice? Although many audit tools used in the field have been developed by subject-matter experts (SMEs), few have been validated using additional scientific methods. As Petersen states, "The practice of accepting audits as a valid measure of excellence is questionable, unless the audits have passed some rigorous tests. Subjecting the organization to an audit that correlates progress with losses, in large enough numbers over time, may be a good indicator of performance."

This article seeks to demonstrate a method to validate safety audit tools by using scientific methods commonly used in the field of psychology (e.g., occupational health psychology, industrial and organizational psychology). To that end, the authors present as an example the validation process of a tool developed by one insurance company's loss prevention unit. The goal is to help the reader understand how to design or select good safety program evaluation tools.

Defining a "Good Audit Tool"

How can one be certain that the right things are being examined in the right way? Two characteris-

Yueng-Hsiang Huang, Ph.D., *is a research scientist at Liberty Mutual Research Institute for Safety in Hopkinton, MA. She earned her Ph.D. in Industrial/Organizational Psychology from Portland State University. Huang's primary research interests are in occupational injury and accident prevention, as well as selection, training and performance evaluation. She is a member of the Society for Industrial and Organizational Psychology, American Psychological Assn. and the ASSE Foundation Research Committee.*

Stanford A. Brubaker, CSP, ARM, ALCM, CIE, CPSM, *has been an insurance loss prevention safety professional with Liberty Mutual for 25 years, serving as a national account coordinator and casualty specialist. He is currently technical director, manufacturing technology and general liability. His work supports national operations of the business market loss prevention field staff and often involves research projects, including some with the Liberty Mutual Research Institute for Safety. Brubaker holds a B.S from Illinois State University. He is a member of ASSE's Badgerland Chapter.*

tics immediately come to mind: reliability and validity. Reliability refers to consistency and stability while validity refers to accuracy and precision (Muchinsky). A good audit tool should include consistency and accuracy, by examining the reliability/validity when applied to workplace safety and health program improvement.

Reliability

Reliability is the extent to which a score from a test/tool is stable and error-free. A score that is not stable or error-free is not useful (Aamodt). Several methods can be used to determine whether a test/tool is reliable (Muchinsky), and three types of reliability are particularly appropriate to consider when assessing the consistency or stability of an audit tool: 1) test/retest reliability; 2) internal-consistency reliability; and 3) inter-rater reliability.

Test/Retest Reliability

This means to measure something at two points in time and compare the scores. A test/tool should yield the same score on repeated use when the measured program or factor has not changed. This correlation is called "a coefficient of stability" because it reflects the stability of the test/tool over time. If the test/tool is reliable, those who scored high the first time would also score high the second time, and vice versa. As a rule, reliability coefficients around +0.70 are professionally acceptable, although some frequently used tests/tools have test/retest reliabilities only in the +0.50 range.

Internal-Consistency Reliability

The internal consistency of the test is the extent to which it has homogeneous content. If a test/tool is homogeneous (the item content is similar), it will have high internal-consistency reliability. If a test is heterogeneous (items cover a wide variety of concepts), it is not internally consistent and the resulting coefficient will be low. Guion concluded that one important characteristic of a reliable measure is that the various parts of a total measure should be so highly interrelated that they can be interpreted as measuring the same thing (Guion). One technique for assessing internal-consistency reliability is to compute Cronbach's alpha (Cronbach 297). The value of alpha ranges from 0 to 1. A higher score means a higher level of internal consistency or "test homogeneity."

Inter-Rater Reliability

When assessments are based on raters' judgments, it is possible for the raters to disagree in their evaluations. Different raters may observe the same program, yet evaluate it differently. Inter-rater reliability is the degree of correspondence between judgments or scores assigned by different raters. In some situations, raters need to exercise some judgment in arriving at a score. Estimation of inter-rater reliability is usually expressed as a correlation and reflects the degree of agreement among the ratings.

Validity

A valid measure is one that yields "correct" estimates of what is being assessed. Validity refers to the tool's appropriateness for predicting or drawing inferences about criteria. There are several different representations of validity (Aamodt; Muchinsky). Two tests are particularly appropriate in testing the validity of an audit tool: criterion-related validity and content validity.

When selecting an off-the-shelf audit tool or working to develop a customized tool, how can one know whether it will be a good choice?

Criterion-Related Validity

Criterion-related validity refers to how well a predictor relates to a criterion. The two major types of criterion-related validity are concurrent and predictive. In measuring concurrent criterion-related validity, one tests "how well a predictor can predict a criterion at the same time"—that is, no time passes between collecting the predictor and criterion data.

