August 2020

Novel Approach in Measuring Training Effectiveness

Thomas Samuel
University of Wisconsin-Milwaukee

NOVEL APPROACH IN MEASURING TRAINING EFFECTIVENESS

by

Thomas Samuel

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin–Milwaukee

August 2020

ABSTRACT

NOVEL APPROACH IN MEASURING TRAINING EFFECTIVENESS

by

Thomas Samuel

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Dr. Naira Campbell-Kyureghyan

Conducting training for employees is a popular method, in industry, to increase awareness and competencies of individuals that helps start the journey towards a change in performance. Globally, fiscal investments in the order of hundreds of billions of dollars annually are made by organizations and governments to train employees in various concepts. Of the various types of training programs, safety related training is specifically important as it increases awareness of the work risks to employees and plays a critical role in reducing safety incidents in the workplace. This has considerable societal and organizational impacts as reduction in safety incidents reduces mortality and injury rates among workers, improves their work environments and benefits the organizations as they have reduced costs and happier employees. Due to the level of investment made and associate positive impacts it is important to ensure that there is an acceptable return on investment for the training provided and that the training is effective.

As with any measurement method, training evaluations have gaps in their ability to determine if the participants are guessing or if they truly have learned the concept. Additionally, the measurement and reporting of training effectiveness can be improved to help industry trainers and organizations quickly determine which of the concepts trained were truly learned and what changes or countermeasures need to be implemented to the content or the delivery to help improve the effectiveness.

The goal of this research is to improve the assessment of training effectiveness by quantifying the effect of guessing and accounting for participant prior knowledge of the concepts. This was achieved by conducting post-hoc analysis on training assessment data collected from 1,474 participants in three major utility industries and quantifying the effect that the inclusion of the IDK option has on learning in a pre-/post-test assessment model by introducing the concept of a Control Question (CQ). The results showed that there was a statistically significant reduction of 28% in the use of the IDK option in the post-test compared to the pre-test for all questions including the CQ. Research was conducted into methods to determine the best and least learned concepts by the participants in the training. The Assessment of Training Effectiveness Adjusted for Learning (ATEAL) method was developed to adjust learning for participant prior knowledge and any negative impact they might have experienced due to the training. The ATEAL method was validated using scenario analysis and simulations and its performance was compared to the most popular metrics (Total Percentage Correct or Post-Pre Percentage Correct) currently used to report training effectiveness. The questions that were administered to the participants were grouped into safety concepts to determine which were the best and least learned concepts in the training for the different training groups and industries. The results of the ATEAL method were compared to the results reported by the commonly used metrics and detailed investigations into the merits and gaps of each method were conducted. It was observed that the ATEAL method performed better at identifying the concepts that were the best learned while compensating for prior knowledge and poor experience during the training. An additional advantage of this method is that the ATEAL method is not limited to MCQ assessments and can be used in any situation with score-based pre-/post-training assessments.

The knowledge gained from this research will enable trainers and organizations to design training assessments that make better use of the IDK option by understanding that it does reduce guessing behavior in the pre-test assessment, but it does not reduce participant guessing in post-test assessments. Including a Control Question in the pre- and post-tests assessments can be helpful with generating estimates of the probability of guessing and allow better estimates of true learning and training effectiveness. As a result of this research a method (ATEAL) was developed to allow trainers to quickly and accurately identify the concepts that the participants need further training on and where improvement to the training is required. A newly introduced concept of Training Effectiveness Matrix can be used for visual assessment of whether the trainees exhibited prior knowledge for a certain concept, if there was evidence of guessing or if the participants experienced any other possibly non-positive effects as a result of training. The results of this research enable the development and implementation of countermeasures to improve the training for the participants and thereby offer some guidance for increasing the training effectiveness. Additionally, these methods can be much more broadly adopted to any environment where there is a transfer of knowledge and thus has far reaching benefits across numerous industries and organizations, as well as for various training and assessment styles.

To,

my parents,

Ava, my daughter & editor,

and especially my wife

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CQ	Control Question
TPC	Total Percent Correct
PPPC	Post-Pre Percent Correct
PK	Prior Knowledge
PTI	Positive Training Impact
NTI	Negative Training Impact
LAC	Learning Adjustment Coefficient
NTIC	Net Training Impact Coefficient
TEM	Training Effectiveness Matrix
MCQ	Multiple Choice Question
T/F	True/False
ATEAL	Assessment of Training Effectiveness Adjusted for Learning
ANOVA	Analysis of Variance
ANCOVA	Analysis of Covariance
IDK	I Don't Know
RQ	Research Question
PPE	Personal Protective Equipment
NIOSH	National Institute for Occupational Safety and Health
DOL	Department of Labor
ROI	Return on Investment
CC	The question is answered correctly in both pre- and post-tests

CI	The question is answered correctly in the pre-test and incorrectly or IDK in the post-test
IC:	The question is answered incorrectly or as IDK in the pre-test and correctly in the post-test
II:	The question is answered incorrectly or as IDK in both pre- and post-test assessments

ACKNOWLEDGEMENTS

"This is a wonderful day. I've never seen this one before."

- Maya Angelou

I would like to extend my sincerest gratitude to Dr. Naira Campbell-Kyureghyan for providing guidance, mentorship and motivation during this ten-year journey to complete my Doctorate in Philosophy. Her understanding and support during my changes in jobs and countries during this time made it all possible, enjoyable and fulfilling.

I would like to thank Dr. Razia Azen for her support in the analysis, input on the contributions and support in developing the manuscripts. I appreciate the guidance and support from Dr. Wilkistar Otieno and Dr. Hamid Seifoddini throughout my degree. Finally, I will be forever grateful to Dr. Michael Lovell for inspiring me to start this journey over a decade ago, at which time it felt like a fool's errand.

I am extremely thankful to my laboratory colleagues, Madiha Ahmed and Blake Johnson, who supported me when I would intermittently visit the lab and made me feel part of the team. And finally, to Betty Warras who was always a cheerful support through my years of blundering questions.

Chapter 1: Introduction & Scope of Research

1.1 Motivation

Formal training of employees is a popular way for organizations to increase, refresh or update the awareness and knowledge of their workforce in specific competencies and use it as an avenue to change the expected behavior of the participants (Tai, 2006). Globally, in 2016, organizations spent $359 billion on training (Glaveski, 2019) with US organizations increasing their spending by 62% from $54 billion in 2000 (Arthur Jr., Bennett Jr., Edens & Bell, 2003) to $87.6 billion in 2018 (Freifeld, 2018). Brunello & Medio (2001) observed that different countries invest varying amounts in employee training based on tenure, but there is an overall approach globally to increase the knowledge of employees in an organization using formal training methods. This significant amount of financial and time investment made across the globe by organizations on developing employee skills necessitates measurement or quantification that the training is effective and will result in the desired changes in their behavior.

Trainings can cover a wide variety of topics based on the industry and their needs, but training on safety related topics is particularly important due to the severity and impact of poor safety behavior. The US Bureau of Labor and Statistics reported that the number of fatal work injuries increased by 2% from 5,147 fatalities in 2017 to 5,250 fatalities in 2018. Ho & Dzeng (2010) reported similar statistics on occupational disasters in Taiwan. This impact on human life and the societies that people live in has given rise to a number of legislative acts and organizations to help reduce occupational injuries. The primary path in impacting occupational injuries is by mandating workers undergo formal safety training as it is a well-documented method to improve the safety outcomes of employees worldwide and reduce safety incidents on the worksite (Bahn & Barratt-Pugh, 2012; Ho et al., 2010; Burke, Sarpy, Smith-Crowe, Chan-

Serafin, Salvador & Islam 2006; Becker & Morawetz, 2004; Demirkesen & Arditi, 2015). Campbell-Kyureghyan, Ahmed & Beschorner (2013) observed that safety training for employees in dynamic work environments, such as construction, is even more important since traditional countermeasures to reduce hazards i.e. workstation redesigns etc. are ineffective or not practical in a constantly changing work environments. Effective training has been observed to increase the knowledge, skills and abilities (KSAs) of employees benefit (Blume, Ford, Baldwin & Huang, 2010 & Tai, 2006) and this is specifically important in the case of safety training as improved understanding and application of these KSAs have significant positive human and societal impact.

To ensure that the training is effective many methods of measuring the effectiveness exist, of which the Kirkpatrick's model (Kirkpatrick, 1967) is the most frequently used by industry trainers (Arthur et al, 2003). Details of the model are described in Section 3.1.1 and an observation of note is the linkage between the amount of learning by the participants and the impact it has on their application of this knowledge in their work environment (Kontoghiorghes, 2001). Thus, it is important for accurate assessments of participant learning to ensure that organizations can expect improved safety performance by the trainees. The most popular assessment method is to administer Multiple Choice Questions (MCQ) assessments to the participants to assess their knowledge of the concepts taught. This is typically done by conducting post-test assessments or by conducting assessments before and after training using pre-/post-test assessments to assess their knowledge gain.

One of the most frequent criticism of an MCQ assessment method, among training professionals is that it enables examinees to more easily achieve the correct answer by guessing compared to other methods. Of the many methods (formula scoring adjustments etc.) that have

been proposed and researched to measure the true score of the individual by accounting for guessing, the introduction of an 'I Don't Know' (IDK) option in MCQ assessments has been shown to be a good way to minimize guessing as it gives the participants an option that they can truthfully answer if they did not know the answer to the question. However, this previous research, which is further explored in Chapter 2, has mainly focused on the use of the IDK option in a post-test assessment construct and predominantly in True (or) False (T/F) type MCQs. Thus, a gap in knowledge still exists on how the IDK option affects guessing on MCQ assessments with more than 2 options in a pre-/post-test assessment model

The other important aspect of measuring training effectiveness is the method by which the scores themselves are calculated and reported. A variety of methods, discussed in detail in Chapter 2, ranging from the simple i.e. reporting on the number correct answers achieved in a post-test assessment, to the statistically rigorous i.e. calculating the score deltas by using ANOVA and/or ANCOVA for analysis, exist to help trainers and organizations quantify the learning of the concepts taught. In conducting pre-/post-test assessments it is observed that learning scores are affected by the positive training impact, prior knowledge of the participants, negative training impact and by zero learning that they might have experienced during the training. Despite all these methods there still does not exist an easy method to help trainers quickly assess training effectiveness for the concepts taught while adjusting for participants' prior knowledge and the negative training impact that they might have experienced.

1.2 Scope of Research

Research was conducted into reduction of occupational injuries for multiple sectors of three major utility industries as part of the DOL Susan Harwood Training Grant program by a team at the University of Wisconsin-Milwaukee over six years (DOL OSHA Grant Numbers: SH-20840-

SH0, #SH-22220-SH1, #SH-23568-SH2, #SH-24880-SH3). The research was primarily focused on small business utilities and utility contractors as they have limited resources and thus were at a disadvantage in improving safety/ergonomic of their employees. The three utility sectors considered for this research were Natural Gas, Electric Transmission and Power Generation utilities which consisted of 16, 15 and 4 companies respectively (Campbell-Kyureghyan, Hernandez & Ahmed 2013). As a result of this research training programs were developed, administered, and evaluated for a total of 1,474 workers and managers across the various utilities, and the detailed demographics of the participants are provided in Table 3-6. The current research involves the further analysis of the training assessment results from these participants to address the research goals.

The methodology used to recruit companies, perform safety audits and assessment, develop novel training materials and deliver them in the form of training to improve the safety knowledge of the employees and managers and develop Train-the-Trainers in the various companies is detailed in Figure 1-1. The process began with the recruitment of utility companies in the target utility sector and was followed up with onsite visits. As the development of the training materials depended on understanding the specific risk factors and concerns found at the utilities, it was important to observe the safety risks experienced during the onsite visits. These observations involved interviews with managers & employees and direct observation of the performed tasks and videotaping. This analysis, in addition to nationwide injury and fatality statistics for each specific utility industry, helped complete the needs assessment to develop the training materials (Campbell-Kyureghyan, Principe & Ahmed, 2013) and the training content was comprised of multiple modules that address specific needs of the various utility sectors. The

scope of the current research solely focuses on post-hoc analysis of Level 2 evaluations done before (pre-test) and after (post-test) training as illustrated by the red outlined area in Figure 1-1.

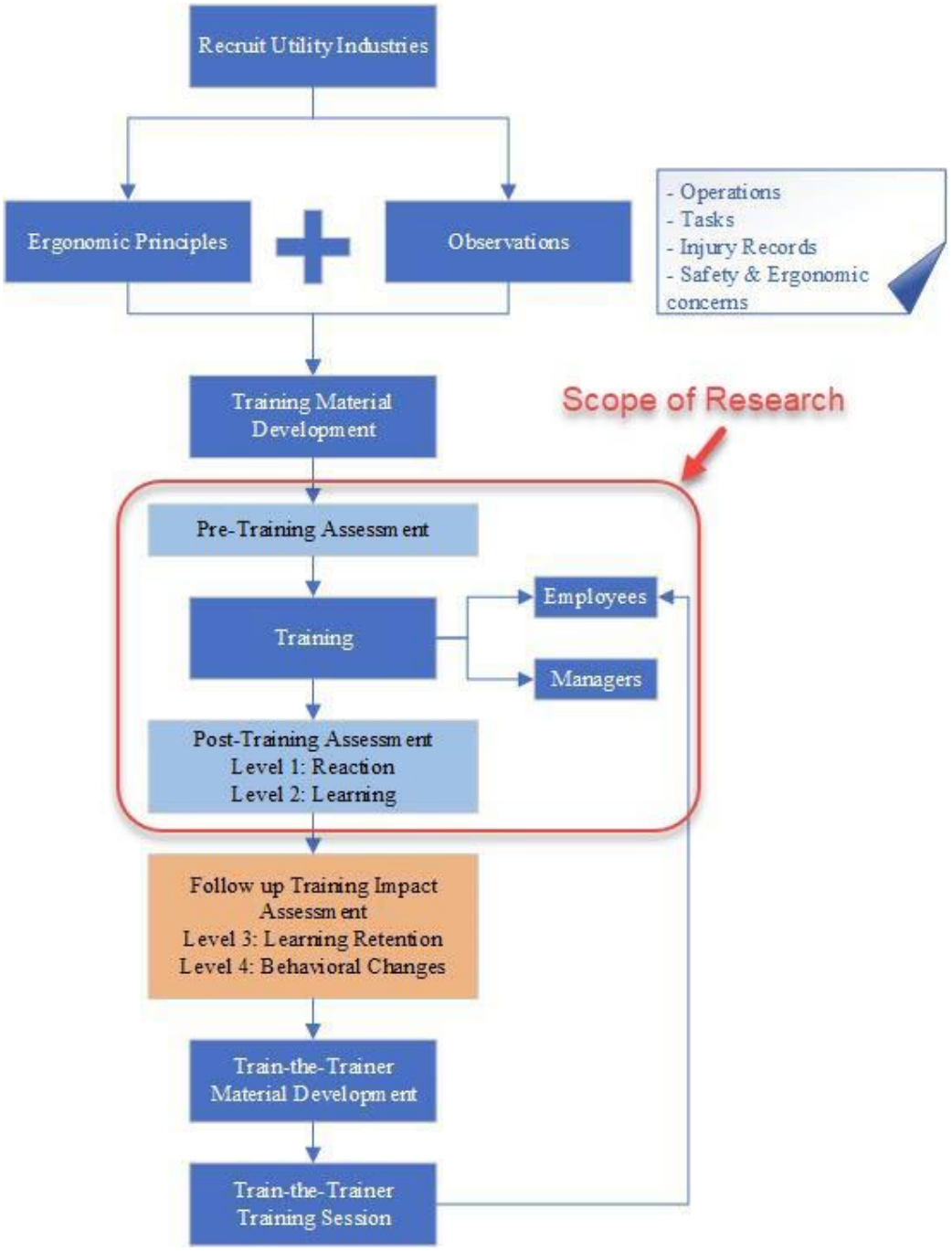

Figure 1-1: Flow diagram of the process used to recruit and assess companies, develop training materials and deliver training to various audiences

The developed training material was split into concepts that were common across and unique to each specific utility industry. Figure 1-2 illustrates the concepts that were common between

all three industries and those that were specific to each specific utility. A total of 7 concepts were common between the three industries. However, it is important to note that the content details within each concept was specific to that industry and company. Similarly, 4 concepts were common between the three industries (italicized in Figure 1-2) and finally 4 concepts were completely unique to a specific industry.

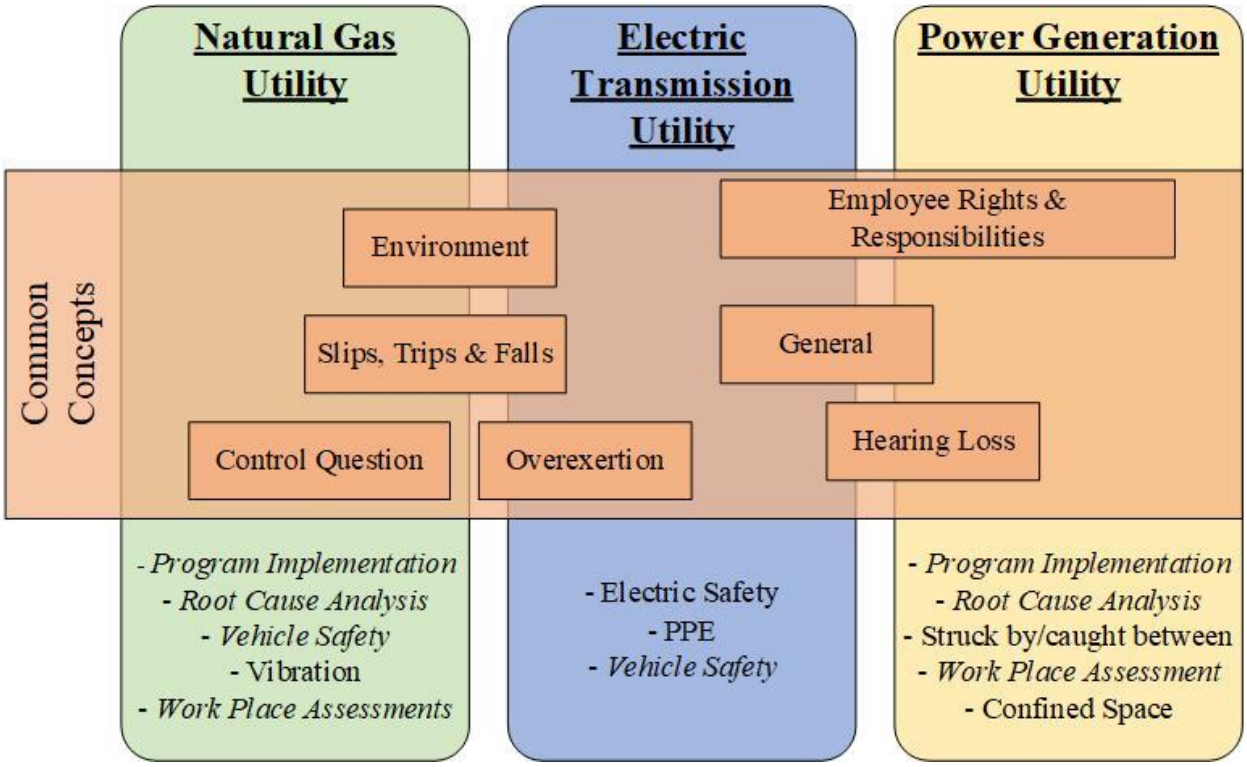

Figure 1-2: Safety concepts taught by utility industry

Training was conducted onsite and split into two categories, Tier 1 & Tier 2. Tier 1 training was conducted by individuals who developed the training content and was targeting employees and managers. Tier 2 sessions were conducted by individuals who attended a nationwide train-the-trainer program led by the Tier 1 trainers and were targeting to train only employees, as illustrated in Figure 1-3 (Campbell-Kyureghyan et al., 2013). The employees received a training of 4-5 hours with the managers receiving an additional 2 hours of content focusing on workplace risk assessment and program implementation.

