Case Study: The Importance of the Assessment Technique in Chemical Safety Training on a College Campus

James H. Withers and Steven A. Freeman

Abstract

Safety training is an integral part of every organization's overall safety program. A variety of delivery methods are used to conduct training with the most common learning outcome being performance on a written exam. The safety professional must consider numerous issues when composing a written exam, including question design and exam difficulty, to establish a meaningful passing level and to assess overall training effectiveness. A research study was undertaken to further explore issues related to question design and exam difficulty relative to a chemical safety course offered in both classroom- and computer-based formats on a college campus. The objectives of this study were to 1) evaluate the potential impact of question difficulty as a part of an assessment technique that measures learning and 2) evaluate the potential impact of exam difficulty and sequence of exam administration as a part of an assessment technique that measures learning. An analysis of question difficulty factors across three different versions of learning assessments used showed differing levels of difficulty. Additionally, the order of administration of the exam was a factor in the amount of measured learning. The implications of these results are discussed. Nuances of assessment techniques, including question difficulty and order of administration, must be evaluated to truly evaluate the effectiveness of any safety training intervention.

Keywords
Safety training, assessment technique, training effectiveness

Introduction & Background

Safety training is conducted using a variety of delivery methods. In addition to traditional classroom offerings, safety professionals have been using new technologies, such as computer-based training, at an increasing rate since the 1980s. An International Data Corp. study projected that 80% of safety training would be conducted via computer by 2003 (Overheul, 2002). Accordingly, studies on training effectiveness began to emerge in the scientific literature that examined differences in learning between the two methods (Bowan, et al., 1995; Coppola & Myre, 2002; Hasselbring, 1986; Kulik & Kulik, 1991; Lawson, 1999; Robson, et al., 2010; Stephenson, 1991; Williams & Zahed, 1996).

Regardless of the delivery method for safety training, learning outcomes must first be defined. Once defined, training effectiveness can be evaluated relative to the success in achieving these learning outcomes. In a recent NIOSH-funded literature review, four categories of learning outcomes were identified: 1) knowledge (typically shown via a written exam covering a particular policy, procedure or hazard); 2) attitudes and beliefs (including perception of risk); 3) behaviors (meaning worker actions that could result in exposure to hazards); and 4) health (referring to early detection of illnesses/injuries) (Robson, et al., 2010). Of the four outcomes, the most common in safety training is showing knowledge via a written exam (Burke, 2006). At X University, the majority of current safety training offerings have a written exam component (R. Book, personal communication, Dec. 6, 2010).

The safety professional has numerous issues to consider when composing a written exam. What are the appropriate questions to ask? Are questions clear? Did the training course cover the topic in sufficient detail to allow the participant to answer the question correctly? At this point, the safety professional is faced with a dilemma. Weidner (2000) stated that while safety regulations with training requirements are based on known scientific principles related to hazards, they often lack the underpinnings of the principles of adult learning and assessment. This becomes increasingly important when considering the measure of success in exam-based safety training: achievement of a minimum passing score (percentage) on a postcourse test. In general, a 70% score is widely accepted as an indicator of "moderate" knowledge, 80% of "moderately higher" knowledge and so forth (Angoff, 1984). However, the safety professional must wrestle with issues related to question design and exam difficulty to establish a meaningful passing level. This is especially important given the prevalence of exam-based safety training. While the concept of this research is not new, the context has not appeared before in the literature. Many higher education institutions routinely provide chemical safety training that could benefit from a more systemic approach to their assessments processes.

Research Objectives

This research is part of a larger study looking at delivery methods of safety training and the resulting knowledge gained and retained over time consistent with NIOSH, OSHA and

James H. Withers is the environmental health and safety manager at Danfoss Power Solutions in Ames, IA. He may be contacted at jwithers@sauer-danfoss.com.

Steven A. Freeman is professor of occupational safety in the Department of Agricultural and Biosystems Engineering at Iowa State University in Ames, IA.

American National Standards Institute training paradigms. See Withers, et al. (2012) for the theoretical explanation behind the training framework and the details of the broader study. This study was undertaken to further explore issues related to question design and exam difficulty.