In measuring predictive criterion-related validity, predictor information is collected and used to forecast criterion performance—that is, a time interval occurs between collecting the predictor and criterion data. For example, one can collect scores from performing an evaluation or audit with such a tool at one point in time, then collect the criterion data one year later.

When predictor scores are correlated with criterion data, the resulting correlation is called a validity coefficient (Muchinsky). While an acceptable reliability

coefficient is in the 0.70 to 0.80 range, a desirable validity coefficient is in the 0.30 to 0.40 range. Validity coefficients below 0.30 are not uncommon, but those above 0.50 are rare. The greater the correlation between the predictor and the criterion, the more one knows about the criterion on the basis of the predictor.

Content Validity

Content validity refers to the extent to which the test (tool) items sample the content that they are supposed to measure (Pannone 507). The main purpose of validating content is to learn whether 1) each item is clear and easily understood; 2) people interpret each item as it was intended; 3) the items have an intuitive relationship to the study's topic and goals; and 4) the intent behind each item is clear to other auditors who are knowledgeable about the subject. Content validity is assessed by SMEs in the field that the test (tool) encompasses. Experts would first define the domain, then write test questions covering it. These experts would then decide how content valid the test (tool) is. Their judgments could range from "not at all" to "highly valid."

Elements of a Safety Audit Tool

Many viewpoints have been offered about how to structure a given operational audit model. Furthermore, research has shown that no single standardized audit tool offers a uniform approach to comprehensively examine a given risk reduction and safety program system. A literature review revealed several studies of elements or dimensions on which to base a comprehensive safety audit, some of which may have experienced changes in emphasis or priority over time.

For example, one published article illustrated the changes that can occur in core [safety plan] elements over a 20-year span: the support for incentive recognition plans faded, while the relative value for an ergonomics emphasis ranked much higher (Swartz 25). And although some audits shared similar content and perspective, the means of data accumulation and content/process validation differed.

In 1989, OSHA produced a guidance document, "Program Evaluation Profile" (PEP), illustrating four major elements needed to manage workplace safety risk: 1) management commitment and employee involvement; 2) worksite analysis; 3) hazard prevention and control; and 4) safety and health training (OSHA). The agency later added a concise, albeit abbreviated, version of an audit in the PEP—a series of six core elements expanded from the 1989 guidance document, each with additional aspects to examine that will lead the auditor to a more structured and repeatable score. This later approach provided a defined selection ranging from one to five points in order to arrive at the corresponding level of safety management audited. This serves as a platform on which to consider expansion and validation of the audit tool being used.

One audit tool was first developed in 1999 as a program analysis tool based on the concept, scoring model and organizational outline logic of OSHA's PEP. It was adapted by one insurance company's loss prevention field in May 2000. It has been used broadly since then by customers to examine various operations.

The original audit tool consisted of an 11-dimension ranking with a descending hierarchy of arbitrarily, but logically, assigned weightings. An optional category of "motor vehicle use" (consistently a top loss source in annual Bureau of Labor Statistics surveys) was based on application and operations of the user. For those dimensions and categories or subelements not applicable to a given situation, only scores entered received numeric accumulation in the total points possible. An improvement in the original scoring system was the primary benchmarking goal.

A user's guide walked the auditor or team through the process of selecting candidates, preparing management and labor groups, organizing interviews, preparing suitable interview techniques and questions, developing a schedule, performing the audit, preparing a final summary and closing confer-

Questions Used in Content Validity Testing

SMEs were asked to evaluate the original audit tool based on the following questions:

- Does each dimension accurately measure the intended safety program?
  __Yes. __ No. If not, why?
- How important is this dimension in terms of measuring the overall safety program?
  1) Not important.
  2) Somewhat important.
  3) Very important.
- Does each item measure its corresponding dimension?
  __Yes. __ No. If not, why? Does it represent another dimension?
- How important is each item in terms of measuring the stated dimension?
  1) Not important.
  2) Somewhat important.
  3) Very important.
- Is each item clear and easily understood?
  __Yes. __ No. If not, how should it be rewritten?
- Are the response choices for each item clear?
  __Yes. __ No. If not, how can they be worded to be more clear?
- Overall, how valid is the tool? Does the content of the tool represent what is being assessed?
  1) Not at all.
  2) Somewhat valid.
  3) Highly valid.

ence, and setting goals for follow-up audits. Selection of scoring levels within each subcategory of the 11 dimensions required considerable work and was achieved within a character-sensitive software platform that prohibited complete definitions of the scores awarded.