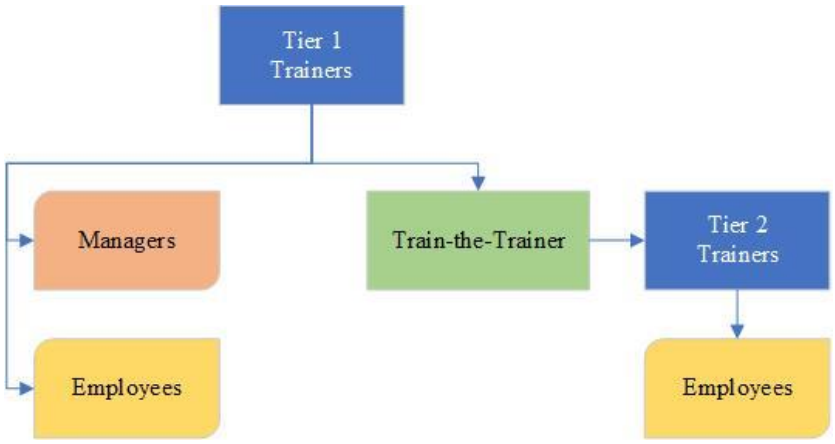

Figure 1-3: Audiences trained by Tier 1 and Tier 2 professionals

Training effectiveness was measured using a modified the Kirkpatrick's model (Campbell-Kyureghyan et al, 2013, Kirkpatrick, 1967) and as illustrated in Figure 1-1, Level 1 evaluations were conducted to capture the participant reaction to the training session immediately after training. Additionally, a survey to capture self-reported demographics was administered at the same time. Level 2 evaluations were made after the training and Level 3 and Level 4 evaluations were made after the participants had an opportunity to apply the learning in their workplace. Results of evaluations from Level 2 were used as input to modify training content and delivery to improve effectiveness. Similarly, results from the Level 3 & 4 evaluations were used to modify training delivery and develop Train-the-Trainer content to improve effectiveness.

The training for all sessions was conducted face to face with class sizes ranging from 6 to 40 participants. The Level 2 pre-/ and post-test assessments consisted of MCQ and T/F questions and the participants were required to complete the pre-test assessment as soon as they arrived, and these were collected prior to the start of training. On completion of the training the same assessment was administered as the post-test assessment. Additionally, a Level 1 assessment was administered to determine the participant's satisfaction level with the training. Figure 1-4 illustrates the various utility industries and the details of the assessments that the employees and the managers received. The assessments consisted of MCQ and T/F questions that related to the

concepts relevant to that specific industry for all the pre-/post-test assessments. In addition to these questions, in 4 out of the 6 cases, a Control Question (CQ) was added to the MCQ. The CQ is a question that is contextually similar to the content trained, however the specific details of this item were not taught in the class (Samuel, Azen & Campbell-Kyureghyan, 2018). Further explanations of the CQ are provided in Section 3.1.2.3. As part of the answer options, an IDK option was provided for all MCQ, CQ and T/F questions in 4 out of the 6 training cases. Figure 1-4 details the assessments administered to the participants in the various utility industries in both the pre- and post-test assessments.

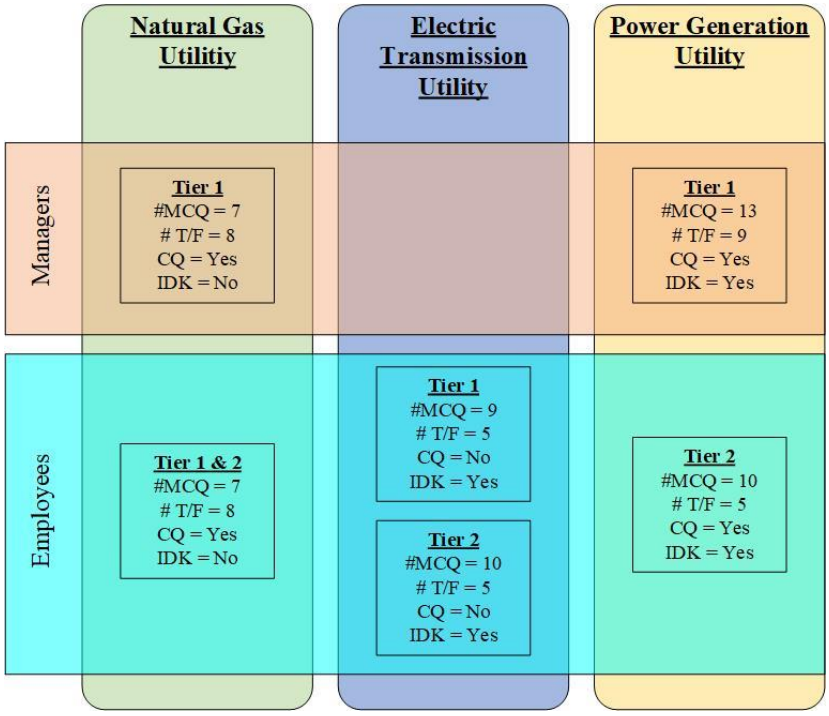

Figure 1-4: Types and number of questions administered in Level 2 pre- and post-training assessments by utility sector and training participants

In Figure 1-4, #MCQ and #T/F indicate the number of Multiple Choice Questions and True/False questions in the assessments respectively, and CQ and IDK indicates if there was or was not a Control Question or 'I Don't Know' option in the assessment.

1.3 Goals of the Research

The overall goal of this research is to improve assessment of training effectiveness by quantifying the effect of guessing and accounting for participant prior knowledge of the concepts delivered during training.

The specific research questions this research aims to answer are:

1. How does the addition of the IDK option in the pre-test Level 2 MCQ assessment change the proportion of correct and incorrect answers?
2. How does the addition of the IDK option in the post-test Level 2 MCQ assessment change the proportion of correct and incorrect answers?
3. Does the addition of the IDK option truly reduce the amount of guessing in pre-test and post-test assessments?
4. If the participant chooses IDK in the pre-test assessment, is there a difference in how that participant responds on the post-test assessment depending on the type of question (MCQ or a Control Question - CQ)
5. How can the learning outcomes for the concepts taught during the training session be assessed?
6. How do the different methods used to measure training effectiveness of concepts in Level 2 assessments in a pre-/ post-test assessment model differ from each other on the concepts they report as best and least learned?

This dissertation is split into six chapters and manuscripts are presented within three of the six chapters that specifically define the goals and outcomes of each study. Chapter 1 introduces the motivation of the research and provides details into the scope of the research. Chapter 2 provides an overview of the current body of literature that exists in measuring training

effectiveness and highlights the research gaps that are addressed in this dissertation. Chapter 3 describes the research that quantifies the effect of the IDK option on guessing in a MCQ pre-/post-test assessment model. Chapter 4 describes a new methodology to measure training effectiveness, Assessment of Training Effectiveness Adjusted for Learning (ATEAL) and Chapter 5 is the application of the ATEAL method to training assessment results from participants that underwent safety training in an industrial setting. Finally, Chapter 6 summarizes the research results and contribution to the field.

1.4 References

Arthur Jr, W., Bennett Jr, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234.

Bahn, S., & Barratt-Pugh, L. (2012). Emerging issues of health and safety training delivery in Australia: Quality and transferability. *Procedia – Social and Behavioral Sciences, 62*, 213-222.

Becker, P., & Morawetz, J. (2004). Impacts of health and safety education: Comparison of worker activities before and after training. *American Journal of Industrial Medicine, 46*(1), 63-70.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065-1105.

Brunello, G., & Medio, A. (2001). An explanation of international differences in education and workplace training. *European Economic Review, 45*(2), 307-322.

Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Salvador, R. O., & Islam, G. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health, 96*(2), 315-324.

Campbell-Kyureghyan, N., Hernandez, A. P., & Ahmed, M. (2013). Effectiveness of first and second tier safety and ergonomics training in power utilities. *Proceedings of the XXVth Annual Occupational Ergonomics and Safety Conference*, Atlanta, GA, USA, June 6-7, 2013.

Campbell-Kyureghyan, N., Ahmed, M., Beschorner, K. (2013, May). *Measuring Training Impact 1-5*. Paper presented at the US DOL Trainer Exchange Meeting, Washington DC, March 12-13, 2013.

Demirkesen, S., & Arditi, D. (2015). Construction safety personnel's perceptions of safety training practices. *International Journal of Project Management, 33*(5), 1160-1169.

Freifeld, L. Training (2018). *2018 Training Industry Report*. Retrieved October 10, 2019 from https://trainingmag.com/trgmag-article/2018-training-industry-report/

Glaveski S. (2019). *Where companies go wrong with learning and development*. Retrieved October 2, 2019 from https://hbr.org/2019/10/where-companies-go-wrong-with-learning-and-development

Ho, C. L., & Dzeng, R. J. (2010). Construction safety training via e-Learning: Learning effectiveness and user satisfaction. *Computers & Education, 55*(2), 858-867.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and Development Handbook* (pp. 40-60). New York: McGraw Hill.

Kontoghiorghes, C. (2001). Factors affecting training effectiveness in the context of the introduction of new technology—a US case study. *International Journal of Training and Development, 5*(4), 248-260.

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2019). Evaluation of learning outcomes through multiple choice pre-and post-training assessments. *Journal of Education and Learning, 8*(3).

Tai, W. T. (2006). Effects of training framing, general self-efficacy and training motivation on trainees' training effectiveness. *Personnel Review, 35*(1), 51-65.

Chapter 2: Critical Literature Review

Understanding and quantifying training effectiveness in any environment where training occurs is of importance to many stakeholders, be it the participants, the trainers or the organization, leaders, parents etc. who have a vested interest in the improvement of the participants undergoing the training.

The following academic and scholarly search engines were used for the literature search: Science Direct, Google Scholar, University of Wisconsin Milwaukee (UWM) Library tools and Research Gate. The process of identifying articles began with broad queries into the above listed databases using a series of relevant keywords. The following keywords were used in pairs or in groups of three or more: training effectiveness, pre-test assessment, post-test assessments, global training, safety training, pre-test post-test assessment, control question, multiple choice questions, true or false, learning assessments, ergonomic training, guessing behavior, formula scoring, number right scoring, I don't know option and adult learning. The review of the literature was conducted on literature written in English only and focused on a time frame from the 1960's to current. Finally, literature related to Item Response Theory (IRT) was excluded from the search due to the large amount of data required for the analysis (Smith & Wagner, 2018) and the fact that test using IRT typically require computer based testing to administer.

A total of over 500 papers were identified through the initial searches and subsequent examination of the articles referenced in those papers. Articles were excluded if there was no quantitative data to illustrate application of a proposed methodology and if the analysis method included the use of IRT, however proposed training effectiveness models without quantitative analysis were included. Literature that involved the use of the Control Question in polygraph tests and associated studies that revolved around detecting incorrect answers in polygraph test

participants were excluded. 48 papers remained after applying the exclusion criteria (Table 2-1) and are included in this literature review.

The purpose of this literature review is to capture the current state of knowledge with regards to measuring and reporting training evaluation and effectiveness. Additionally, even considering the extensive body of knowledge, due to over 50 plus years of research in this field, there is room for research into methods for measuring and reporting on training evaluation and effectiveness to help organizations and training practitioners assess and improve the training process to benefit the associated stakeholders.

2.1 Training Effectiveness

Alliger & Janak (1989) noted that researchers agree that evaluating the effectiveness of the training is important in validating the benefits of the training to the organization. Of the various methods that can be used to measure training effectiveness, Alvarez, Salas & Garofano (2004) and Simkins & Allen (2000) noted that training evaluations or assessments are predominantly used to quantify the benefits of training and to determine its deficiencies. Additionally, these assessment results are used to develop countermeasures to address any observed gaps. Alvarez et al. (2004) also noted that although training evaluation and training effectiveness are distinct concepts, they are closely related and need to be considered together to provide an overall picture.

Arthur et al. (2003) noted that of the many methods and models used to measure training effectiveness, industry trainers most frequently use the Kirkpatrick's model (Kirkpatrick, 1967) that consists of 4 evaluation levels. The model links the reaction (Level 1) of the participants to the training and learning (Level 2) of the participant during the training activity to changes in their behavior (Level 3) in their job performance which leads to results (Level 4) for the

organization through productivity gains etc. The linkage between learning and the application of the knowledge in a work environment was also observed by Kontoghiorghes (2001) and was noted to be a significant finding as it statistically validated the training evaluation component. Post-training assessment of learning, Level 2 assessments, using Multiple-Choice Questions (MCQ) is the most popular training practice to assess participants knowledge of the concepts taught (Alliger & Janak, 1989; Bar-Hillel, Budescu & Attali, 2005; Dimitrov & Rumrill, 2003; Newble, Baxter & Elmslie, 1979). The assessments conducted after the training are labeled as post-test assessments and the performance of the participants is thus measured by the number of questions answered correctly. A refinement of the post-test assessment model is the pre-/post-test model (Level 2) that assess the training participants twice by administering the same assessment before the training (pre-test assessment) and after delivery of the training (post-test assessment). This is done to gage the initial knowledge level of the participants and measure the increase due to the training. This measured change in score was observed to be a preferable method to a single post-training assessment as it identified individual learning and how different trainees have changed due to their prior levels of competencies (Dimitrov & Rumrill, 2003; Warr, Allan & Birdi, 1999).

Baldwin & Ford (1988) conduced an exhaustive review of the literature, identified critical gaps and proposed future direction for the concern of 'transfer problem' during training. The transfer problem is defined as the gap between the amount that industries spend on training and the amount of these expenditures that result in transfer to the job. The researchers state that no more than 10% has been effectively transferred to the job. Through the extensive and detailed reviews done, one of the key gaps identified as it relates to organizational training effectiveness is the need for training research is to ensure job relevance of the training content versus

implicitly assuming relevance of the content without clearly understanding the behaviors and skills clearly that need to be acquired. Additionally, they encourage that all future study on transfer should provide evidence of job relevance of training content prior to examining its effect. These observations of job relevance/usefulness as being an important for transfer sustainment was further confirmed by Axtell, Maitlis & Yearta (2014) who conducted an exploratory investigation by researching the sustainment of transfer of interpersonal skill training in a workplace among 45 trainee respondents assessed from when the training was completed to multiple times within a one year timeframe after completion of training. The key variables explored in the research were (1) the perceived relevance/usefulness of the training (2) self-efficacy (3) motivation (4) managerial support & (5) autonomy existing in the job. In assessing the trainees at multiple times, the researchers were able to determine if there was any correlation and changes to how the participants answered the assessments as time progressed and they spent more time in their job. The results suggested that if the trainees transferred their skills to the work environment within one month after training, it was more sustainable to observe that performance after one year. That is, early skill transfer is a key predictor of long term retention. Contrary to some studies, the participants in this study indicated that the managers did not play a major role on transfer of the training to the work environment. One of the main limitations of this study is that the assessments are qualitative in nature and based on telephonic survey conducted by the researchers with the participants. There were no direct observations of the participants in their job environments and there was no unbiased review of their performance to determine if the training actually resulted in better interpersonal behavior of the participants.

Blume, Ford, Baldwin and Huang (2010) conducted another exhaustive review of "Transfer of training" to update the research conducted 22 years prior by two of the original authors. They

conducted an exhaustive analysis on 89 empirical studies of transfer of training. The researchers coded the studies using 10 different criteria (predictor variables) and focused on whether training transfer was measured as the use of trained skill or the effectiveness of the performance of the trained skill. The Hunter & Schmidt's meta analytic procedure was used to conduct the overall analysis on the relationship between the predictor variables and the transfer. The researchers observed that transfer climate had the highest relationship to training transfer followed by supervisor support and peer support. The observation of supervisor support is at odds with the results reported by Axtell, Maitlis & Yearta (2014), however this could be due to the content that was trained or the methods of measurement. One consistent finding by the researchers was that transfer measured immediately following training yielded consistently stronger relationship with the predictor variables than transfer measured after a time lag. The researchers conclude that the most promising avenues to improve training transfer is to be mindful in selecting training cohorts as peer support is important, focus on increasing the motivation of trainees and induce higher level of supervisor support in the work environment. One of the weaknesses in this research is that it does not compare or critique the various methods of training assessments and how these assessments impact / predict the training transfer of the participants.

Alliger, Tannenbaum, Bennett Jr, Traver & Shotland (1997) conducted an extensive meta-analysis of 111 correlations on the training criteria that affects training effectiveness. The analysis was based on a modified Kirkpatrick's and the authors identified that the reaction measures were most strongly related to on-the-job performance. Although the authors state that the reaction measure showed strong relationship to transfer, they do stress that reaction measures cannot be used as a replacement for other measures such as learning performance. They observe that reaction measures that capture if the participants liked or did not like the training is unrelated

to measures of learning and transfer. However, reaction measures that capture if the participant is more or less likely to apply the training at work demonstrated a moderate relationship between learning and transfer. The main weakness with this meta-analysis, as noted by the authors, is that it is based on Kirkpatrick's model. Alliger and Janak (1989) observed in their meta-analysis that Kirkpatrick's model provides generalized classifications for criteria, however, the easily adopted terminology tends to cause misunderstanding and overgeneralizations in its application. Despite its shortcoming, it was used as it is the most popular model used in industry. Additionally, the meta-analysis did not discuss if different Level 2 assessments methods (only post-test, pre-/post-test, etc..) would be a better or worse predictor of Level 3 performance of the participants.