The study focused on a chemical safety training course offered at X University that is an example of exam-based safety training. The course is offered in both classroom and computer-based formats and is considered the backbone of the university's chemical safety program. The course provides basic chemical safety programmatic information to the learner and provides a "roadmap" by which a research group-specific safety program can be developed and implemented. Course topics covered include regulations, terminology, roles and responsibilities, exposure controls and prevention, recordkeeping, exposure monitoring, MSDSs, emergency preparedness, PPE and lab maintenance and inspection.

The first topic evaluated was question difficulty. A specific, associated research objective was as follows:

Evaluate the potential impact of question difficulty as a part of an assessment technique that measures learning.

The larger issue of overall exam difficulty was also explored in relation to question difficulty. The specific associated research objective was as follows:

Evaluate the potential impact of exam difficulty and sequence of exam administration as a part of an assessment technique that measures learning.

Data were collected from participants in a required university chemical safety training course. The 243 participants represented a broad cross-section of university employees and students [for a detailed description of the population and the objectives of the larger study see Withers, et al. (2012)]. Study results were used to identify lessons learned that could be applied to programmatic and course improvements. An additional purpose was to demonstrate simple techniques that other safety professionals can use or adapt for use when evaluating the issue of question and exam difficulty relative to an exam-based safety training course.

Research Methods

The data collection mechanism used was a learning assessment tool (LAT). The LAT consisted of 16 multiple-choice questions, each testing knowledge of a specific topical area. To measure knowledge gained and knowledge retention, LATs were given to participants prior to training, after training and 1 year after training (Withers, et al., 2012). Three versions of the LAT were developed in consultation with a panel of experts with extensive chemical safety and regulatory experience with responsibilities for managing all aspects of chemical safety in a university environment. Question consistency across the three versions of the LAT was tested using a Wilk's Lamda calculation to determine how well each of the three questions tested the student on a particular learning outcome (Hinkel, et al., 2003). In other words, if the three questions were clearly written and the participant had salient knowledge of the topic, all questions should be answered correctly. Conversely, in a situation in which the participant did not have knowledge of the concept, all three questions would be answered incorrectly.

To measure knowledge gained as a result of the training experience, the LAT was administered prior to and after training. In classroom sessions, the pretest and posttests were handed out to participants. In computer-based sessions, the pretests and posttests were presented to the participant automatically on the computer. In each case, the version (1, 2 or 3) was randomly selected by the instructor or computer program. Upon completion of the course, a second and different version of the LAT was administered. Upon completion, each LAT was scored for number of questions correct. In addition, the number of individuals getting a particular question correct (or not) was also collated for each question on the three versions of the LAT.

Results & Discussion

Question set analysis via Wilk's Lambda test statistic revealed three of the 16 topical areas had one of three questions that was not consistently answered correctly relative to the other two. The three discrepancies were in the areas of training records, regulations and laboratory audits. A review of the individual questions did not reveal any apparent issues with clarity (as described before) that would warrant restructuring of the question. This information was used to review the content of both versions (computer and classroom) to ensure that it was delivered clearly prior to the study's commencement.

A common method for evaluating question difficulty is by evaluating the "difficulty factor" (DF) (Knauper, et al., 1997). DF is calculated by taking the number of individuals answering the question correctly divided by the total number of participants answering the question. In general, a calculated DF of > 0.7 is considered to be an "easy question"; a DF of < 0.3 is generally regarded as a difficult question. If a test's purpose is to discriminate between different levels of achievement, items with difficulty values between 0.3 and 0.7 are most effective. The optimal level should be 0.5 (Arizona State University, 2004). For the purpose of assessing exam question difficulty, a DF was calculated for each question on each LAT when taken as a pretest. The pretest was chosen so as to minimize any learning effect caused by participation in the training. Results are shown in Table 1.

An analysis of the data for each LAT shows that each version had a majority of questions that had a DF > 0.7 (denoted in green). Specifically, LAT Version 1 had 11 of 16, LAT Version 2 had 9 of 16 and LAT Version 3 had 10 of 16 questions with calculated DFs that were greater than 0.7. Conversely, each LAT also had some questions that fit the difficult criteria (< 0.3) (denoted in red). Specifically, LAT Version 1 had 2 of 16, LAT Version 2 had 3 of 16 and LAT Version 3 had 2 of 16. Data tend to support an overall conclusion that the exams are weighted on the "too easy" side. Given that data were generated by a group of participants who had no prior work experience with chemicals or any prior chemical safety training further supports that conclusion.