This original tool was well-received by both business owners and consultants involved in administering the process. This process was a new benchmarking tool, yet it lacked suitable statistical evidence that a given score would necessarily net a positive result, or that content language served as a valid indication of safe performance criteria. Furthermore, score data were not automatically entered into an electronic database; instead, they were manually summarized. Because the original version had received considerable use and input, ample evidence suggested that its limitations needed to be addressed and that it would be worthwhile to provide research evidence to validate, prove reliability, and correlate the construct and scoring system of the tool with actual loss experience.

Validating an Audit Tool: An Example

As noted, scientific methods commonly used in the field of psychology were used in the validation process, which encompassed three phases: 1) content validity testing and revision process to finalize the content of the tool for further testing; 2) reliability testing (inter-rater and internal-consistency reliabilities); and 3) criterion-related validity.

Phase I: Content Validity Testing

By following the guideline in testing content validity, SMEs from an insurance company were selected to participate based on their employment tenure, experience level with safety management and demonstrated capabilities examining organizational performance at various levels. Among the 40 SMEs recruited were past tool users, technical directors or specialists, and field SH&E consultants. Their skills of assessment, along with a cross-section of education (credentials and designations) and work with customers, netted a well-rounded pool of talent and skills, allowing for varied viewpoints and extensive comments, thereby fulfilling the intent.

The sidebar on page 38 lists the questions used to assess content. SMEs were asked to provide a "weighting" for each dimension's score, based on the overall importance of one dimension versus another and to provide additional suggestions for wording improvements that would clarify or define that dimension more effectively.

Overall, 36 SMEs (90 percent) provided comments and scores. Of these SMEs, 33 were male and their job titles included account service directors, loss prevention managers, technical directors in safety specialized disciplines and researchers. Four were new to the field, but held advanced occupational safety degrees; the remainder had on average more than 15 years' experience in the SH&E field.

Several changes were made based on findings in Phase I. The initial scoring model and ranking multipliers were revised, dimensions were reorganized and consolidated, and score definitions were clarified to increase reliability. A new audit tool was released in March 2003. The user's guide was also revised, and new interview templates were developed to provide better direction with strategic questions for analyzing labor and management perception gaps. These improvements were the models used for Phase II and III of the study. The final dimensions of the audit tool were:

1) safety and health management leadership and administration;
2) behavior safety performance and work expectations;
3) maintenance of safe working conditions;
4) injury management, claims reporting and analysis;
5) industrial hygiene hazards surveillance;
6) orientation and continuing education;
7) crisis management and life safety;
8) regulatory compliance activity;
9) occupational health protection;
10) motor vehicle safety.

Phase II: Reliability Testing

As noted, reliability testing estimates the consistency and stability of the test and whether a score from a test is constant and error-free. Many methods can be used to estimate reliability. Due to practical constraints, two kinds of reliability tests were conducted in this study: inter-rater and internal-consistency reliability tests.

Inter-Rater Reliability Test

Determination of the consistency among raters has been termed inter-rater reliability (Martinko). Ideally, auditors should visit the same site independently and compare their scores afterward. In the current project, test audits were conducted by two insurance loss prevention consultants for 40 different worksites in a manufacturing industry. Due to practical constraints, the auditors were only able to visit the site in the same trip; however, they performed their audits independently, then discussed the results jointly with the participating company at the end. The correlation for inter-rater reliability from 40 different sites was 0.98, which shows a high level of inter-rater reliability.

Internal-Consistency Reliability Test

Internal consistency is the extent to which the tool has homogeneous content (the item content is similar). Results showed that the internal-consistency reliabilities for most of the dimensions of the tool were above 0.70, which indicated reasonable levels of consistency. However, two dimensions, "behavior and health management, leadership and administration" and "occupational health protection," had low alphas of 0.51 and 0.47, respectively, compared to other dimensions.

The research team discussed ways to improve the internal consistency for these two dimensions. First, the analysis results suggested that if some particular items were deleted, the alphas would be higher. For

these items, SMEs were again asked for their opinions. In this case, after evaluating the content and relative value of these items to a given customer application, the SMEs suggested these items be retained. Another solution would be to add more items within a given dimension—in short, give more opportunities to score, thereby improving the odds of reliability. Since the items (questions and topics to audit) were developed by SMEs, it was decided not to add more items. Overall, the team concluded that these dimensions showed reasonable reliabilities.