Alvarez et al., (2004) stated that training effectiveness is a study of variables that influence training outcomes of which training assessments are an integral part and that training experts typically study training effectiveness through the targets of evaluation. Based on the critique of four prior training evaluation models and three prior training effectiveness models, the authors introduced the Integrated Model for Training Evaluation & Effectiveness (IMTEE). A detailed description of the model is beyond the scope of this review, however it is important to note that the authors state that the model incorporates training effectiveness variables from the past 10 years of research with all six training evaluation measures. The model identifies cognitive learning (measured through paper-and-pencil or electronically administered tests) as one of the key inputs to training and transfer performance. However, the model does not elaborate on the differences on the impact a pre-/post-test assessment outcome would have compared to a post-test only evaluation outcome to the training and transfer performance. Additionally, the literature and learning assessment models reviewed did not take into consideration the reporting

of learning results in a disaggregated manner, described later in this section. A critical review of the training effectiveness model developed by Tai (2006) was forthcoming, however, the IMTEE model incorporates all the elements put forward by Tai and has additional resolution to understand many more variables that significantly affect training effectiveness.

As stated by Alvarez et. al, (2004) a critical element in measuring training effectiveness is the method by which the scores themselves are calculated and reported. Significant research detailing methods to calculate score gains (Campbell & Stanley, 1963; Herbig, 1976; Hendrix, Carter & Hintze, 1978; Brogan & Kutner, 1980; van der Linden, 1981; Warr et al., 1999; Dimitrov & Rumrill, 2003; Arthur Jr. et al., 2003) has been conducted and a variety of statistics such as absolute test scores, test score deltas, ANOVA, ANCOVA etc. have been detailed by Dimitrov and Rumrill (2003), Bonate (2000) and Tannebaum and Yukl (1992) to measure the effectiveness of the training. The research by Bonate (2000) is extensive in its explanation of the pre-test / post-test models and the associated statistics that can be conducted along with their assumptions, benefits, expected results and shortcomings. The following is a brief description of the more popular statistical methods to measure change in a pre-test/post-test design:

- ANOVA on gain scores:

  In this analysis, the score delta (D = Post-test score – Pre-test score) represents the dependent variable in ANOVA comparisons of two or more groups. This method can be used to test the null hypothesis of zero mean gain score in the population of test takers. The use of this measurement has been criticized by some researchers due to their assertion that the difference between scores is less reliable than the scores themselves. However, this assertion is only true if the pre-test and post-test scores have the same variances.

- ANCOVA with pre-test & post-test data:

The ANCOVA method with a pre-test/post-test assessment design is to reduce the error variance in non-randomized assignment of subjects to groups. This is especially important when pre-test scores are not reliable as then the treatment effects can be seriously biased in a nonrandomized design. When the pre-test and post-test scores are the same and the regression slope is 1, the F ratio of the ANCOVA and the ANOVA on gains scores is the same. When the slope does not equal 1, the ANCOVA is a much more powerful test.

- ANOVA on residual scores:

  Residual scores are the delta between an observed post-test score and the predicted value from a simple regression using the pre-test score as a predictor. It is observed that residual scores contain less error than gain scores when the variances of the test scores are different. It is important to note that this method is less powerful and much too conservative than the ANCOVA method when the regression coefficient is calculated using the total sample for all groups combined.

It is important to note a practical aspect associated with statistical analysis at this point. All the analysis detailed above require sophisticated statistical software such as Minitab, SAS etc. to conduct the computation and the investment in time and training to use these software packages effectively is not insignificant, even for large organizations. Thus, analysis of this nature is typically limited to a few individuals within the organization. This is a considerable drawback as trainers need to be able to assess the results and derive guidance on improvements that need to be made to the training without having to send the data into an analysis group and wait for weeks to get the results of the analysis, the results of which might be too late to apply in the training environment. In addition to these methods, Walstad and Wagner (2016) introduced a method to deconstruct the pre-/post-test assessment results into four quadrants (positive, negative, retained

and zero) of learning and was able to determine the effectiveness of the training as a whole and separated by each question or concept taught. They also argued that only using post-test scores or delta scores between pre- and post-test is misleading as the scores are influenced by the interaction of the results in the four learning quadrants which confounds the measurement of the results. The post-hoc analysis was conducted on micro and macro test result data from the Test of Understanding in College Economics (TUCE) for a total of 10,997 participants. Regression analysis was conducted by the researchers to model the test score results (pre-test, post-test, Post-Pre and the results of the four learning quadrants) based on student characteristics (demographics) and school variables (colleges, universities etc.). The study focuses on trying to predict the outcomes of the test results based on the demographics and the types of education that the participants possess. It, however, does not provide insight into the how effective the training was for the concepts that were taught within the Micro and Macro Economics courses and how the results from the quadrant analysis can be used to improve the training. Additionally, it does not have an "I don't know" (IDK) option available in the model and there is gap in how this model can be applied to MCQ assessments where IDK is an option.

2.2 Guessing in MCQ Training Assessments

The frequent and predominant criticism of Multiple Choice Question (MCQ) assessments is that it allows for the respondents to choose the correct answer by guessing correctly. This was observed by Newble, Baxter & Elmslie (1979) where they noted significantly higher scores in tests with MCQ assessments than in free response tests. Additionally, this difference was observed to be greater for students with lower level of education and seniority among other medical students. 5th year students were observed to achieve a 61% better score in MCQ than free response compared to 37% for 6th year students. This score gain due to guessing gives an

exaggerated representation of the respondents knowledge level and makes it difficult to accurately assess participant performance based on the gain in knowledge. A large amount of work has been conducted into methods to reduce the effect of guessing through scoring methods, instructions given to participants or by giving options that enable a participant to answer truthfully in the assessment. The next sections evaluate these methods.

2.2.1 Scoring Adjustments: Formula Scoring

"Religion, politics and formula scoring are areas where two informed people often hold opposing ideas with great assurance" (Lord, 1975). Frary (1988) describes formula scoring as a procedure designed to reduce multiple choice score irregularities due to guessing. He states that formula scoring is not designed to penalize guessing but to adjust scores due to random guessing as a result of complete ignorance. Lord (1975) noted that the number-right score differs from formula score by the lucky or unlucky guesses that affect the number-right score. The clarity of the instructions given by the trainers and the interpretation of the participants is critical for the formula-scoring method to have its desired effect. Current formula scoring instructions advice participants against blind guessing but does encourage them to guess whenever they can eliminate a wrong choice. The interpretation of the instructions by participants is critical, since if they believe that they will be penalized for guessing, then they will tend to follow a more conservative strategy than instructed by the formula-scoring instructions Frary (1988).

Edgington (1965) observed trends where the formula scoring of R-W (# of correct answers - # of incorrect answers) on post-test assessments with instructions to not guess, produced a bias in the participants based on their propensity to follow the instructions or not. Based in these observations, he recommended that formulae that correct for guessing should not be used and scoring should be in terms of number of correct answers. Little (1966) conducted research into

the application of formula scoring of R-W/(N-1) where N is the number of possibilities for each of the items in a test, in a pre-test (prior to teaching course)/post-test (at final exam of course)/re-test (8 months after course completion) assessment model and observed that the formula scoring over-corrects the score when the participants are unfamiliar with the concepts and under-corrects the scores when the participants are familiar with the concepts. These over and under corrections are compared to a metric defined as sure-correct response which can only be calculated once all the pre/post and re-test assessments have been conducted. This method is time consuming and is difficult to be applied in industrial training environments where the time is of the essence and results of participant performance and recommendations for improvements need to be made quite quickly after completion of training. Similarly, Davis (1967) introduced a far more complex formula to calculate the score of participants while correcting for guessing. Despite the various scoring methods, it is noted that participant risk tolerance has a measurable effect on their answering characteristics and thus their ability to guess or not guess in the MCQ assessments (Budescu & Bar-Hillel, 1993).

Ebel (1968) devised a method to determine if the participants in objective achievement tests were blindly guessing. For the MCQ assessment questions, the participants were initially asked to answer the questions. Following that the students were given a new answer sheet for the second time, and were asked to choose the response that would be considered equal to blind guessing. This method is tedious and adds little value in an industrial training environment as it is difficult to justify the participants taking an assessment twice just to capture responses for guessing. Additionally, the response choices were limited to T/F.

Collett (1971) introduced the concept of Elimination Scoring and compared the results to Classical number right scoring and Weighted Choice scoring to compensate for guessing. It was

observed that the elimination scoring method (the participant eliminates all the incorrect answers instead of choosing the correct answer), along with the associated formula scoring, enabled it to assess partial knowledge. In reviewing the methodology used and the instructions given to the participants on the grading schemes, it is observed to be cumbersome for the participants. Additionally, the grading instructions would be difficult to explain and be understood by industry participants who are accustomed to choosing the correct answer in a MCQ assessment to show they have learned the concept.

Espinosa and Gardeazabal (2005) conducted research to determine if the answering behavior of participants changes based on the scoring methods that are used to assess their performance. A comparison of number-right scoring and formula-scoring was conducted on post-test assessments of participants attending an undergraduate macroeconomics course. The students were split up into three groups and each group was graded based on a different scoring rule. A total of five assessments were given over the period of the course and three of the assessments were evaluated using some form of formula-scoring while two were evaluated using number-right scoring. Considerable communication and education of the scoring methods was provided to the student groups as their final course grade depended on the results. The results of the assessments were normalized based on the formula-scoring methods adopted and it was observed that the participant behavior was not affected by the scoring rules and the results were consistent with the rational behavior of students to maximize value with the difference between the scoring rules is due to risk aversion by the participants. This method involved significant coaching of the participants on the different scoring methods that would be used over the period of a macroeconomics course in a university setting. The scoring was only made on post-test assessments. Applying a similar approach in an industrial setting could be difficult as the

training durations are much shorter and it is difficult to communicate and convince the participants of the fairness of using different scoring rules for different participants. Additionally, the normalization and the statistics required to determine the scoring are time consuming and do not allow for immediate feedback to the participants on their performance.

Hammond, McIndoe, Sansome & Spargo (1998) conducted MCQ assessments using T/F and an IDK option on examinees. The scoring involved a positive score for a correct answer, a negative score for a wrong answer and a zero score for an IDK. In conducting the MCQ assessment in combination with other assessment methods such as essays and oral examinations, it was observed that the candidates performed considerably worse in MCQ compared to the other methods (30.5% and 50-60% respectively). In order to understand the contributors for the score loss, a confidence option was added for the participants to indicate if they were confident about the answer, making educated guesses or wildly guessing. The results indicated that the participants did not understand the effect of negative scoring and the impact it is meant to have on discouraging guessing and its impact on the overall score. Thus, it was possible for a participant to fail by not answering enough questions rather than losing marks by guessing. Although this study illustrates the overall trends with participant understanding of scoring method other than a number-right scoring, it did not conduct any pre-test assessments and the assessment questions were limited to T/F questions only. A similar study was conducted by Bereby-Meyer, Meyer & Flascher (2002) on using prospect theory to analyze guessing of participants in MCQs. They observed that the participants tended to guess more if they saw that the potential for loss was higher than by omitting a response i.e. if the participants felt that they were not going to be able to meet the minimum requirements score, they guessed more than participants who felt that they were going to meet the minimum score. .

Bar-Hillel, Budescu & Attali (2005) wrote a research paper on the "irrationality" of scoring and keying multiple choice questions and they offer much of the same arguments stated so far in this review. In critiquing they observe the confusion that the formula scoring process creates among test takers and how it benefits the bold test taker who is less concerned about the negative grading. The researchers also state the formula scoring "evokes lay logic in its simple-mindedness and myopia: 'If you want to decrease guessing – penalize it'". However, even under formula scoring, it is never better to omit than to guess thereby the concepts are observed by the researchers as being somewhat self-defeating. Thus, a majority of the drawbacks of formula scoring revolve around the ability of the assessment participants to understand and follow the instructions. Additionally, the participants should understand the relatively complex instructions on when to guess and when not to guess. Formula scoring therefore requires additional time and effort to inform the score users (participants and other stakeholders) of the nature of the effect of scoring and not doing so accurately may be doing them a disservice. These deficiencies are also noted by Budescu & Bar-Hillel (1993) in their critique of formula-scoring and its impact on guessing. Finally, it is of note that Ebel (1965) noted that assessment scores that have been corrected for guessing usually rank students in about the same relative positions as the uncorrected scores, this implies that although the overall score may change, it is not possible to compensate for guessing by using formula scoring when comparing results within participants or within different concepts taught in the same class of participants.

2.2.2 I Don't Know (IDK) Option

Sanderson (1973) assessed the impact of the 'I Don't Know' option in the post-test assessment of T/F questions for medical examinations. Unlike the actual medical examination that included the IDK option, there were only two answer options that the participants could

choose from, True or False. A lack of an answer indicated that the participant did not know the concept and was interpreted as the IDK option. The scoring methodology penalized the participants for an incorrect answer but did not give any score for an IDK choice. The participants were given the examination three successive times and were asked to indicate how they would have answered the question if there had not been an IDK option. At the first assessment participants were asked to mark the answers for which they were sure (definitive), in the second they were asked to also mark the answers for which they were not sure enough to make a firm decision (tentative). It was observed that in the post-test assessment the participants were not guessing at random as the tentative score was significantly increased over the definitive score of the participants. This indicated that the IDK option potentially conceals a small amount of correct knowledge based on the participants personality. Cautious participants would be more likely to choose the IDK option than to guess like the bolder participants would. This research was conducted only on post-test assessments with T/F questions available as answer options. Although the research provides useful insights into the effect the IDK option has on the participants, it does not provide guidance on how the participants would have answered if they did have an IDK option to choose from.

Courtenay and Weidemann (1985) conducted research on the 'Palmore's Facts of Aging Quiz' specifically to determine if the IDK option would reduce guessing among participants. The participants were split into groups, two of which had an IDK option and two that did not. Results showed that the addition of the IDK option significantly reduced the amount of guessing since if there was no IDK option, the participants would have no other option but to guess, and the addition of the IDK option was observed to reduce the number of incorrect answers. The

major gaps in the research is that the assessments were done only on post-test assessments with T/F response options in the MCQ.

Mameren and Vleuten (1999) conducted research into the effect the IDK option has on number-right and formula scoring. Through the assessments conducted on 363 medical students answering 150-180 T/F items, the participants were scored using formula scoring (R-W) initially and then the participants were asked to choose either T/F for items that they had answered IDK to achieve the number right scoring. It was observed that the score was 2.5% - 3.4% lower for the formula scoring than for the number right scoring. This indicated that addition of the IDK option reduced the propensity of the participants to guess in the assessments. They also observed, like Sanderson (1973), that participants who were less willing to guess obtained lower scores. The limitation with the method used is that it only conducted post-test assessments with T/F options. Additionally, it is difficult to make industry participants to go back and change their IDK options into their best guess of an answer as it takes up valuable time and gives the impression that the assessment results are being manipulated.

Spears and Wilson (2010) conducted research into evaluations in an extension education program. They proposed a pre-/post-test assessment model with the inclusion of the IDK option and proposed the usage of the marginal homogeneity test to determine if the distribution of answers in the pre-test differs significantly from the distribution of answers in the post-test in the same subject. They also note that it is important to capture the information of the participants who changed their answer to or from IDK in the assessments. The method analyses one question with T/F & IDK option and a second question which has four answer options without an IDK option. The method proposed does provide a strong ground to understand how the participants

learned during the training, however, it does not explore the effect of the IDK option on guessing and the study is limited to only the T/F assessment type.

Research on the IDK option has predominantly revolved around True or False (T/F) type questions (Sanderson, 1973; Newble et al., 1979; Courtenay & Weidemann, 1985; Hammond et al., 1998; Mameren & Vleuten, 1999; Spears & Wilson, 2010) and there is a lack of a holistic picture regarding the impact of introducing the IDK option in a pre-/post-test assessment method that uses MCQs. Understanding the impact of the IDK option on MCQ testing is of importance as it will help validate if introduction of the IDK option truly reduces guessing and gives a clearer view of the true knowledge level of the participants after training.

Smith & Wagner (2018) explored a method to adjust for guessing in the disaggregation model that was introduced by Walstad & Wagner (2016). The researchers developed mathematical models to calculate expected values of the positive learning, negative learning, retained learning and zero learning. They conducted a Monte Carlo simulation on class sizes from 15 to 300 participants and applied the method to a class of 90 students taught by the author and to the TUCE data discussed previously in this section and used by Walstad & Wagner (2016). They observed that the method adjusts the values in each of the disaggregated quadrants and inferences on improvements to the parts of the training that need improvement are proposed. One of the main drawbacks in this model is that it assumes that the probability of guessing is identical across all students and occurs independently on the pre-test and post-test. The second assumption is based on the fact that the authors assume that enough time passes between the pre-test and post-test assessment for the participants to not remember the responses from the pre-test. The second assumption is particularly troublesome in the case of training and testing industry participants as they have limited time and the organization would prefer that extra time not be

taken from the participants solely for the purpose of an assessment. Thus, pre-/post-test assessments are done immediately before and after the training is conducted in the same session. Additionally, this method does not take into consideration the use of an IDK option in the assessment and the calculation process to determine the adjusted scores needs to be done using a computer as the calculation is intensive. This does not allow for quick feedback to the trainers and the organization on the areas where the participants had the best and worst performance.

2.3 Industry Training:

As detailed prior in Chapter 1, the importance of ensuring that training is effective for safety related content is paramount due to the personal, societal, and financial impacts associated with poor safety in work-environments. As a linkage to training effectiveness and safety training, Demirkesen and Arditi (2015) stressed that safety improvements may not be achieved unless special attention is paid to the effectiveness of learning during the training session.

Becker and Morawetz (2004) conducted research into the determining the effectiveness of a hazardous waste training program as it related to changes in attitudes and post-training activities. They surveyed 55 workers prior to training and after 14-18 months following training. The training program was also intended to create train-the-trainers within the organizations in which the individuals worked. The surveys revolved around interest and involvement in safety activities, attitudes, their ability to make workplace improvements and their involvement in training others in their workplace. The researchers observed that the interest in making changes in safety conditions reduced significantly among the participants as time progressed after training, however, they did observe that the trainees were more active in safety training in the workplace and that the trainees experienced an increased success rate in attempting to implement safety improvements. This study and method, however, does not assess the amount of learning

that the participants had during the training program as part of the post-test assessment and focuses mainly on the long-term effects of the training on the behavior of the participants. Due to the many confounding elements it will be very difficult to make specific changes to the training to improve the learning outcomes based on results obtained using this method.