To further evaluate the issue of LAT difficulty, an analysis was conducted of overall pass rate for each LAT for the

TOPICAL AREA	LAT 1	LAT 2	LAT 3
Regulations	1.0	.23	.37
Laboratory Practices	.58	.46	1.0
Emergencies	.50	.38	.50
Exposure Control	.92	.15	.50
Training	.75	.38	.75
Material Safety Data Sheet	.25	.92	1.0
Personal Protective Equipment	.92	1.0	.75
Inspections	1.0	.92	.13
Postings	.58	.92	.75
Lab Procedures	.92	.15	.75
Labels	.83	.58	.63
Transportation	1.0	.92	.75
Behaviors	1.0	1.0	.88
Spills	.92	.85	.88
Standard Operating Procedures	.98	1.0	.25
Waste Disposal	.17	1.0	.88

NOTES: LAT = Learning Assessment Tool; values >0.7 denoted in green; values <0.3 denoted in red.

Table 1 Pretest Difficulty Factor Data: Participants With No Prior Work Experience or Previous Chemical Safety Training

same group, participants with no prior work experience with chemicals or any prior chemical safety training. For LAT Version 1 taken as a pretest, 83% of participants achieved a 70% or greater; the passing rates were 54% for LAT Version 2 and 75% for Version 3. These data suggest that the difficulty of each version might be different (i.e., Version 2 is more difficult that the other two). The implications of question and LAT difficulty are discussed in the Summary and Conclusions sections.

Order of assessment of the LAT was also explored. Inherent in the development of the three versions of the LAT was an assumption that all three were of equal difficulty. Given the previously described methodology, there were several possible combinations of administering the three versions of the LAT as pretests and posttests.

To evaluate the question of whether or not all LAT versions were equivalent in terms of difficulty, all possible combinations of the three versions were evaluated for amount of learning (defined as Delta 1). This evaluation was completed using an analysis of variance (ANOVA) model where Delta 1 was defined as the dependent variable and LAT order (Version Group) and computer or classroom (Delivery Method) were defined as the independent variables. Table 2 shows the results.

The *p-value* data show that both the version group and delivery method are significant in terms of explaining differences in learning. The calculated value of R^2 was 0.397, which indicates a strong model [defined as: Learning (Delta 1) = Version Group + Delivery Method]. The least squares mean data indicate two interesting trends. Study participants taking Version 2 as a pretest and Versions 1 or 3 as a posttest showed the greatest increase in learning of all possible combinations. A possible explanation of this result is that participants scored low initially on Version 2 because of increased difficulty. When Versions 1 or 3 were taken as the posttest, the amount of measured learning was greater than the other combinations.

Conversely, study participants who took Versions 1 or 3 as a pretest may have scored higher initially because they were easier and then showed less learning (or even a decrease) due to Version 2, as the posttest, being more difficult. The combination of these two observations suggests that Version 2 is a more difficult LAT than Versions 1 or 3. The implications of this finding are discussed in Summary and Conclusions.

Summary

When considering the previous data, it should be obvious that the safety professional needs to consider assessment technique early in the training development process. Reliability

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Statistic	Probability > F
Version Group	5	589.387	117.877	28.88	<0.0001
Delivery Method	1	20.392	20.392	5.00	0.026
R-Square	0.397				

NOTES:

Version Group 1 = LAT 1 then LAT 2
Version Group 2 = LAT 1 then LAT 3
Version Group 3 = LAT 2 then LAT 1
Version Group 4 = LAT 2 then LAT 3
Version Group 5 = LAT 3 then LAT 1
Version Group 6 = LAT 3 then LAT 2

LEAST SQUARES MEAN:

Version Group 1 = -0.233
Version Group 2 = 0.265
Version Group 3 = 3.538
Version Group 4 = 3.466
Version Group 5 = 2.182
Version Group 6 = -0.020

Table 2 ANOVA for LAT Order

testing conducted during the development of the LAT provided valuable feedback that was a catalyst for a review of training content. An analysis of difficulty factor data, the overall pass rate for each LAT and the influence of exam order suggested that Version 2 of the LAT was more difficult than the other two.