Phase III: Criterion-Related Validity

Criterion-related validity refers to the extent to which a test score is related to some measure of performance (Barrett, et al 1). Since this audit tool served as the basis on which to evaluate safety performance, each company's injury frequency and days away from work, restricted work activity or job transfer (DART) rates were used as comparative criteria. Recordable injury frequency rates were based on the OSHA definition of recordable incidence rates (per Bureau of Labor Statistics data) by industry class. The DART rate is calculated based on (N/EH) x (200,000) where N is the number of cases involving days away and/or job transfer or restriction, and EH is the total number of hours worked by all employees during the calendar year.

Results showed that the score from the tool was significantly (negatively) correlated to customer-reported injury frequency rates in the prior year ($r = -0.28$, $p = 0.04$, significant). This means that when injury frequency is higher, corresponding scores on the audit tool are lower. This finding provided evidence of the tool's validity.

Conversely, many reported DART rates were not conclusive and did not consistently correlate to the scores. The nonsignificant relationship between DART rates and tool scores may be due to factors such as 1) recordkeeping discrepancies; 2) interpretation of OSHA standard changes in recording workplace incidents with the new recordkeeping criteria in 2001; 3) details of data collected or missing from SME audits; and 4) inaccurate reporting of worker hours where overtime and shift work may not be accurately or consistently reflected in the data collected for this study.

Overall, however, the scores were significantly correlated to company reported injury frequency rates for the prior year, which provided some evidence of criterion-related validity. According to Muchinsky, validity coefficients less than 0.30 are not uncommon.

Conclusion

As this example shows, scientific methods can be used to validate an audit tool. The tool assessed was developed and leveraged as a practical application of safety and risk management concepts and deployed over a large number of workplace operations, each offering significant input into the overall improvement of this audit process.

After the various phases of validating and subsequent revision processes, the new version of the audit tool:

- provides a uniform guidance and numeric scoring model that allows for repeatable conclusions;
- provides a comparative benchmark on which to compare future progression or regression, based on organizational use of improvement guidance;
- allows for industrial-based comparisons to similar business operations focusing on the measurement of valid components of safety performance systems and programs;
- uses a validated ranking model designed to focus attention where improvement emphasis is most needed;
- compares labor perceptions (employee support through focus group interviews) with program elements (managerial and tactical systems) to help validate scores;
- emphasizes a model of core program dimensions and subelements where risk and loss trends are most common.

Although the validation process led to these improvements, the process itself had some limitations. Since this was a field project, only certain methods were employed to test reliability and validity. Other methods/validation practices, such as test/retest reliability and predictive criterion-related validity, were not available for this project. As this was a practical project, the raters were only able to visit the sites at the same time although they performed their audits individually. When conducting inter-rater reliability tests in the future, raters should conduct the rating independently. Furthermore, all worksites involved in the current study were involved in the manufacturing industry.

No tool is perfect and the process of testing its reliability and validity is an ongoing process. Through the demonstration of the evaluation process, it is hoped that the reader will have gained knowledge about how to design or select good tools. ■

References

Aamodt, M.G. *Applied Industrial/Organizational Psychology.* Belmont, CA: Wadsworth Publishing Co., 1991.

Barrett, G.V., et al. "Concurrent and Predictive Validity Designs: A Critical Reanalysis." *Journal of Applied Psychology.* 66(1981): 1-6.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika.* 16(1951): 297-334.

Guion, R.M. *Personnel Testing.* New York: McGraw-Hill, 1965.

Martinko, M.J. *The Job Analysis Handbook for Business, Industry and Government.* New York: Wiley, 1987.

Muchinsky, P.M. *Psychology Applied to Work.* Belmont, CA: Brooks/Cole Publishing Co., 1997.

OSHA. "Program Evaluation Profile." Washington, DC: U.S. Dept. of Labor, OSHA, 1989. <http://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=FEDERAL_REGISTER&p_id=12909>.

Pannone, R.D. "Predicting Test Performance: A Content Valid Approach to Screening Applicants." *Personnel Psychology.* 37(1984): 507-514.

Petersen, D. "Safety Management 2000: Our Strengths & Weaknesses." *Professional Safety.* Jan. 2000: 16-19.

Swartz, G. "Safety Audits: Comparing the Results of Two Studies." *Professional Safety.* Feb. 2002: 25-31.