Bahn and Barratt-Pugh (2012) conducted research into determining how new regulatory frameworks impact organizations specifically in terms of safety training mechanisms in Australia. Additionally, the research was to determine the training mechanisms that had the greatest ability to impact work-related injury rates. The researchers conducted nine semi-structured 30-60 minute interviews with representatives from training offices, unions and accrediting boards. The analysis was more qualitative in nature with the compilation of the conversations into groups that identified areas of structural failure and successes in Australian health and safety training delivery. They observed that due to the complex nature, variety of training providers, level of skills taught and variety of delivery methods, the participants questioned the value of such training and would like to understand if the time and money is well spent on these efforts. This study does not explore the effectiveness of the safety training through the use of quantitative assessments and analysis of participant responses in the assessments. Thus, despite the important qualitative information compiled about perception of the safety training system, there is little in the results regarding specific training concepts that can be said to be improved to improve work-related injury rates.

Ho and Dzeng (2010) investigated the learning effectiveness of safety training delivered via e-Learning due to the increasing popularity in organizations embracing e-Learning methodologies to reduce cost, increase participation and be more readily approachable to the participants. The research they conducted involved surveying the participants on the e-Learning

platform function, content design, network quality, user interface and other factors related to the e-Learning environment and conducting statistical analysis on the results of the surveys versus the specific safety concepts taught in the training session. However, learning effectiveness was defined more based on performance metrics such as learning satisfaction, operation safety and time cost. The researchers concluded that the e-Learning mode proved highly feasible and can reinforce the safety behavior of labor operation for the researched ranged of participant age and educational degree. The research does not delve into the actual safety concepts being taught and how the effectiveness of that training can be improved via specific assessments made on concepts taught in the e-Learning environment. This makes it difficult to use the results to improve the training effectiveness of specific concepts as this research only looks at the overall operation safety metric as a guideline for if the content was relevant and well taught.

Demirkesen and Arditi (2015) noted that high safety performance in construction work environments is achieved with intense safety training and that the efficiency of the safety training programs depends on organizational issues (organizational structure and management's commitment to safety) and the effectiveness of the safety trainers in improving the quality of the training session. The researchers developed a questionnaire to investigate safety personnel perceptions of training strategies in the top 400 contractors in the US. The questionnaire addressed two main themes, first, how the individuals achieve and reinforce learning in worker safety training, and second how language issues were resolved when training non-English speakers on safety content. The first theme had 18 sub-questions and the second had eight. Analysis was conducted on the demographics of the companies (age, size, union or not, location in the US etc.) to determine if there were statistical trends that could be determined between the groups based on their answers to the questions. The researchers found that the surveyed group

(large contractors in the US) were sensitive to organizational, feedback, content, process and worker issues when it related to effectiveness of safety training. Although they state that the organizations need to pay special attention to the effectiveness of learning during the training session, they do not provide a structure (or) methodology to measure and improve the effectiveness in the training session. Further research to improve training effectiveness on specific safety concepts that the participants and organization find critical will be useful to increase the impact of this research and provide tactical guidance to organization and trainers.

Research conducted by Campbell-Kyureghyan and Cooper. (2012) investigated the effectiveness of ergonomic and safety training and the impact that tailoring the training has towards a target demographic group. They conducted a participant recruitment of 11 companies followed by an onsite needs assessment to determine specific ergonomic and safety needs of each company. Training material (booklets, tests & feedback questionnaires) were developed in both Spanish and English for employees and managers based on the onsite needs assessment. The learning assessment consisted of a pre-/post-test model with the same identical MCQs being assessed before and after the training session. Assessment results from 635 participants were analyzed based on overall performance and demographic differences (sex, first language, ethnicity & education). It was observed that all demographic groups increased their test score by an average of 10%. A deeper analysis was conducted between participants for whom English was and was not a first language. This analysis showed that although the pre-test scores were different between native and non-native English speakers, they both had similar test score improvements. One of the main drawbacks of this method is that it although there were pre-/ and post-test assessments conducted that resulted in paired data, a detailed analysis on the change in answers by question or concept was not analyzed. The analysis was based on the overall delta

performance and it is not possible to determine if the primary English speaking participants were more willing to take more risk by guessing in the assessments as compared to the non- English speaking participants. Additionally, an IDK option was not presented in the assessment for the participants to choose from.

Burke, Sarpy, Smith-Crowe, Chan-Serafin, Salvador, & Islam (2006) conducted an extensive meta-analysis to determine the effectiveness of the different methods of safety training in improving safety knowledge and performance. A total of 95 quasi-experimental studies were analyzed from 1971–2003 and the work has been cited in 169 subsequent research studies. The researchers state that this is the first analysis to be conducted on studies published since 1971 in the field of safety training with a scientifically rigorous approach. The studies researched were coded across six criteria ranging from method of safety and health training to country of study. Statistical analysis was conducted to calculate significance in gains and losses because of training and a meta-analysis procedure was used to allow mean effects to be compared across different types of dependent variables. The major finding in this study is that the researchers observed a statistically significant improvement in the effect of safety training as we progress from least engaging to most engaging (3 times more effective) training interventions. However, even the least engaging training (lectures, passive training methods) resulted in an improvement in safety outcomes. The authors note that designing and implementing effective training programs is central to the effort of improving safety behavior among participants. The study does not extend to examining safety and health training specific to safety knowledge (types of injuries etc.) and individual performance. Hence, this is an area where more research is required to quantify the role of specific safety knowledge and how it impacts safety behavior.

Park, Kwak & Chang (2010) researched the effect of food safety training on employees in the hospitality industry in Korea. The effectiveness of the safety training was measured with respect to food safety knowledge, safety practice of employees and inspection of food safety performance. The researchers designed a nonequivalent pre-test and post-test control group method with 41 food handlers in the intervention group and 49 in the control group. The training procedure for the intervention group involved conducting a pre-test assessment along with safety inspection by the researchers and a survey completed by the managers. Following this, a 30-minute food safety training was conducted along with a feedback questionnaire. Approximately 2 weeks later a retraining was conducted for the intervention group using the same content as the first training. Following the retraining the post-test assessment was completed. Statistical analysis was conducted on the difference between the pre-test results between the two groups (no difference) and the post-test results (significant difference) between them. There was also an increase in sanitation knowledge in the intervention group as shown by a 17.3 point score increase from pre-test to post-test scores. However, the intervention group did not show any significant changes in food safety practices. The researchers also conducted a correlation study between the safety knowledge and safety practices and observed a negative correlation in many cases. Despite the limitations detailed by the authors, the study does not probe the reason, as related to training assessment results, as to why there is no impact to the safety behavior of the participants despite showing an increase in safety knowledge. This study would benefit significantly by conducting a detailed analysis on disaggregating the pre-/post-test results and representing them in quadrants as detailed by Walstad and Wagner (2016). Additionally, the authors do not specify the use of an IDK option in the assessments which would have potentially helped determine if there were significant gaps in learning the safety content.

Campbell-Kyureghyan (2013) investigated the effectiveness of hearing conservation training programs. The study introduced an expanded model of the Kirkpatrick four stage model with the inclusion of an assessment step for learning retention. Pre-/ and post-test assessments were conducted on 875 participants (employees and managers) and it was analyzed that 84% of the trainees were able to correctly answer the question related to hearing conservation. However, when the learning retention evaluation was conducted 3-6 months after training, the percentage of trainees who were able to correctly answer the question had dropped to 41% and was the question that had the lowest learning retention of all the questions taught. It was observed that although the participants answered the hearing impact question incorrectly, 98% of the trainees were applying learned skills to their jobs. Further, management implemented administrative rule changes with regards to hearing protection following the initial training and reduction in assessment scores may not be the same as observed behavior by the participants. A drawback in this method is that although the data is available, the analysis does not evaluate the change in answers by the participants for the concept of 'Noise and Hearing Protection' to help determine if the participants were guessing using the IDK option. The analysis is focused solely on the results of the post-test assessments and the follow-up evaluation conducted 3-6 months after training. A description of change in the assessment scored from pre-test to post-test to follow-up evaluation would have been valuable. The answer option for the question under consideration was a T/F question and there was no IDK option in the assessments for the participants to choose from.

Caston, Cooper & Campbell-Kyureghyan (2009) conducted research into the effectiveness of safety training in small business foundries in reducing musculoskeletal injuries by analyzing results of pre- and post-test evaluations. Companies with 250 employees or less were recruited,

training content was developed after on-site evaluations were conducted with managers and employees. The training developed consisted of content for ergonomic awareness (all employees) and deploying ergonomic programs (managers, engineering and production leaders). Pre- and post-test evaluations were conducted before and following the training sessions for the participants. As part of the assessment, a control question was included. This is a question that is contextually similar to the content trained however was not specifically taught in the training sessions. The control question was used to calculate an odds ratio to determine the improvement of knowledge of each participant in each question as compared to control question. A total of 37 participants were trained as part of this program and the results of their assessments on 15 MCQs were analyzed. The questions were grouped into four concepts and the scoring was conducted as delta between the post-test and pre-test scores within the groups. The odds ratio was calculated by question type, however, there is no explicit conclusion on the best and worst learned concepts. Pre- and post-test score analysis was conducted and it was observed that overall the participants maintained or improved a passing score. One drawback in this study is that although it is not explicitly stated, it appears that the researchers are trying to use the control question and the odds ratio methodology to account for guessing by the participants. An explicit statement for the use of this novel methodology would have been helpful to determine the purpose and a scenario analysis would be helpful to further explain the mechanics and interpretation of the odds ratio and the results reported. Finally, the study does not include an IDK option in the assessment for the participants to choose from.

Ahmed and Campbell-Kyureghyan (2014) conducted research into the effectiveness of ergonomic training for small business electric utilities by analyzing training results for 34 trainees. Analysis was done to determine the best and least learned concept and the test question

reliability determined by the measurement of true learning of the participants. Pre- and post-test evaluations were conducted on the trainees using the identical questions and an IDK option was available for the participants with the purpose of eliminating guessing. Stratification of the trainees into groups based on correct and incorrect answers in the pre- and post-test was done. Trainee pass rate improved from 41% to 85% before and after training. Analysis was conducted to determine the best learned concept by grouping participants who answered incorrectly in the pre-test and correctly in the post-test. Additionally, an assessment of the reliability of each training concept was conducted by calculating the proportion of trainees answering the questions correctly divided by the number who answered the question, while accounting for guessing. The study could benefit from specific definition as to how the effect of guessing was determined.

Ahmed and Campbell-Kyureghyan (2017) researched the effectiveness of safety and ergonomic training in the wind energy sector. Training content was developed based on onsite observations of specific risks, interviews, ergonomic principles and review of injury records. A pre-/post-test training assessment model was followed and an IDK option was available for the participants to choose in all cases. A total of 16 trainees participated in the training and a feedback questionnaire was administered to determine the participant's satisfaction with the training. Average score gains for the training program were reported and an analysis was conducted to determine the trainees who did not answer correctly on one of the key concepts (struck by/caught between). Follow-up interviews were conducted 18 months after the training to evaluate its effectiveness in preventing injuries. The results indicated that the training was effective, with no struck by/caught between injuries occurring in the period after training. One weakness of the study is that no on-site observations were performed to see if the training resulted in behavioral changes that lead to the reduction in injuries. Additionally, as the IDK

option was used in the assessments, there is an opportunity to determine the amount of guessing as was done by the same authors in 2014.

2.4 Identified Literature Gaps

The previously published research body reviewed is not intended to be exhaustive of all existing research on the subject of measuring and improving training effectiveness. However, it is believed to be a comprehensive summary of the research that has been performed in this discipline related to the inclusion criteria such as training effectiveness, safety training, IDK option etc. as stated in the beginning of this chapter. As mentioned, IRT, publications in language other than English and literature prior to the 1960's has been omitted due to the reasons stated before and not addressed in this critical review.

The gaps in literature vary as each study had a different focus and a different limitation. Two of the six research questions (#1 and #2) of this study involve the impact that addition of the IDK option has on how participants respond in pre- and post-test MCQ assessments. Two other research questions (#3 and #4) relate to measuring the effect of the IDK option on participant guessing in MCQ assessments. Based on the literature reviewed, it can be seen that the IDK option has been proposed as a way to minimize guessing among the training participants and produce a score that is more representative of the participants true knowledge level. However, all the studies except one reviewed only focus on the use of the IDK option where the assessment responses were T/F. Thus, there is also a lack of research in the quantification in the amount of guessing that the addition of the IDK option reduces in a pre-test/post-test assessment model where there are more than just two assessment options. In addition to the reduction in guessing, there is limited understanding of how the addition of the IDK option changes the proportion of correct and incorrect answers in pre-test and post-test Level 2 MCQ assessments. Finally, there

is a lack of research on how participants who choose the IDK option in the pre-test assessment respond in the post-test assessments based on concepts that they have learned (MCQ) or not learned (CQ).

Two of the research questions of this study (#5 and #6) relate to assessing the learning outcomes of the concepts taught and to determine how the methods used to measure training effectiveness of concepts in Level 2 assessments in a pre-/ post-test assessment model differ from each other on the concepts they report as best and least learned. The literature reviewed illustrates the measurement of training effectiveness using assessments as either how much a student knows at a certain time or change in knowledge over time as measured by pre-test/post-test assessments. The works by Walstad & Wagner (2016) and subsequently refined by Smith and Wagner (2018) are the only research that breaks down the answers to create new learning variables that give more resolution into the learning of participants. However, even these research works do not take into account the effect of an IDK option in the pre-test and post-test assessments. Additionally, there is no research on how the positive learning scores can be adjusted for prior knowledge and negative training impacts to help the trainers and the organizations determine the best and worst learned concepts.

A core aspect of the goal of this research revolves around improving the assessment of training effectiveness by quantifying the effect of guessing and accounting for participant prior knowledge of the concepts trained. In the literature reviewed with regards to safety related training that has been done in industry, it is observed that none of the studies researched the effect of addition of the IDK option on guessing in the assessments among the participants. In the instances when scores are measured and reported, only post-test assessment scores (or) delta scores between pre- and post-test assessments are reported. Applying the method of

disaggregation to safety training in an industrial setting will have great benefits helping the trainers and the organizations best utilize their resources to improve training effectiveness. Thus, there is a research gap that when addressed will help to improve the assessment of training effectiveness by applying the method of disaggregation and accounting for the participant prior knowledge of concepts delivered during training. This quantification of training effectiveness will also enable determination of the best and least learned concepts

Addressing the aforementioned gaps in the research will help improve training effectiveness by helping trainers choose the best method to assess the training, quantifying the effect of guessing and accounting for prior knowledge of the concepts trained. These improvements will enable organizations to focus on the concepts for which the participants have the most knowledge gaps and focus on those rather than on concepts for which the participants had high prior knowledge. This will thereby maximize the impact of the time spent in training by the participants and improve the return on investment of the training for the organization.

Table 2-1 illustrates the compilation of the literature review into the categories pertinent to this research. These categories were chosen based on the research goals of this study. The categories of training effectiveness and predicting/measuring guessing, pre-test assessment, post-test assessment, MCQ, T/F & IDK option are related to the overall research goal of improving the assessment of training effectiveness by quantifying the effect of guessing and accounting for participant prior knowledge of the concepts delivered during training. The topic of scoring analysis method is related to the research question of how the learning outcomes for the concepts taught during the training session are best assessed. Finally, the topics of training industry participants and safety / ergonomic training is related to the funding grants to reduce occupational industries. An 'X' indicates that the paper reviewed had some content presented

with regards to that specific category. As can be seen from the table, no literature that has been reviewed from post 1960, with the inclusion of numerous meta-analyses, addresses all the categories chosen. A majority of the reviewed literature does not consider the effect of guessing while training effectiveness is being investigated. 40% of the studies focus solely on post-test assessments to determine learning of participants, 44% of the studies focus on both pre-test and post-test assessments and 16% do not focus on type of assessments. Only 15% of the literature included an investigation of the IDK option and measured its effects on reducing guessing among training participants. Of the studies that included the use of the IDK option, a majority (71%) used only T/F post-test training assessments and did not consider training for industry participants or for safety and ergonomic related training. It was observed that only a small portion (27%) of the literature reviewed included the application of the methodology for industry participants. The research presented here aims to fill these gaps and create a new body of knowledge to improve training effectiveness in industry training applications.

Table 2-1: Summary of topics covered for papers included in the literature review

Author(s)	Year	Training Effectiveness	Predicting Guessing	Scoring Analysis Method	Pre-Test Assessment	Post-Test Assessment	MCQ	T/F	I Don't Know	Training to Industry Participants	Safety / Ergonomic Training
Edgington	1965		X	X		X	X	X			
Little	1966		X	X	X	X	X	X			
Davis	1967		X	X		X	X	X			
Kirkpatrick	1967	X									
Ebel	1968		X			X	X	X			
Collet	1971		X	X		X	X	X			
Sanderson	1973					X		X	X		
Lord	1975		X			X	X	X			
Herbig	1976			X	X	X	X	X			
Hendrix et al.	1979			X	X	X	X	X			
Newble et al	1979	X		X		X	X	X			
Brogan et al	1980			X	X	X					
van der Linden	1981		X	X	X	X	X	X			
Courtenay et al.	1985		X		X	X		X	X		
Frary	1988		X			X	X	X			
Alliger et al.	1989	X									
Budescu et al.	1993		X	X		X	X	X			
Baldwin et al.	1994	X									
Axtell et al.	1997	X			X	X	X	X			
Alliger et al.	1997	X									
Hammond et al.	1998		X			X		X	X		
Mameren et al.	1999		X			X		X	X		
Warr et al.	1999	X			X	X	X	X		X	
Kontoghiorghes	2001	X				X	X			X	

Table 2-1: Summary of topics covered for papers included in the literature review (cont.)

Bereby-Meyer et al.	2002		X			X	X				
Arthur Jr. et al.	2003	X									
Dimitrov et al.	2003			X	X	X					
Espinosa et al.	2005		X			X	X	X			
Spears et al.	2010				X	X		X	X		
Walstad et al.	2016	X		X	X	X	X	X			
Smith et al.	2018	X	X	X	X	X	X	X			
Simkins et al.	2000				X	X	X				
Tai	2004	X				X	X	X			
Alvarez et al.	2004	X									
Blume et al.	2010	X									
Sung-Hee et al.	2010	X			X	X	X			X	X
Burke et al.	2006	X								X	X
Bar-Hillel et al.	2005		X			X					
Demirkesen et al.	2015	X				X	X			X	X
Ho et al.	2010	X				X	X			X	X
Bahn et al.	2012	X				X				X	X
Bonate	2000			X	X	X	X	X			
Becker et al.	2004	X			X	X	X			X	X
Campbell-Kyureghyan et al.	2012	X		X	X	X	X	X		X	X
Campbell-Kyureghyan	2013	X			X	X		X		X	X
Caston et al.	2009	X		X	X	X	X	X		X	X
Ahmed et al.	2014	X			X	X	X	X	X	X	X
Ahmed et al.	2017	X			X	X	X	X	X	X	X

2.5 References

Ahmed, M., Campbell-Kyureghyan, N. (2014). *Reliability of learning assessment.* Proceedings of the XXVIth Annual International Occupational Ergonomics & Safety Conference, El Paso, TX, June 5-6, 2014

Ahmed, M., Campbell-Kyureghyan, N. (2017). *Injury-specific novel safety training helps to reduce injuries in the renewable energy generation sector*. Proceedings of XXIXth Annual Occupational Ergonomics and Safety Conference, Seattle, WA USA, 2017

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel psychology*, *42*(2), 331-342.