However, at this juncture, the safety professional must consider another issue: establishing a passing level. As mentioned, 70% is a commonly used passing level in safety training, but how can the safety professional establish a passing level without consideration of question and exam difficulty as well as order of administration?

In the example, a majority of questions had a DF > 0.7 (LAT Version 1: 11 of 16, LAT Version 2: 9 of 16, LAT Version 3:10 of 16). Conversely, each LAT also has several questions that fit the difficult criterion (< 0.3) (LAT Version 1: 2 of 16, LAT Version 2: 3 of 16, LAT Version 3: 2 of 16). Without an understanding of LAT composition, in terms of the distribution of difficult or easy questions, the safety training's impact and value are difficult to determine. Organization management might look at the high rate of safety training completion and falsely conclude that workers, because of participation in safety training, are now "qualified" when, in reality, the assessment technique did not have sufficient rigor. Conversely, the safety professional might look at low pass rates for a given safety course and conclude that some aspect of the course (e.g., content) needs improving when, in reality, the assessment technique used was too difficult.

A similar discussion is necessary related to exam difficulty and order of administration. As was shown in this study, both exam difficulty and order of administration played a key role in the measured amount of learning. A false assumption was made that each exam had the same amount of difficulty when, in fact, one version was more difficult than the other two. A training participant who took the more difficult version of the exam as a pretest and then showed a significant gain in knowledge on a posttest might lead the safety professional to conclude that the training intervention was highly effective. Conversely, if the participant took the more difficult version of the exam as the posttest, the false conclusion would be that the training intervention was not effective (i.e., the participant did not learn much).

It should be obvious that data related to question and exam difficulty are necessary for the safety professional to evaluate safety training course effectiveness. Data generated in this study indicate a need to further evaluate the composition of LAT Version 2. Any changes made in individual questions would necessitate the need to reevaluate issues related to pass rate, etc. If the safety professional can show equivalent difficulty with each version of the LAT, then improvements in the assessment technique can be made. For example, raising the passing rate to 80% or higher might be evaluated as an option. However, what additional issues will that present in terms of ensuring the adequacy of content, length of course and other variables related to delivery methods? Will the safety professional spend more time with participants who do not achieve a passing grade outside of class and, therefore, devote more of his/her limited time to supporting the overall training program?

Developing an effective safety training program is challenging in any work environment. Clearly, many complexities are associated with evaluating safety training effectiveness. Sugure and Rivera (2005) reported that only about 50% of companies measure learning outcomes from training, and less than 25% make any attempt to assess potential programmatic improvements resulting from training. Today, the predominate type of safety training includes administration of a written exam and the achievement of a minimal score as a measure of success. To properly evaluate this type of assessment technique, it is imperative that the safety professional have the necessary data collection mechanisms in place. Evaluation of these data and resulting training enhancements will be an ongoing and iterative process.

Conclusions

This study has demonstrated the usefulness of several straightforward analytical techniques that can be used to assess issues related to both question and exam difficulty. It should be noted that the issue of exam difficulty was done within a specific chemical safety course. The results presented and discussed in this study cannot be used to predict potential outcomes of evaluations of other courses. The only way to truly shed light on issues related to the value of the assessment technique used is to implement a process by which course and exam-specific data can be collected and analyzed. The need to include this important step in the developmental process is directly related to the significance of the training course subject matter and the intended learning outcomes. Finally, there must be a clear indication of learning that results from the training experience that is not influenced by nuances (e.g., exam difficulty and exam order) associated with the assessment technique. ✪

References

Angoff, W. (1984). Scales, norms and equivalent scores. In A. Ward, H. Stoker, & M. Murray Ward (eds.), *Educational Measurement: Theories and Applications*. Princeton, NJ: Educational Testing Service. Retrieved from http://books.google.com/books?hl=en&lr=&id=dWTx7RXc028C&oi=fnd&pg=PA121&dq=Angoff++scales+norms+and+equivalent+scores&ots=hZGUHA236m&sig=6KR4CakpIWkqTsWmMa3a_99MFgc#v=onepage&q=Angoff%20%20scales%20norms%20and%20equivalent%20scores&f=false