Alliger, G. M., Tannenbaum, S. I., Bennett Jr, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel psychology*, *50*(2), 341-358.

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human resource development Review*, *3*(4), 385-416.

Arthur Jr, W., Bennett Jr, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied psychology*, *88*(2), 234.

Axtell, C. M., Maitlis, S., & Yearta, S. K. (1997). Predicting immediate and longer-term transfer of training. *Personnel Review*.

Bahn, S., & Barratt-Pugh, L. (2012). Emerging issues of Health and Safety training delivery in Australia: Quality and transferability. *Procedia-Social and Behavioral Sciences*, *62*, 213-222.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*(1), 63-105.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, *4*(1), 3-12.

Becker, P., & Morawetz, J. (2004). Impacts of health and safety education: Comparison of worker activities before and after training. *American Journal of Industrial Medicine*, *46*(1), 63-70.

Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, *15*(4), 313-327.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065-1105.

Bonate, P. L. (2000). Analysis of pretest-posttest designs. ISBN 1-58488-173-9

Brogan, D. R., & Kutner, M. H. (1980). Comparative analyses of pretest-posttest research designs. *The American Statistician*, *34*(4), 229-232.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277-291.

Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Salvador, R. O., & Islam, G. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health*, *96*(2), 315-324.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research* (No. 04; Q175, C3.). Boston: Houghton Mifflin.

Campbell-Kyureghyan, N., Cooper, K. (2012). *Impact of customized training on learning across demographic groups*. JPIIE, 9(1): 25-31.

Campbell-Kyureghyan, N. (2013). *Evaluation of the impact of hearing conservation training*. CAOHS, 4:1-4.

Caston, S., Campbell-Kyureghyan, N. , Cooper, K. (2009) *Assessment of ergonomic training in small business foundries.* Proceedings of International Occupational Ergonomics and Safety Society Meeting, pp: 1-6 , 2009

Collet, L. S. (1971). Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, *8*(3), 209-214.

Courtenay, B. C., & Weidemann, C. (1985). The effects of a “don't know” response on Palmore's Facts on Aging quizzes. *The Gerontologist*, *25*(2), 177-181.

Davis, F. B. (1967). A note on the correction for chance success. *The Journal of Experimental Education*, *35*(3), 42-47.

Demirkesen, S., & Arditi, D. (2015). Construction safety personnel's perceptions of safety training practices. *International Journal of Project Management*, *33*(5), 1160-1169.

Dimitrov, D. M., & Rumrill Jr, P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*(2), 159-165.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs.

Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement*, *5*(4), 321-325.

Edgington, E. S. (1965). Scoring formulas that “Correct for Guessing”. *The Journal of Experimental Education*, *33*(4), 345-346.

Espinosa, M. P., & Gardeazabal, J. (2005). Do students behave rationally in multiple choice tests. *Evidence from a Field Experiment. V Mimeograph, Universidad del Pais Vasco*.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33-38.

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: Adopting an evidence-based approach to exam technique. *Anaesthesia*, *53*(11), 1105-110 8.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1978). A comparison of five statistical methods for analyzing pretest-posttest designs. *The Journal of Experimental Education*, *47*(2), 96-102.

Herbig, M. (1976). Item analysis by use in pre-tests and post-tests: A comparison of different coefficients. *Programmed Learning and Educational Technology*, *13*(2), 49-54.

Ho, C. L., & Dzeng, R. J. (2010). Construction safety training via e-Learning: Learning effectiveness and user satisfaction. *Computers & Education*, *55*(2), 858-867.

Kirkpatrick, D. L. (1967). Evaluation of training. *In R. L. Craig & L. R. Bittel (Eds.), Training and Development Handbook (pp. 40–60)*. New York: McGraw Hill.

Kontoghiorghes, C. (2001). Factors affecting training effectiveness in the context of the introduction of new technology—a US case study. *International Journal of Training and Development*, *5*(4), 248-260.

Little, E. B. (1966). Overcorrection and undercorrection in multiple-choice test scoring. *The Journal of Experimental Education*, *35*(1), 44-47.

Lord, F. M. (1975). Formula scoring and number right scoring. *Journal of Educational Measurement*, *12*(1), 7-11.

Mameren, H. V., & Vleuten, C. V. D. (1999). The effect of a 'don't know' option on test scores: Number-right and formula scoring compared. *Medical education*, *33*(4), 267-275.

Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education*, *13*(4), 263-268.

Park, S. H., Kwak, T. K., & Chang, H. J. (2010). Evaluation of the food safety training for food handlers in restaurant operations. *Nutrition Research and Practice*, *4*(1), 58-68.

Sanderson, P. H. (1973). The 'don't know' option in MCQ examinations. *Medical Education*, *7*(1), 25-29.

Simkins, S., & Allen, S. (2000). Pretesting students to improve teaching and learning. *International Advances in Economic Research*, *6*(1), 100-112.

Smith, B. O., & Wagner, J. (2018). Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores. *The Journal of Economic Education*, *49*(4), 307-323.

Spears, K., & Wilson, M. (2010). "I don't know" and multiple choice analysis of pre- and post-tests. *Journal of Extension*, *48*(6tt2)

Tai, W. T. (2006). Effects of training framing, general self-efficacy and training motivation on trainees' training effectiveness. *Personnel Review, 35*(1), 51-65.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*(1), 399-441.

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research*, *51*(3), 379-402.

Walstad, W. B., & Wagner, J. (2016). The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education*, *47*(2), 121-131.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology*, *72*(3), 351-375.

Chapter 3: Evaluation of learning outcomes through multiple choice pre- and post- training assessments

Thomas Samuel, Razia Azen & Naira Campbell-Kyureghyan

Abstract

Training programs, in industry, are a common way to increase awareness and change the behavior of individuals. The most popular way to determine the effectiveness of the training on learning outcomes is to administer assessments with Multiple Choice Questions (MCQ) to the participants, despite the fact that in this type of assessment it is difficult to separate true learning from guessing. This study specifically aims to quantify the effect of the inclusion of the 'I don't know' (IDK) option on learning outcomes in a pre-/post-test assessment construct by introducing a 'Control Question' (CQ). The analysis was performed on training conducted for 1,474 participants. Results show a statistically significant reduction in the usage of the IDK option in the post-test assessment as compared to the pre-test assessment for all questions including the Control Question. This illustrates that participants are learning concepts taught in the training sessions but are also prone to guess more in the post-test assessment as compared to the pre-test assessment.

Keywords: pre-test assessment, post-test assessment, multiple choice question, I don't know, control question, training assessment, adult learning, guessing behavior.

3.1 Introduction

Training individuals is a common way for organizations to increase the knowledge of their workforce in specific competencies. Based on the Industry Report from 2000, US organizations with 100 or more employees budgeted to spend $54 billion in formal training (Arthur Jr., Bennett Jr., Edens & Bell, 2003). These trends for formal training are also observed in Australia (Bahn & Barratt-Pugh, 2012) and have been reported to play an important role in how companies

perceive that they can improve the safety of their employees and reduce incident rates. Overall in 2014 worldwide corporate spending on training was estimated at $170 billion (Bersin, 2014). As a significant amount of money is being dedicated annually around the globe to employee skill development and required changes in behavior, it is important to measure and verify the impact of the training. As a best practice for validating the benefits of training to the organizations, researchers agree on the importance of evaluating training effectiveness (Alliger & Janak, 1998). Although training programs are utilized worldwide (Calle, 2016), evaluation of the training methods is limited in non-Western countries (Ningdyah, 2018).

Of the many methods that can be used to measure the effectiveness of training, Kirkpatrick's model (Kirkpatrick, 1967) remains the one most frequently utilized by trainers (Arthur Jr. et al., 2003). The model consists of 4 evaluation levels as follows:

Level 1 – Reaction: Assessed by asking the trainees how they liked and felt about the training

Level 2 – Learning: Assessed by results of traditional tests of declarative knowledge

Level 3 – Behavior: Assessed by on-the-job performance (i.e., work samples, work outputs and outcomes)

Level 4 – Results: Assessed by organizational impact (i.e., productivity gains, customer satisfaction, cost savings).

Kontoghiorghes (2001) demonstrated that learning in a training setting, as measured by post-test assessments, is a good predictor of how people will apply their knowledge in their work environment. It was also shown that there is a high correlation between the retention of the training material after training and follow up post-test scores. The author concluded that this can be considered a significant finding given that it statistically validates the importance of the training evaluation component, as has been advocated by many human resource development

theorists. This finding also suggests that trainees will be more motivated to learn during training if they know that they are accountable for the training that they receive (Kontoghiorghes, 2001). Similarly, the methods used in the training are also important to help drive the required change in knowledge, attitude and behavior among the trainees. A meta-analysis of training related literature conducted by Burke & Baldwin (1999) concluded that any method that encourages engagement, dialog, and participation of the training participants was more effective than passive methods of training delivery like lectures, online training, and so on. A study by Campbell-Kyureghyan, Principe & Ahmed (2013) found that this method of participatory training, where participants can directly relate the learned material to their jobs, was shown to be effective at reducing work-related injuries. Campbell-Kyureghyan, Ahmed & Beschorner (2013) more importantly observed that dynamic work environments, where traditional approaches of workstation redesigns are not effective, are environments where there is an increased need for contextualized safety and ergonomic training to provide awareness, enhance knowledge, and change the attitude and behavior of the participants as it relates to job site safety.

Immediate post-training assessments of learning, Kirkpatrick's Level 2 assessment, are a common training practice. Knowledge is assessed by multiple choice test responses, answers to open-ended questions, listing of facts, and so forth. That is, trainees are asked to indicate, in one of several ways, how much they know about the topics trained. Alliger & Janak (1998) and Newble, Baxter & Elmslie (1979) indicate that traditional tests in the form of multiple choice questions are by far the most common to assess the knowledge gained

One of the frequent criticisms of Multiple Choice Question (MCQ) assessments is that they enable examinees to answer correctly by guessing. Many trainers and companies view any score gain from guessing as an incorrect representation of the participant's knowledge, which can

negatively affect the participant's performance in a job environment. Also, multiple choice scores are generally perceived to be too high because scores from comparable short-answer or fill-in-the-blanks tests were found to be lower (Newble et al., 1979). Thus, it is important to have a grading procedure that accurately estimates the true score of the individual by accounting for guessing (Frary, 1998). Guessing can be interpreted by the illustration in Figure 3-1 and is defined here by the scenario where a participant does not know the answer yet answers the MCQ correctly. This is troublesome because guessing the correct answer artificially increases the score of the participant and is not an accurate measure of the participant's knowledge level of the subject. Hence, in any MCQ assessment, it is desirable to minimize the cases where the participant does not know the answer and yet answers correctly.

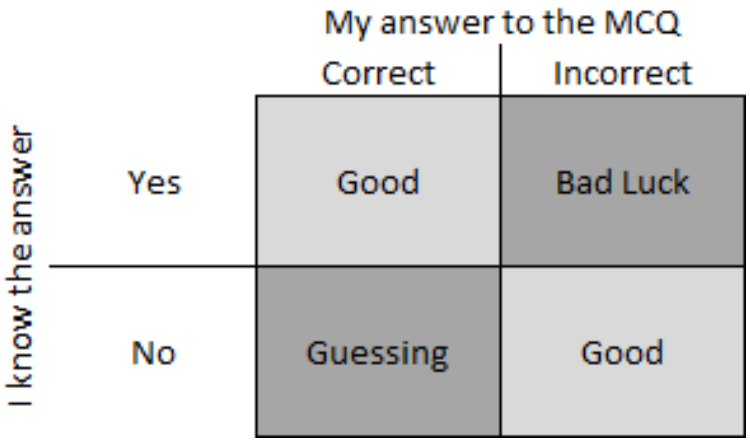

Figure 3-1: Outcomes of MCQ based answers based on the participant knowledge level

An extension of the post-test assessment model is defined by a pre-/post-test model (Level 2) that is essentially assessing participants twice. The pre-test assessment is administered before the training to gage the initial level of knowledge the participant has (baseline), and the post-test assessment is administered after the delivery of the training to gage the increase in knowledge due to training. Initial and final scores of the participants are tracked to determine change in assessment scores. Warr, Allan & Birdi (1999) observed that it is preferable to measure training outcomes in terms of changes from pre-test to post-test, rather than merely through post-test only

scores, as this explains individual learning and an understanding of how different trainees have changed as a result of their experiences. This is because there are often prior differences between trainees in the level of competence that they bring to the training. Although there is literature to illustrate methods to calculate score gains (Campbell, Stanley & Gage, 1963; Herbig, 1976; Hendrix, Carter & Hintze, 1978; Brogan & Kutner, 1980; van der Linden, 1981; Warr et al., 1999; Dimitrov & Rumrill, 2003; Arthur Jr. et al., 2003), there is a gap in the body of knowledge on using the pre-test/post-test method to predict correct guessing of answers on training assessments.

As a method to minimize guessing, a number of authors have suggested adding an 'I don't know' (IDK) option to the true-false answer choices in MCQ assessments (Sanderson, 1973; Newble et. al., 1979; Courtenay & Weidemann, 1985; Hammond, McIndoe, Sansome & Spargo, 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). For example, van Mameren & van der Vleuten (1999) suggested the formula (total # correct answers) – (total # incorrect answers) for the score, with no penalty for IDK answers. Research conducted by Courtenay & Weidemann (1985) indicates that inclusion of the IDK option reduces the overall score of the respondents by 2.5% to 3.4% depending on the tests that were administered and decreases the percentage of questions that are answered incorrectly. Thus, the use of the IDK option is believed to compensate for guessing and increase the likelihood of a more accurate score.

A majority of the research on the IDK option has been conducted in the context of True or False (T/F) type questions (Sanderson, 1973; Newble et al., 1979; Courtenay & Weidemann, 1985; Hammond et al., 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). The work by Newble et al. (1979) included 19 multiple choice items in a post-test only

assessment with an IDK option, but a gap in knowledge still exists on how the IDK option applies to MCQ with more than 2 options in a pre-/post-test assessment model. Therefore, the main goal of the research paper is to investigate and quantify the effect of the IDK option on guessing in a MCQ pre-/post- training assessment model.

The specific research questions (RQ) this study aims to answer are:

- RQ #1: How does the addition of the IDK option in the pre-test Level 2 MCQ assessment changes the proportion of correct and incorrect answers?
    - With the addition of the IDK option, we would expect the percentage of correct answers to stay the same and the percentage of incorrect answers to be reduced.
- RQ #2: How does the addition of the IDK option in the post-test Level 2 MCQ assessment changes the proportion of correct and incorrect answers?
    - With the addition of the IDK option, we would expect a reduction in the percentage of correct answers and a reduction in the percentage of incorrect answers.
- RQ #3: Does the addition of the IDK option truly reduce the amount of guessing in pre-test and post-test assessments?
    - With the addition of the IDK option, we would expect participants to choose the IDK option instead of guessing on questions to which they do not know the answer.
- RQ #4: If the participant chooses IDK in the pre-test assessment, is there a difference in how that participant responds on the post-test assessment depending on the type of question (MCQ or a Control Question - CQ) – Details of the CQ are discussed in detail in the 'Methods' section below.
    - For an MCQ, we would expect most of the participants to answer correctly in the post-test assessment if they answered IDK in the pre-test assessment.

- For a CQ, we would expect that most of the participants to answer IDK in the post-test assessment if they answered IDK in the pre-test assessment.

3.2 Method

A novel training method on workplace safety and ergonomics was developed for multiple sectors of the utility industry under a DOL Susan Harwood Training Grant by the team of researchers at the University of Wisconsin-Milwaukee. Training content was developed from a combination of onsite assessment observations, employee and management interviews, management concerns, ergonomic principles, nationwide injury and fatality records specific to the utility industry and known problematic operations and tasks. Table 3-1 illustrates the number of companies and participants that were trained in the three energy utility sectors.

Table 3-1: List of the number of companies and training participants in each industry

UTILITY SECTOR	# OF COMPANIES	# OF PARTICIPANTS	PARTICIPANT ROLE
Natural Gas	16	*Tier 1:* 500	Employee: 414 Manager: 86
		Tier 2: 375	Employee: 375
Electric Transmission	15	*Tier 1:* 61	Employee: 54 Manager: 7
		Tier 2: 359	Employee: 359
Power Generation	4	*Tier 1:* 22	Employee: 8 Manager: 14
		Tier 2: 157	Employee: 157

To understand and re-define the ergonomic risks, particularly specific to small business utilities, onsite visits were conducted rather than relying solely on general ergonomic principles that are relevant to that utility. Data was gathered from managers/employee interviews and direct observation of all performed tasks using videotaping methods. Since the recruited utilities provide different services, utilize different tools, and are exposed to various ranges of risk-factors, the onsite visits identified the specific ergonomic risks and safety concerns of interest for each facility. The collected information was analyzed and combined with information acquired

from nationwide injury and fatality statistics for the utility industry. The basic ergonomic risk factors and safety concerns present in utilities were identified from the observational data (Campbell-Kyureghyan et al., 2013).

The onsite training was split up into two distinct categories. Tier 1 training was conducted by the individuals who conducted the onsite visit and developed the training content. Tier 2 training was conducted by individuals who had participated in a train-the-trainer program conducted by the Tier 1 trainers. In each company both employees and managers were trained and their respective counts are detailed in Table 3-1. All employees received a base training of 4-5 hours. In addition, managers received an extra 2 hours of training specific to workplace risk assessment and program implementation. It is to be noted that Tier 1 trainers delivered first-hand training to both employees and managers, and Tier 2 trainers conducted primarily employee training.

3.2.1 Training Content

Newly developed content was based on research that specifically targeted the areas of safety and ergonomics of companies, utilities and contractors. All examples and applications in the training were based on the medium to high risk of injury utility-specific tasks that were observed and assessed with the applicable ergonomic methods and tools during onsite visits. Risk factors were classified into the following categories: physical factors such as lifting heavy loads, pushing/pulling, exposure to vibration, or awkward postures, and environmental factors such as exposure to heat or cold, noise, or slippery conditions. The training materials were organized in separate modules: slips/trips/falls, overexertion/repetitive injuries, noise, environment, PPE, and vehicle safety. The materials were developed with a diverse audience in mind, including some employees with less than a high school education or with English as a second language.

3.2.2 Training Assessments

Out of Kirkpatrick's 4 levels of assessments mentioned previously, only the first 2 levels are used in the current study. Due to a very diverse range of trainees with respect to prior competence on ergonomic concepts, years of experience, learning skills, etc., a pre-test and post-test model of training assessment was used.

The mode of training for all session was face to face with the number of participants ranging from 6-40. Both pre-test (baseline) and post-test assessments, using MCQ items, were administered to determine the knowledge of the delivered content that each individual acquired. Participants for all the training sessions were required to complete a 10-15 minute pre-training assessment (pre-test) as soon as they arrived for the training. Once the pre-test assessment was completed by all the participants, they were collected by a training team member for further analysis and the training session commenced. Upon completion of the training the same assessment items were administered to the participants post-test. Table 3-2 illustrates the number of multiple choice questions in the pre-test and post-test training assessments for each of the utility sectors based on the role of the participant.

Table 3-2 List of the number of assessment questions for managers and employees in each utility sector

UTILITY SECTORS	PARTICIPANT ROLE	# OF MCQs IN ASSESSMENT
Natural Gas	Employee	7
	Manager	7
Electric Transmission	Employee	9
	Manager	12
Power Generation	Employee	10
	Manager	13

Finally, the participants were given a Level 1 training reaction assessment that consisted of eight questions to determine the training quality, trainer quality, training material, training process, and the intent of the individuals to apply their new knowledge to their work environment.

3.2.3 Knowledge Testing

Control question (CQ) and IDK option: One question in both the pre- and post-test assessments was a question that was contextually similar to the content being trained in the session; however, that specific item was not covered in the training class. For example, the content of the training consisted of information applicable to most common risk factors present in every energy utility sector (natural gas and electric transmission and power generation) such as: slips/trips/falls, overexertion/repetitive injuries, noise, environment, PPE, and vehicle safety. For the assessment, the control question was NOT related to the content of the training, such as application of the NIOSH lifting equation in the case of employee training, and the selection of appropriate anthropometric measurements for office furniture design in the case of management training. In the CQ model, it is reasonable to assume that a correctly answered Control Question is not a consequence of the training, but rather can be explained by prior knowledge, or guessing.

During the pre-test and post-test assessments for the electric transmission and power generation utility sectors, participants were given an additional 'I don't know' option for each MCQ in addition to the CQ. Participants were instructed to choose the 'I don't know' options instead of guessing at the answers in both assessments. Table 3-3 summarizes the usage of the CQ and the 'I don't know' option in the various assessments for each energy utility sector.

Table 3-3: Usage of CQ and IDK option in MCQ assessments by utility sector

UTILITY SECTOR	TRAINEE TYPE	MCQ ASSESSMENT	
		CQ	IDK
Natural Gas	Tier 1 employee	x	
	Tier 1 Manager	x	
	Tier 2 employee	x	
Electric Transmission	Tier 1 employee		x
	Tier 1 Manager		x
	Tier 2 employee	x	x
Power Generation	Tier 1 employee	x	x
	Tier 1 Manager	x	x
	Tier 2 employee	x	x

3.2.4 Analysis

The data from the all pre-and post-test results (Level 2), as well as the feedback questionnaire (Level 1) were compiled for analysis, and the percentages of correct, incorrect and IDK usage were calculated for the MCQs and the CQs for all the utility sectors.

For research questions 1-3, we define 'P' as the proportion of correct answers out of the total number of questions answered. The first subscript (Y or N) indicates whether the IDK option was available and the second subscript (1 or 2) indicates whether the assessment was pre-test or post-test assessment, respectively. We define 'Q' as the proportion of incorrect answers out of the total, using the same subscripts. In cases (such as research question 3) where only control questions (CQs) were analyzed, this is indicated by a third subscript (C). So, for example, P_{Y2C} would indicate the proportion of CQs answered correctly (of the total number of CQs) on the post-test where there was an IDK option. We define 'I' as the proportion of IDK option chosen using the same subscripts. These definitions are summarized in Table 3-4.

Table 3-4: Summary of proportions used for the analysis

QUESTION TYPE	ASSESSMENT	IDK	PROPORTION CORRECT	PROPORTION INCORRECT	PROPORTION IDK
MCQs	Pre-Test	Yes	P_{Y1}	Q_{Y1}	I_{Y1}
		No	P_{N1}	Q_{N1}	
	Post-Test	Yes	P_{Y2}	Q_{Y2}	I_{Y2}
		No	P_{N2}	Q_{N2}	
CQs Only	Pre-Test	Yes	P_{Y1C}	Q_{Y1C}	I_{Y1C}
		No	P_{N1C}	Q_{N1C}	
	Post-Test	Yes	P_{Y2C}	Q_{Y2C}	I_{Y2C}
		No	P_{N2C}	Q_{N2C}	

Statistical analysis was performed using Minitab 17 (State College, PA, USA). Two-tailed two-proportion z-tests were conducted with a level of significance (α) of 0.05 for statistical analysis of all hypothesis that are detailed for each RQ below.

RQ#1: In order to quantify the impact of IDK addition to all MCQs on the pre-test assessment, the percentage of correct and incorrect answers were compared between two training groups, one of which did not have the IDK option in the pre-tests. Statistical analysis was performed for difference in percentage of correct (H_0: $P_{Y1} - P_{N1} = 0$) and incorrect (H_0: $Q_{Y1} - Q_{N1} = 0$) answers on the pre-tests with and without the IDK option.

RQ #2: Similar to research question 1, the effectiveness of IDK addition to all MSQs on the post-test was evaluated by comparing the percentage of correct and incorrect answers in the post-training assessment of two groups, one of which didn't have the IDK option. Statistical analysis was performed for two hypotheses: (H_0: $P_{Y2} - P_{N2} = 0$) and (H_0: $Q_{Y2} - Q_{N2} = 0$).

RQ #3: To understand if the addition of the IDK option truly reduces the amount of guessing in pre- and post-training assessments, the percentage of correct, incorrect and IDK answers for the CQ in the pre- and post-training tests were compared between two groups, one of which did not have the IDK option on their tests. Statistical analysis of difference between the percentage of correct (H_0: $P_{Y1C} - P_{N1C} = 0$) and incorrect (H_0: $Q_{Y1C} - Q_{N1C} = 0$) answers on the pre-tests for the CQ with and without the IDK option was conducted. Similar analysis was performed on the posts-tests between the percentage of correct (H_0: $P_{Y2C} - P_{N2C} = 0$) and incorrect (H_0: $Q_{Y2C} - Q_{N2C} = 0$) answers. Finally, statistical significance was tested for a difference in the percentage of IDK answers between the pre-test and the post-test for the CQ with and without the IDK option (H_0: $I_{Y1C} - I_{Y2C} = 0$).

RQ #4: To determine the difference in post-test response between MCQ and CQ if IDK was chosen during the pre-test we define P as a proportion out of the total pre-test questions answered IDK. The first subscript indicates whether the post-test answer (which was IDK on the pre-test) was correct (a), incorrect (b), or IDK (c). When only control questions (CQs) were analyzed,

this is indicated by a second subscript (C). So, for example, if $P_{bC} = 0.3$, this would indicate that 30% of CQs answered IDK on the pre-test were changed to an incorrect answer on the post-test. These definitions are summarized in Table 3-5.

Table 3-5: Summary of proportions used for research question 4.

QUESTION TYPE	POST-TEST ANSWER	PROPORTION CHANGED FROM PRE-TEST IDK
MCQs answered IDK on pre-test	Correct	P_a
	Incorrect	P_b
	IDK	P_c
CQs answered IDK on pre-test	Correct	P_{aC}
	Incorrect	P_{bC}
	IDK	P_{cC}

Then, based on this smaller data set, we examined each participant's response on the same question in the post-test assessment, and grouped them into 3 groups: 'Pre-IDK to post-Correct', 'Pre-IDK to post-Incorrect' and 'Pre-IDK to post-IDK'. Statistical analysis was conducted to test the difference in the percentage of IDK answers on the pre-tests that changed to correct (H_0: $P_a - P_{aC} = 0$), incorrect (H_0: $P_b - P_{bC} = 0$) or IDK (H_0: $P_c - P_{cC} = 0$) answers on the post-tests for all MCQs and CQ.

3.3 Results

The 1474 study participants well represented general demographics of the utility workforce in the US, with a majority (over 90%) males and none of the participants had an issue with literacy. More than half (54.3%) of participants reported having no prior ergonomic training, and most (71%) worked at the same company more than five years. The detailed demographics of the participants in the various training sessions are provided in Table 3-6.

Table 3-6: Demographic information of the training participants from each utility sector

	UTILITY SECTOR						Total (n)
	Natural Gas		Electric Transmission		Power Generation		
	Tier 1	Tier 2	Tier 1	Tier 2	Tier 1	Tier 2	
Number of Participants (n)	500	375	61	359	22	157	1474
Gender							
Male	94.9%	86.8%	100%	94.9%	90%	91.6%	1365
Female	5.10%	13.2%	0%	5.1%	10%	8.4%	99
Ethnicity							
African American	1.5%	0%	3.4%	0%	0%	0%	10
American Indian	0%	0%	3.5%	0%	0%	0%	2
White, Non-Hispanic	94.8%	95.5%	91.2%	93.5%	95%	96.1%	1395
Multi-ethnic Background	0%	0%	0%	3.1%	0%	0%	11
Other	3.7%	4.5%	0%	3.4%	5%	3.9%	55
Level of education							
HS Diploma / GED	42.5%	25.9%	20%	35.9%	10%	9.9%	468
Some college	27.2%	34.2%	43.6%	32.4%	20%	8.6%	425
2-Year degree	20%	36.9%	23.6%	18.7%	25%	59.3%	419
4-Year degree	3.7%	0%	9.1%	6.9%	40%	16%	83
Higher degree	2.8%	0%	0%	3.1%	0%	3.7%	31
Other	3.9%	3%	3.6%	3.1%	5%	2.5%	49
Prior Ergo Training							
No	58.2%	52.1%	52.7%	53.5%	55%	54.3%	808
Yes	41.8%	47.9%	47.3%	44.1%	45%	40.7%	650
Years with Company							
<1	3.4%	8.6%	10.5%	19.9%	0%	7.5%	139
1-5	19.6%	25.4%	0%	27%	20%	10%	310
6-10	13.1%	19.2%	15.8%	23.2%	10%	35%	288
11-15	12.3%	11.5%	10.5%	8.2%	15%	13.8%	165
16-20	11.2%	6.2%	12.3%	17.2%	15%	8.8%	166
20+	38.7%	28.9%	38.6%	0%	40%	22.5%	370

To understand the trends in answering the MCQs in the pre- and post-test assessments, Table 3-7 details the percentage and counts of the answers that had been answered correctly, incorrectly, and IDK (when applicable) and these percentages have been aligned with the previously defined variables

Table 3-7: Percentage of correct, incorrect and IDK answers in pre-test assessment

QUESTION TYPE	ASSESSMENT	IDK	PROPORTION CORRECT*	PROPORTION INCORRECT*	PROPORTION IDK*
MCQs	Pre-Test	Yes	P_{Y1} = 66% (n = 1661)	Q_{Y1} = 30% (n = 765)	I_{Y1} = 3% (n = 87)
		No	P_{N1} = 42% (n = 2111)	Q_{N1} = 58% (n = 2929)	
	Post-Test	Yes	P_{Y2} = 83% (n = 2094)	Q_{Y2} = 16% (n = 402)	I_{Y2} = 1% (n = 17)
		No	P_{N2} = 80% (n = 4031)	Q_{N2} = 20% (n = 1009)	
CQs Only	Pre-Test	Yes	P_{Y1C} = 14% (n = 68)	Q_{Y1C} = 24% (n = 116)	I_{Y1C} = 62% (n = 297)
		No	P_{N1C} = 12% (n = 103)	Q_{N1C} = 88% (n = 727)	
	Post-Test	Yes	P_{Y2C} = 40% (n = 190)	Q_{Y2C} = 27% (n = 128)	I_{Y2C} = 34% (n = 163)
		No	P_{N2C} = 24% (n = 203)	Q_{N2C} = 76% (n = 627)	

**Where 'n' is the number of questions*

The results for RQ #1 indicate that there was a statistically significant difference ($z = 20.65$; $p < 0.05$) of 24% between the percentage of correct pre-test MCQ answers with (P_{Y1} = 66%) and without (P_{N1} = 42%) the IDK option. In addition, there was on average a 28% statistically significant difference ($z = -24.04$; $p < 0.05$) observed in the percentage of incorrect pre-test MCQ answers with (Q_{Y1} = 30%) and without (Q_{N1} = 58%) the IDK option. Figure 3-2 illustrates the trends in the percentage changes of correct, incorrect, and IDK answers in the pre-test assessment for the MCQ with the addition of the IDK option.

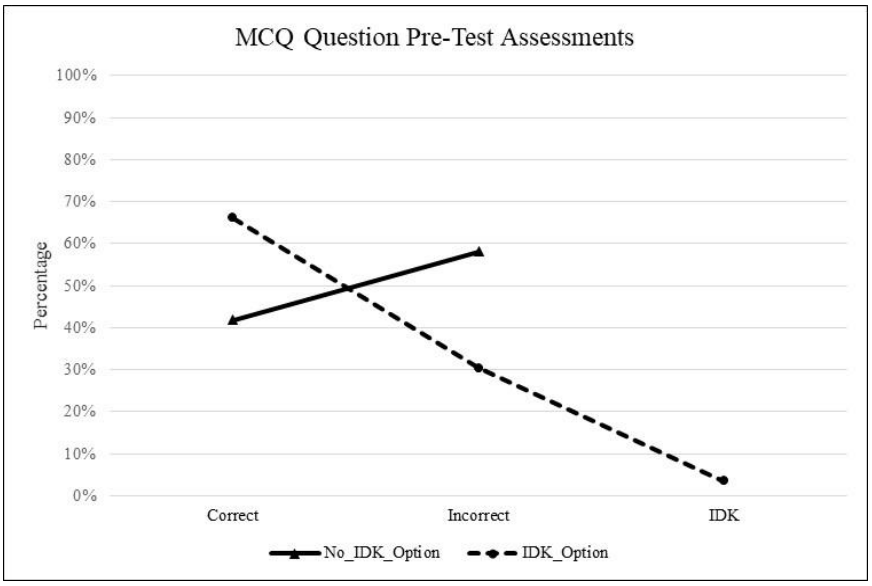

Figure 3-2: Percentage of questions that were answered Correct, Incorrect and IDK in the pre-test assessment for MCQs

While the difference between two groups of trainees (with or without IDK option) were similar, the results for RQ #2 indicate that, there was a 3% statistically significant difference ($z = 3.59$; $p < 0.05$) in correct post-test MCQ answers with ($P_{Y2} = 83\%$) and without ($P_{N2} = 80\%$) the IDK option. Furthermore, a 4% difference ($z = -4.36$; $p < 0.05$) was observed in the percentage of incorrect post-test MCQ answers with ($Q_{Y2} = 16\%$) and without ($Q_{N2} = 20\%$) the IDK option. The trends in in the percentage changes of correct, incorrect, and IDK answers in the post-test assessment for the MCQ with the addition of the IDK option are illustrated in Figure 3-3.

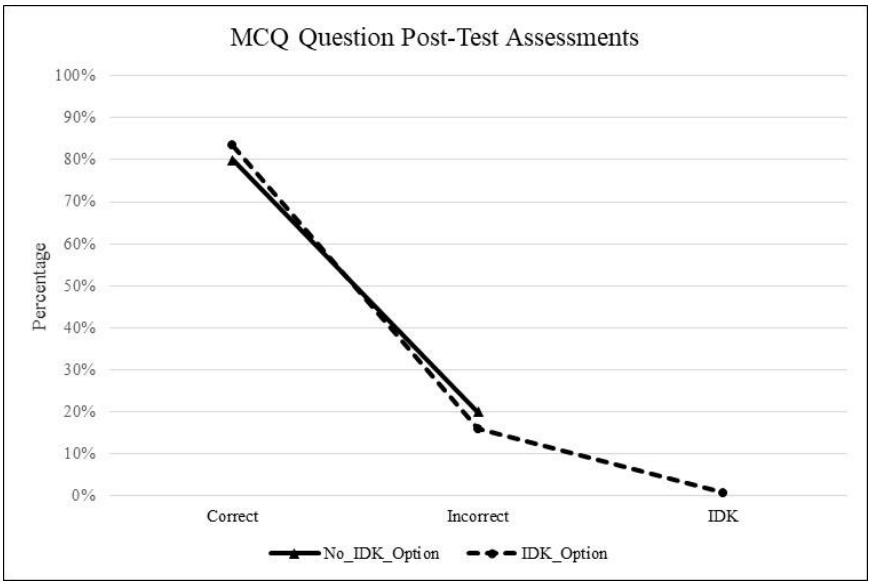

Figure 3-3: Percentage of questions that were answered Correct, Incorrect and IDK in the post-test assessment for MCQs

The pre-test assessment results for RQ #3 revealed no statistically significant difference ($z = 0.88$; $p > 0.05$) in the percentage of correct pre-test CQ answers with ($P_{Y1C} = 14\%$) and without ($P_{N1C} = 12\%$) the IDK option. Nevertheless, a 63.4% difference ($z = -28.07$; $p < 0.05$) was detected in the percentage of incorrect pre-test CQ answers with ($Q_{Y1C} = 24\%$) and without ($Q_{N1C} = 88\%$) the IDK option. The trends in percentages of correct, incorrect, and IDK answers in the pre-test assessment for the CQ with the addition of the IDK options are illustrated in Figure 3-4.