Arizona State University, University Testing and Scanning Services. (2004). Exam scores: How to interpret your statistical analysis reports. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCkQFjAA&url=https%3A%2F%2Fuoeee.asu.edu%2Fsites%2Fdefault%2Ffiles%2Fdocs%2FGuide_stat_analy_exam_scores.pdf&ei=tE15UuTdBKecyQHurYC4Aw&usg=AFQjCNH6FXqlmZP-NEHf-nEr3FypKw9Pgw&bvm=bv.55980276,d.aWc&cad=rja

Bowman, B.J., Grupe, F.H. & Simkin, M.G. (1995). Teaching end-user applications with computer-based training: Theory and an empirical investigation. *Journal of End User Computing, 7*(2), 12-18.

Burke, M.J. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health, 96*(2), 315-324.

Coppola, N. & Myre, R. (2002). Corporate software training: Is web-based training as effective as instructor-led training? *IEEE Transactions of Professional Communication, 45*(3), 170-184.

Hasselbring, T. (1986). Research on the effectiveness of computer-based instruction: A review. *International Review of Education, 32*(3), 313-324.

Hinkel, D., Wiersma, W. & Jurs, S. (2003). *Applied statistics for the behavioral sciences.* Boston, MA: Houghton Mifflin Co.

Knauper, B., Belli, R., Hill, D. & Herzog, R. (1997). Question difficulty and respondent's cognitive ability: The effect on data quality. *Journal of Official Statistics, 13*(2), 181-199.

Kulik, C.-L.C. & Kulik, J.A. (1991). Metaanalytical studies of findings on computer-based instruction. In E. Baker & H. O'Neil (eds.), *Technology assessment in education and training.* New Jersey: Lawrence Erlbaum Associates Inc. Retrieved from http://books.google.com/books?hl=en&lr=&id=7spWIqyYdVIC&oi=fnd&pg=PA9&dq=kulik+1991&ots=UKakOb2h-f&sig=Le5fK6-YI08kaDKxGctnF8c-Nec#v=onepage&q=kulik%201991&f=false

Lawson, S.R. (1999). Computer-based training: Is it the next wave?. *Professional Safety, 44,* 30-33.

Overheul, V. (2002). Does online training live up to its promises? *Occupational Health & Safety, 71*(6), 100.

Robson, L., Stephenson, C., Schulte, P., Amick, B., Chan S., Bielecky A., . . . Grubb, P. (2010). A systematic review of the effectiveness of training & education for the protection of workers. Toronto: Institute for Work & Health. Cincinnati: NIOSH.

Rudman, W.B. (2003). *Emergency responders: Drastically underfunded, dangerously unprepared.* New York, NY: Council on Foreign Relations.

Stephenson, S.J. (1991). *The effect of instructor-student interaction on achievement in computer-based training* (AL-TP-1991-0002). Brooks Air Base, TX: Armstrong Laboratory, Technical Training Research Division.

Sugure, B. & Rivera, R.J. (2005). 2005 state of the industry: ASTD's annual review of trends in workplace learning and performance. Alexandria, VA: American Society for Training and Development.

Thalheimer, W. (2006). *Practical wisdom from learning research.* Somerville, MA: Work-Learn Research.

Wiedner, B. (2000). Testing as a measure of worker health and safety training: Perspectives from a hazardous materials program. *American Journal of Industrial Medicine, 37,* 221-228. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-0274%28200002%2937:2%3C221::AID-AJIM8%3E3.0.CO;2-W/pdf

Williams, T.C. & Zahed, H. (1996). Computer-based training versus traditional lecture: Effect on learning and retention. *Journal of Business and Psychology, 11*(2), 297-310.

Withers, J.H., Freeman, S.A. & Kim, E. (2012, Sept.-Oct.). Learning and retention of chemical safety training information: A comparison of classroom versus computer-based formats on a college campus. *Journal of Chemical Health & Safety,* 47-55.