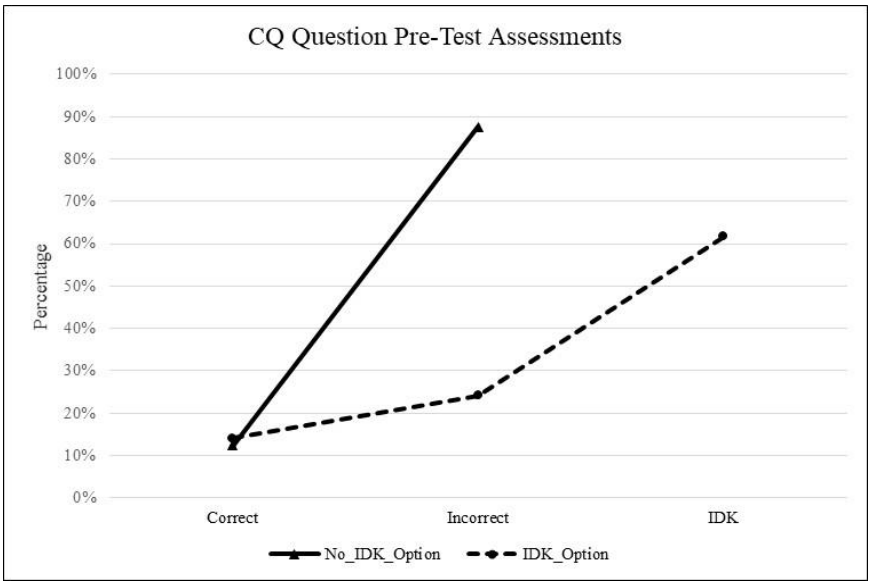

Figure 3-4: Percentage of Correct, Incorrect and IDK answers for the control question for pre-test assessments

In the post-test assessments there was a statistically significant difference ($z = 5.61$; $p < 0.05$) of 16% in the percentage of correct post-test CQ answers observed with ($P_{Y2C} = 40\%$) and without ($P_{N2C} = 24\%$) the IDK option. In addition, there was a 49% difference ($z = -19.52$; $p < 0.05$) observed in the percentage of incorrect post-test CQ answers with ($Q_{Y2C} = 27\%$) and without ($Q_{N2C} = 76\%$) the IDK option. The trends in the percentage changes of correct, incorrect, and IDK answers in the post-test assessment for the CQ with the addition of the IDK options are presented in Figure 3-5.

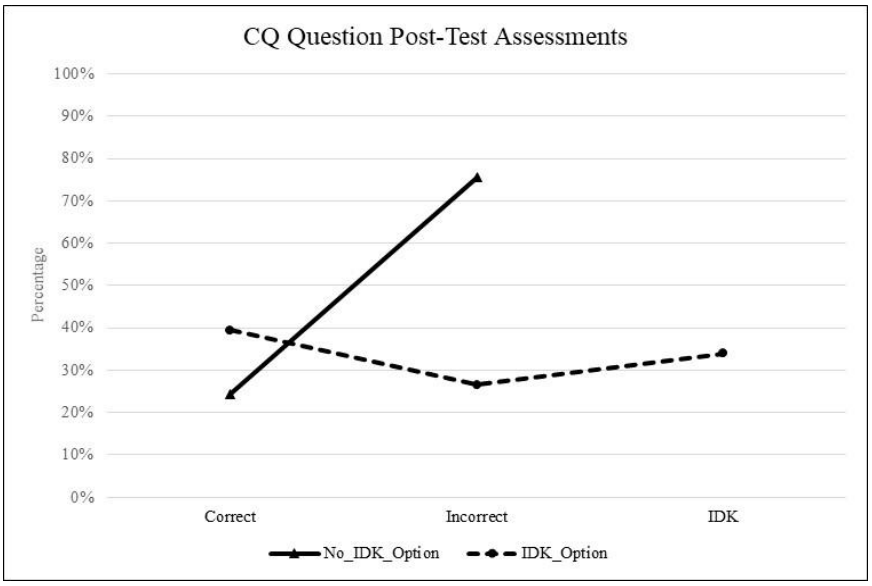

Figure 3-5: Percentage of Correct, Incorrect and IDK answers for the control question in the various training groups for post-test assessments

Comparing the selection of the IDK option in the CQ between the pre- and post-test assessment results indicated that there was a statistically significant difference ($z = 9.01$; $p < 0.05$) of 28% in the percentage of the IDK answers in CQ between the pre-test ($I_{Y1C} = 62\%$) post-test ($I_{Y2C} = 34\%$) assessments.

In summary, we observe that:

- The addition of the IDK option decreases the percentage of incorrect answers in the pre-test assessment for both the MCQ and the CQ.
- There is a statistically significant reduction in the usage of the IDK option in the post-test assessment for both MCQ and CQ. This is expected for MCQ as the contents were taught in the training session. This is not expected in the CQ and the content was not taught to the participants.

Further analysis of the post-test assessments with respect to RQ #4 revealed some interesting insights that are helpful for understanding the trainees in post-training reaction to IDK option on the test. Table 3-8 summarizes the percentage and counts of the questions that were answered as

IDK in the pre-test assessment and then changed to either correct, incorrect or IDK in the post-test assessment.

Table 3-8: Change of state for questions answered as IDK in the pre-test assessment

QUESTION TYPE	POST-TEST ANSWER	PROPORTION CHANGED FROM PRE-TEST IDK*
MCQs answered IDK on pre-test	Correct	P_a = 60% (n = 52)
	Incorrect	P_b = 28% (n = 24)
	IDK	P_c = 13% (n = 11)
CQs answered IDK on pre-test	Correct	P_{aC} = 31% (n = 91)
	Incorrect	P_{bC} = 21% (n = 61)
	IDK	P_{cC} = 49% (n = 145)

**Where 'n' is the number of participants*

A statistically significant difference ($z = 4.94$; $p < 0.05$) of 29% in the percentage of answers that changed from IDK in the pre-test assessment to correct in the post-test assessment for MCQ (P_a = 60%) and CQ (P_{aC} = 31%) was observed. However, there was no statistically significant difference ($z = 1.32$; $p > 0.05$) in the percentage of answers that changed from IDK in the pre-test assessment to incorrect in the post-test assessment for MCQ (P_b = 28%) and CQ (P_{bC} = 21%). Finally, a 36% difference ($z = -7.87$; $p < 0.05$) was observed in the percentage of answers that did not change from IDK in the pre-test and post-test assessments for MCQ (P_c = 13%) and CQ (P_{cC} = 49%).

In summary we observe that:

- For MCQs and CQ, 61% and 30% of the participants respectively, changed from IDK in the pre-test assessment to the correct answer in the post-test assessment. This is expected in the case of the MCQ but not expected in the case of the CQ. Thus it illustrates that some of the

participants are able to guess the right answer instead of answering IDK in the post-test assessment.

- For MCQs and CQ, 28% and 21% of the participants respectively, changed from IDK in the pre-test assessment to the incorrect answer in the post-test assessment. This implies that about the same percentage of individuals are not attentive in the training and answer the questions incorrectly in the post-test assessments or choose not to use the IDK option.
- For MCQs and CQ, 13% and 49% of the participants respectively, did not change their IDK choice in the pre-test and the post-test assessment. This implies that for MCQs a small percentage of participants did not learn the concepts taught and were honest in answering IDK in the post-test assessment. For the CQ, a large portion of the participants were honest in answering IDK in the post-test assessment.
- It is of note that in the CQ, 51% of the participants still chose to change their answer from IDK in the pre-test to either correct or incorrect in the post-test even though the concept was not taught. i.e. 51% of the participants would rather guess at an answer in the post-test assessment rather than answer IDK even though they answered as IDK in the pre-test assessment.

3.4 Discussion

The analysis conducted illustrates some interesting behavioral trends observed in participants with respect to guessing on MCQ pre- and post-training assessments. Several prior studies demonstrated that the concept of adding IDK to only a True/False assessment model helped to minimize guessing on the post-tests (Sanderson, 1973; Newble et. al., 1979; Courtenay & Weidemann, 1985; Hammond, McIndoe, Sansome & Spargo, 1998; van Mameren & van der Vleuten, 1999; Spears & Wilson, 2010). As mentioned before, the major issue with the previous

studies is that their methodology does not allow for true assessment of the training effectiveness. Additionally, since the baseline knowledge was not assessed prior to the training, and control questions were not utilized, it was impossible to separate true learning from guessing on the same group of participants.

The current study design allows these gaps to be filled-in through investigating four main research questions. The first two were specifically addressing the "benefits" of adding an IDK option in pre- and post-test assessments respectively. Based on the results of this study a significant decrease in the percentage of incorrect answers (27%) with the addition of the IDK option to pre-tests is observed. This can simply be explained by a behavioral change, since there is no expectation for a participant to know the correct answer, therefore IDK becomes the best option for the questions about which they have no prior knowledge. In the post-test assessment for MCQs we see a much smaller, although statistically significant, difference (approx. 3% - 4%) in the percentage of correct and incorrect answers with the addition of the IDK option.
While it is expected that the proportion of IDK answers on the post-training assessment will be reduced due to gained knowledge, the participants who did not get a perfect score on the post-test did not chose the IDK option instead of guessing. This became further evident while analyzing the response to the CQ and comparing the difference between pre-test and post-test assessments. For a MCQ we expected most of the participants to answer correctly in the post-test assessment if they answered IDK in the pre-test assessment. Nevertheless, for a CQ, we expected most of the participants to answer IDK in the post-test assessment if they answered IDK in the pre-test assessment.

In the pre-test assessment for the CQ, with the addition of the IDK option, we observe no statistically significant difference in the percentage of correct answers but observe a significant

decrease (63.4%) in the percentage of incorrect answers. This implies that participants, in the pre-test assessment, are very open to answering IDK to a question to which they do not know the answer. In the post-test assessment we observe a 15.1% increase in the percentage of correct answers and a 50% reduction in the percentage of incorrect answers. Additionally, we observe a 28% reduction in the usage of the IDK option between the pre-test and post-test assessments in the case of the CQ. From years of conducting training for adults in various utility industries, this is completely expected as it would indicate that the participants learned the concepts taught and were able to correctly answer the MCQs in the post-test assessment. However, a concerning observation is that we see a significant reduction in the percentage usage of the IDK option from the pre-test to post-test assessment for the CQ as well. Since this question was not taught during any of the training sessions, it helps expose participant guessing behaviors while answering MCQs.

To quantify how participants who answered IDK in the pre-test assessment for MCQs and CQ changed their answers in the post-test assessment, thus answering research question 4, we observe that the MCQs have a 29% higher conversion from IDK to a correct answer than the CQ. There was no difference in the percentage of conversion from IDK to incorrect answers and participants are 36% more likely to answer IDK again in the post-test analysis in the case of a CQ. This implies that most of the participants are learning the concepts taught if they come into the training session not knowing the concept. The troubling finding is that 51% of the participants who answered IDK to the CQ in the pre-test assessment changed their answer and were willing to guess on the post-test assessment.

The findings with regards to the CQ are at odds with what one would typically expect in a training environment. Since the concept in the CQ is not taught in the class, we would expect a

similar percentage of IDK option usage in both the pre-test assessment and the post-test assessment. To get a better understanding of what is occurring in the CQ, the comparisons made between the assessments with and without the IDK option is very telling on participant behavior. In the pre-test assessment, for the CQ, we see that the addition of the IDK option does not impact the percentage of correct answers but helps significantly reduce the percentage of incorrect answers. So, although there is some guessing, it gives an opportunity for the participant to truly express their knowledge level. In the post-test assessment, for the CQ, addition of the IDK option does not have the same impact. There is a significant reduction in the usage of the IDK option, even though the CQ tests a concept that is not taught in the training session. This implies that participants would rather guess at an answer in the post-test assessment than answer IDK, even if they did not know the correct answer. This behavior has been observed and reported among adults and children (Waterman, Blades & Spenser, 2004, Howie & O'Neill, 1996) and was discussed as a significant impactor of business decisions and reported in a Freakonomics radio podcast (Lechberg, 2014).

The more important interpretation of the overall results is that the addition of the IDK option does not significantly reduce the amount of guessing in the post-test assessment and is at odds with the findings from the various authors detailed in the literature review (Sanderson, 1973, Newble et al., 1979, Courtenay & Weidemann, 1985, Hammond et. al. 1998, van Mameren & van der Vleuten), who have stated that incorporation of the IDK option minimizes guessing and can be used as an alternate method to formula scoring. The IDK option, however, is quite effective at helping understand the incoming knowledge level of the participants when administered in the pre-test assessment and can be viewed as a powerful tool to help the instructors modify course content and delivery methods to suit the individual class group needs.

One of the limitations of this study is that the results of different groups (with IDK option and without IDK option) are compared. The commonality is that the training content is related to safety in their utility industry and that the CQ in all cases was not taught during the training session. Also, in the current study it was not possible to conduct Computer Based Testing (CBT) for the participants as the training was conducted at various site locations with some level of computer illiteracy, as well as due to the time constraints available to conduct the training which made setting up computers for each training session not a viable option. Finally, in this study it was not possible to use a formula scoring model to minimize guessing mainly due to the confusing nature of the Formula scoring models and the associated risk of confusing the participants. The time constraints in the training sessions was rather short, and it was not possible to clearly explain the Formula Scoring method to the participants in the assessment.

3.5 Conclusion / Future Direction

This research study investigated and quantified the impact of the IDK option on learning outcomes through MCQ pre- and post-training assessments. A concept called the 'Control Question (CQ)' was introduced in both the pre- and the post-test assessments and is akin to the administration of a placebo treatment since the concept tested by the CQ was not covered in the training sessions. The trends in answers seen in the CQ were compared to those seen in the other MCQs that were taught in the training sessions.

The introduction of the IDK option in the pre-test assessment was observed to statistically significantly reduce incorrect answers by 63% and can be used to help trainers cater the content and delivery to focus on the concepts in which the participants have the largest gaps of knowledge. Nevertheless, the IDK option was not observed to significantly reduce the amount of guessing in the post-test assessment as shown by the change in states measured in the CQ.

Some recommendations that can be derived from this study are:

- Both pre- test assessment before the training and post-test assessment after the training should be administered in order to allow for better assessment of training effectiveness.
- Utilizing MCQs instead of T/F questions decreases the probability of getting a correct answer due to guessing on both pre- and post-test assessments and therefore improves true estimate of learning.
- Conducting the pre-test assessment with the participants prior to the training session and allowing some time to analyze the results before the training may be helpful for the trainers to assess the specific topics that should given greater emphasis during the training.
- Having a dialog on the knowledge gaps to help the training session be more interactive and pertinent to each class will ensure that the trainees get the most out of the training session.
- Being aware that adding an IDK option to the pre-tests was shown to significantly reduce guessing, while the on the post-tests the effect was not as pronounced.
- Using a control (placebo) question(s) on pre- and post-tests can be helpful with generating estimates of the probability of guessing and allow better estimates of true learning.

Acknowledgments

This study was partially funded by the US DOL Susan Harwood Grants: SH-20840-SH0; SH-22220-SH1; SH-23568-SH2. The authors also express their gratitude to Karen Cooper, Sruthi Boda and Madiha Ahmed for assisting with the test development and administration. We would additionally like to thank all the companies and employees who participated in this study.

3.6 References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42*(2), 331-342.

Arciniegas Calle, M. C., Lobelo, F., Jimenez, M. A., Paez, D. C., Cores, S., de Lima, A., & Duperly, J. (2016). One-day workshop-based training improves physical activity prescription knowledge in Latin American physicians: a pre-test posttest study. *BMC Public Health* 16, 1224-35.

Arthur Jr, W., Bennett Jr, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234.

Brogan, D. R., & Kutner, M. H. (1980). Comparative analyses of pretest-posttest research designs. *The American Statistician, 34*(4), 229-232.

Bahn, S., & Barratt-Pugh, L. (2012). Emerging issues of health and safety training delivery in Australia: Quality and transferability. *Procedia – Social and Behavioral Sciences, 62*, 213-222.

Bersin, J. (2014). *Spending on corporate training soars: Employee capabilities now a priority*. Retrieved October 31 , 2018 from https://www.forbes.com/sites/joshbersin/2014/02/04/the-recovery-arrives-corporate-training-spend-skyrockets/#3e38e97dc5a7

Burke, L. A., & Baldwin, T. T. (1999). Workforce training transfer: a study of the effect of relapse prevention training and transfer climate. *Human Resource Management, 38*(3), 227-242.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research* (No. 04; Q175, C3.). Boston: Houghton Mifflin.

Campbell-Kyureghyan, N., Principe, A. H., & Ahmed, M. (2013). Effectiveness of first and second tier safety and ergonomics training in power utilities. *Proceedings of the XXVth Annual Occupational Ergonomics and Safety Conference*, Atlanta, GA, USA, June 6-7, 2013.

Campbell-Kyureghyan, N., Ahmed, M., Beschorner, K. *Measuring Training Impact 1-5*. Paper presented at the US DOL Trainer Exchange Meeting, Washington DC, March 12-13, 2013.

Courtenay, B. C., & Weidemann, C. (1985). The effects of a “don't know” response on Palmore's facts on aging Quizzes. *The Gerontologist, 25*(2), 177-181.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work-Andover Medical Publishers Incorporated. IOS PRESS-, 20*(2), 159-165.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, 7*(2), 33-38.

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: adopting an evidence-based approach to exam technique. *Anaesthesia, 53*(11), 1105-1108.

Hendrix, L. J., Carter, M. W., & Hintze, J. L. (1978). A comparison of five statistical methods for analyzing pretest-posttest designs. *The Journal of Experimental Education, 47*(2), 96-102.

Herbig, M. (1976). Item analysis by use in pre-tests and post-tests: A comparison of different coefficients. *Programmed Learning and Educational Technology, 13*(2), 49-54.

Howie, P., & O'Neill, A. (1996). *Monitoring and reporting lack of knowledge: Developmental changes in the ability to say "I don't know" when appropriate*. Paper presented at the 31st Annual Conference of the Australian Psychological Society, Sydney, Australia.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and Development Handbook* (pp. 40-60). New York: McGraw Hill.

Kontoghiorghes, C. (2001). Factors affecting training effectiveness in the context of the introduction of new technology—a US case study. *International Journal of Training and Development*, 5(4), 248-260.

Lechberg, S. (2014). *The three hardest words in the English language: a new freakonomics radio podcast.* Retrieved June 10, 2017, from http://freakonomics.com/podcast/the-three-hardest-words-in-the-english-language-a-new-freakonomics-radio-podcast/

Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*(4), 263-268.

Ningdyah, A. E., Greenwood, K. M., & Kidd, G. (2018). A training-model scale's validity and reliability coefficients: expert evaluation in Indonesian professional psychology programs. *Makara Human Behavior Studies in Asia*, 22(1), 56-66.

Sanderson, P. H. (1973). The 'don't know' option in MCQ examinations. *Medical Education, 7*(1), 25-29.

Spears, K., & Wilson, M. (2010). *"I don't know" and Multiple Choice Analysis of Pre- and Post-Tests.* Retrieved August 6, 2015, from

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research, 51*(3), 379-402.

van Mameren, H., & van der Vleuten, C. P. M. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education, 33*(4), 267-275.

Waterman, A. H., Blades, M., Spencer, C., (2004). Indicating when you do not know the answer: The effect of question format and interviewer knowledge on children's 'don't know' responses. *British Journal of Developmental Psychology, 22*, 335-348.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*(3), 351-375.

Chapter 4: Assessment of Training Effectiveness Adjusted for Learning (ATEAL). Part I: Method Development and Validation

Thomas Samuel, Razia Azen & Naira Campbell-Kyureghyan

Abstract

Training programs are a popular method, in industry globally, to increase awareness of desired concepts to employees and employers and play a critical part in changing or supporting performance improvements. The predominant method to assess the effectiveness of training programs is to have the participants answer Multiple Choice Question (MCQ) and True/False (T/F) questions after the training; however, the metrics typically used to report the outcome of such assessments have drawbacks that make it difficult for the trainer and organization to easily identify the concepts that need more focus and those that do not. This study introduces measures of the Assessment of Training Effectiveness Adjusted for Learning (ATEAL) method, which compensate the assessment scores for prior knowledge and negative training impact in quantifying the effectiveness of each concept taught. The results of this method are compared to the results of the most popular methods currently used. A simulation of various scenarios and the training effectiveness metrics that result from them is used to illustrate the sensitivity and limitation of each method. Results show that the proposed coefficients are more sensitive in detecting prior knowledge and negative training impact. Additionally, the proposed ATEAL method provides a quick and easy way to assess the effectiveness of the training concept based on the assessment results and provides a directional guide on the changes that need to be made to improve the training program for the participants. A companion paper expands the concepts using results from actual training sessions in multiple industries.

Keywords: multiple choice question, learning assessment, prior knowledge, training effectiveness

4.1 Introduction

Employee training in work environments is a popular way to increase competency and/or change expected behavior (Tai, 2006). Globally, organizations spent $359 billion on training in 2016 (Glaveski, 2019) with the US spending a total of $87.6 billion in 2018 (Freifeld, 2018). With this substantial amount of resources being invested, it is critical that organizations are able to ensure that the training is effective and is leading to the expected changes. Measuring training effectiveness using training evaluations or assessments is a the most widely used method to understand and quantify the deficiencies in the training programs and in developing prescriptions for improving (Alvarez, Salas & Garofano, 2004; Simkins & Allen, 2000). Of the various models presented by Campbell-Kyureghyan, Ahmed & Beschorner (2013) and in the meta-analysis conducted by Alvarez et. al., (2004), the Kirpartick's model (Kirkpartick, 1967), remains one of the most frequently used in training environments to measure training effectiveness (Arthur Jr., Bennett Jr., Edens & Bell, 2003; Salas & Cannon-Bowers, 2001).

The Kirkpatrick's model is comprised of four evaluation levels that measure participants Reaction (Level 1), Learning (Level 2), Behavior (Level 3) and Results (Level 4). The evaluation of Learning (Level 2) by participants, is typically measured by scores attained in post-test assessments or by score changes between pre- and post-test assessments (Dimitrov & Rumrill Jr., 2003). The tests that are administered are typically Multiple Choice Question (MCQ) tests as they are the most expeditious to administer (Bar-Hillel & Budescu, 2005). As we assesses the scores, it is important to clearly be able to measure if the training of the concepts has been effective, if the participants needed to have the training at all and/or if the participants regressed in their knowledge of any of the concepts due to the training.

A pre-test / post-test assessment model is effective at measuring the change in the score of the participants between the pre-test and post-test assessments (Samuel, Azen & Campbell-Kyureghyan, 2019). A variety of different statistics, such as score deltas, ANOVA, ANCOVA, have been employed to measure the effectiveness of the training and the extensive reviews of their benefits and drawbacks are detailed by Dimitrov and Rumrill Jr. (2003), Bonate (2000) and Tannebaum and Yukl (1992). A novel method to break down the pre- / post-test assessments results into quadrants of study was conducted on Economics students by Walstad and Wagner (2016). These measures give an overall understanding into the effectiveness of the training as a whole and the performance of the participants in each question or concept trained. Walstad and Wagner (2016) defined the four quadrants of learning as positive, negative, retained and zero and argued that solely using post-test scores, or the difference in pre- and post-test scores may produce misleading results as each of the scores is influenced by these four learning concepts and their interactions that cannot be discerned easily.

Despite all the information that can be determined from the available assessment methods, there does not exist an easy method to help trainers quickly and effectively understand the learning gaps by concept and give directional guidance on the countermeasures to be taken to improve the learning effectiveness of the participants for each concept trained. Hence, there exists a need for a new methodology to help assess the training effectiveness of concepts:

- Quickly, accurately & repeatably
- Easily interpreted, understood and acted upon to improve outcomes
- Visually impactful to communicate easily to industry stakeholders
- Usable in Multiple Choice Question (MCQ) and True/False (T/F) instances when an I Don't Know (IDK) option is present

This paper introduces the Assessment of Training Effectiveness Adjusted for Learning (ATEAL) methodology that satisfies the gaps stated above and validates the methodology using scenarios and simulation results.

4.2 Method

4.2.1 Learning Assessment Notation

The evaluation of training effectiveness in a pre- and post-test assessment begins with the understanding of the various possible outcomes of the answers, as shown in Figure 4-1. Additionally, Figure 4-1 summarizes the terminology that will be used in this paper. Each combination of pre- and post-test answers is described with two letters, the first being the pre-test result and the second the post-test result. "C" indicates a correct answer, and "I" indicates an incorrect answer or the selection of I Don't Know (IDK). Thus, for example, "CC" indicates a correct answer on both tests, while "IC" represents an incorrect answer on the pre-test and a correct answer on the post-test.

		Post-Test: Correct	Post-Test: Incorrect + IDK
Pre-Test	Correct	CC	CI
	Incorrect + IDK	IC	II

Figure 4-1: Terminology describing pattern of responses in a pre-/ post-test assessment model

In Figure 4-1, each quadrant contains a frequency (or percentage) of respondents and can be interpreted as follows:

- CC: The question is answered correctly in both the pre-test and post-test, indicating that the participants had pre-knowledge of the question or concept
- CI: The question is answered correctly in the pre-test and incorrectly or as IDK in the post-test, indicating that the participants experienced negative learning of the question or concept

- IC: The question is answered incorrectly or as IDK in the pre-test and correctly in the post-test, indicating that the participants learned the concept
- II: The question is answered incorrectly or as IDK in both the pre- and post-test, indicating that the participants did not learn the question or concept

4.2.2 Traditional Assessment Metrics

Training metrics are used to measure the effectiveness of the training and to help determine if there has been an increase in the level of knowledge for the learning objectives among the participants. There are several traditional metrics used to assess pre- / post-training effectiveness.

The most common method to assess testing results for a certain question or concept is to report the number of participants who answered a certain question correctly compared to the total number of participants who answered the question. It can be used both in a pre- /post-training assessment model or in a post-training only assessment model. The formula (4.1) below illustrates the calculation in the case of a pre- /post-training assessment model with an IDK option, and computes the number of correct post-test responses as a proportion of the total responses:

$$\text{Total Percent Correct (TPC)} = \frac{CC+IC}{CC+IC+CI+II} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

The key benefits of TPC are that it can be easily calculated, explained, and understood by the training participants and other organizational stakeholders. However, it gives broad stroke representations of the learning of the participants and thus the performance of the trainee. It is very difficult to discern participant pre-knowledge from actual learning and to use this metric to make improvements to the training programs. Additionally, this metric does not provide an understanding of the negative learning that any of the participants may have experienced, where

negative learning is defined as answering the pre-training question correctly and answering incorrectly on the post-training assessment (CI).

Another method to assess learning is to examine the difference between the number of participants who answered the question correctly in the post-test and the pre-test, which can only be used when the same questions are administered before and after the training. The formula (4.2) below illustrates the calculation in the case of a pre-/post-training assessment model with or without an IDK option. As seen in Figure 4-1, both the IDK and an incorrect answer are treated identically.

$$\text{Post – Pre-Training Percent Correct (PPPC)} = \frac{CC+IC}{CC+IC+CI+II} - \frac{CC+CI}{CC+IC+CI+II} = \frac{IC-CI}{CC+IC+CI+II} \quad \ldots(4.2)$$

Similar to the TPC metric, this measure is easy to calculate, explain and understand. It can also be used to determine the number of participants who answered a certain question correctly. Additionally, it compensates for participants who might have experienced negative learning. However, it is difficult to easily discern what percentage of the participants actually learned the new concept as this measure is insensitive to the prior knowledge of the participants. This also means that it does not allow for determination of the total knowledge of the participants.

4.2.3 Assessment of Training Effectiveness Adjusted for Learning (ATEAL)

The main contributions of this paper are the introduction and validation of the ATEAL method, which starts with the introduction of the Learning Adjustment Coefficient (LAC) and the Net Training Impact Coefficient (NTIC). A number of intermediate metrics and parameters, which will be subsequently used in the calculation of these two coefficients, are defined first.

4.2.3.1 Prior Knowledge (PK):

This metric represents the proportion of all participants who answered a question correctly in the post-training assessment who also answered correctly in the pre-training assessment; it is calculated using the formula (4.3) below.

$$\text{Prior Knowledge (PK)} = \frac{CC}{CC+IC} \quad \text{.....(4.3)}$$

This metric ranges from 0 to 1, where a 0 implies that none of the participants who answered the question correctly in the post-training assessment had any prior knowledge of the concept and 1 implies that all of the participants who answered the question correctly in the post-training assessment had prior knowledge of the concept. That is, a higher PK indicates greater prior knowledge among the participants. This metric is specifically different from CC as a fraction of all the participants answering the question since it helps better estimate the proportion of correctly answering participants with prior knowledge.

4.2.3.2 Positive Training Impact (PTI):

This metric represents the proportion of all the participants who needed to learn the concept (responded incorrectly or IDK in the pre-test assessment) who actually did learn the concept as indicated by their response changing to correct in the post-test. It is described below in (4.4).

$$\text{Positive Training Impact (PTI)} = \frac{IC}{IC+II} \quad \text{.....(4.4)}$$

This metric ranges from 0 to 1, where a 0 implies that none of the participants who could potentially learn actually learned the concept, and a 1 implies that all of the participants who could potentially learn actually learned the concept. That is, a higher PTI indicates more learning among the participants who did not know the concept prior to training. This metric is specifically different from IC as a fraction of all the participants answering the question since it helps better estimate the proportion of participants who did not know the concept prior to training who learned the concept.

4.2.3.3 Negative Training Impact (NTI):

This metric represents the proportion of participants who presumably knew the concept prior to training (answered correctly in the pre-training assessment) who answered incorrectly or IDK in the post-test assessment, potentially due to confusion during the training or guessing. It is described below in (4.5).

$$\text{Negative Training Impact (NTI)} = \frac{CI}{CC+CI} \quad \text{.....(4.5)}$$

This metric ranges from 0 to 1, where 0 implies that none of the participants were negatively impacted by the training and 1 implies that all of the participants (who knew the material prior to training) were negatively impacted by the training. That is, a higher NTI indicates that more participants "unlearned" the material after training. This metric is specifically different from CI as a fraction of all the participants answering the question since it helps better estimate the proportion of participants who had a negative impact from the training.

4.2.3.4 Learning Adjustment Coefficient (LAC):

The LAC is intended to measure the necessity of the training. That is, it compares the positive impacts of the training, determined through the PTI, to the prior knowledge (PK) of the participants. This difference between (actual) learning and prior knowledge is calculated (4.6) as:

$$\text{PTI} - \text{PK} = \frac{IC}{IC+II} - \frac{CC}{CC+IC} \quad \text{.....(4.6)}$$

This metric ranges from -1 to +1. To make the scale more intuitive, it is transformed to represent a proportional change by the following transformation, resulting in the Learning Adjustment Coefficient as shown in (4.7):

$$\text{LAC} = \frac{1+(\frac{IC}{IC+II} - \frac{CC}{CC+IC})}{2} \quad \text{.....(4.7)}$$

The LAC coefficient ranges from 0 to 1, where a 0 implies that all the respondents had prior knowledge so there was no actual learning for that specific concept / question, and 1 implies that there was no prior knowledge and all the respondents who needed to learn the concept did learn the concept. Higher values of LAC thus indicate that the training was needed, and effective, for a higher proportion of the respondents. Lower values indicate that either the training was ineffective, or a substantial number of respondents had previous knowledge and did not require training on the concept.

4.2.3.5 Net Training Impact Coefficient (NTIC):

The NTIC is intended to measure the negative impact of the training session. That is, it compares the positive impacts of the training, determined through PTI, to the negative impact of training (NTI) of the respondents. The difference in the learning and negative impact is calculated in (8) as:

$$\text{NTIC} = \text{PTI} - \text{NTI} = \frac{IC}{IC+II} - \frac{CI}{CC+CI} \quad \text{.....(4.8)}$$

This metric ranges from -1 to +1, where a -1 implies that all the respondents experienced negative training and lost knowledge for that specific concept / question, and a 1 implies that there was no negative training impact and all the respondents who needed to learn the concept did learn the concept. Values of NTIC higher than 0 indicate that there were more positive than negative effects from the training. Values lower than zero indicate greater negative effects, and a value of 0 means the positive and negative effects were equal.

4.2.3.6 Training Effectiveness Matrix (TEM):

To summarize these measures and allow for visual identification of the training effectiveness for a concept/question (as well as determine appropriate adjustment if the training was ineffective), the LAC and the NTIC are plotted together as illustrated in Figure 4-2. The results

regarding effectiveness can be determined based on the quadrant an item is in, where the quadrants for which NTIC is below 0 are combined.

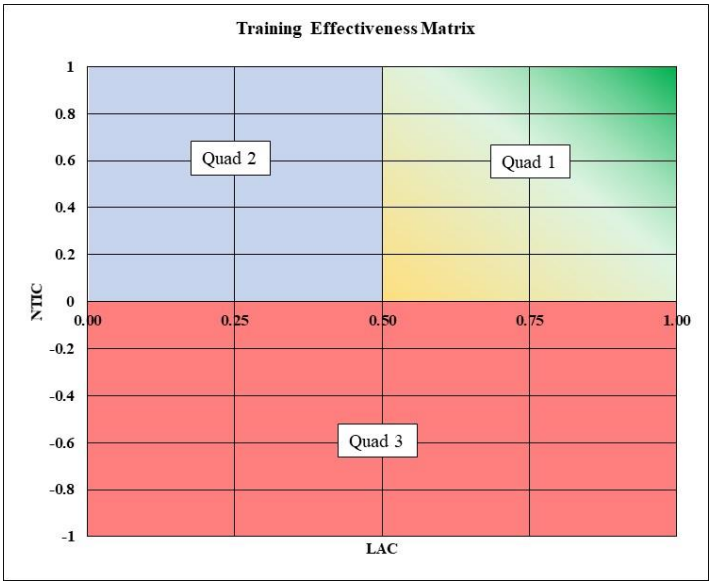

Figure 4-2: Training Effectiveness Matrix with the quadrant layout

Quad 1 contains the questions/concepts for which the percentage of participants had more positive training impact than either prior knowledge or negative learning impact. That is, the percentage of participants who learned the concept from the training is larger than the percentage who had knowledge before training and is also larger than the percentage who experienced negative learning. In the perfect case scenario, if all participants had only positive learning impact and no prior knowledge or negative learning impact, then the question would score as (1,1) on the axes in Figure 4-2. This effectiveness decreases in magnitude as a question scores closer to (0.5,0). This is illustrated with the change in the gradient of color from dark green to yellow. Quad 2 contains the questions/concepts for which the participants had more prior knowledge than positive training impact; however, the question did not experience more negative training than positive training. While for items in this quadrant the training was effective, it indicates that there was significant prior knowledge so training time could potentially be better utilized on other topics. Quad 3 contains the questions for which the participants had

substantial negative training impact and it outweighs any positive impact. It is undesirable for questions to land in this quadrant as it implies that the participants had a reduction in the level of knowledge for the concept based on the training or were forced to guess. In both cases, it indicates a deficiency in the training content, assessment question, or method of training.

4.2.4 Methods to evaluate measures

Two approaches will be used to compare the traditional and proposed metrics. First, meaningful hypothetical scenarios will be used to illustrate the meaning of each metric and their relationship. The use of these scenarios allows for clear expectations and intuitive insight into the meaning of the metrics. Second, a simulation was performed to allow for investigating a larger number of possible outcomes and scenarios, across the range of possibilities. The results of the traditional and proposed metrics were compared to determine their relationship and aid in interpretation of all metrics.

4.2.4.1 Hypothetical Scenarios:

The scenarios, detailed in Table 4-1, were developed to represent the responses (using the categories from Figure 4-1) of a hypothetical group of 100 training participants. These scenarios were chosen as they represent the extremes of learning outcomes in a Pre- / Post-Test assessment model as well as a middle ground of participant performance during a training assessment. The scenarios shown in Table 4-1 included various combinations of complete (C), high (H), moderate (M), and zero (Z) levels of Baseline knowledge, Positive learning, and Negative learning.

Table 4-1: Scenario model data sets, where C=complete, H=high, M=moderate, L=low, Z=zero.

Scenario	Baseline	Positive	Negative	CC	CI	IC	II
1	C	Z	Z	100	0	0	0
2	Z	C	Z	0	0	100	0
3	Z	Z	C	0	100	0	0
4	Z	Z	Z	0	0	0	100
5	M	H	Z	30	0	70	0

6	H	M	Z	70	0	30	0
7	L	H	L	20	10	60	10
8	H	L	L	60	10	20	10
9	L	L	H	10	60	5	25
10	L	L	M	10	25	5	60
11	L	L	H	5	60	10	25
12	L	L	M	5	25	10	60

The LAC and NTIC were calculated for each one of these scenarios and plotted on the matrix in Figure 4-3 (see Results section) to illustrate their quadrant placement and how they can be interpreted. Additionally, the TPC and PPPC are also calculated for each of the scenarios so a comparison can made in terms of how each metric reports the effectiveness of the training (see Table 4-3 in the Results section).

4.2.4.2 Data Simulation:

To further expand on the scenarios modelled and examine a larger population of questions and students, a random number generator (in MS Excel) was used to generate 100 participant responses on 1000 questions for both pre- and post-training. The MS Excel random number generator generates numbers from a uniform distribution, ranging from 0 to 1, and the generation technique produced data for CC, CI, IC & II. The uniform distribution was considered a good way to generate the data as it does not make any preconceived assumptions on how participants would respond in an assessment and if they would learn or not learn a concept. That is, it allows for equal probabilities of the possible outcomes. The data points generated ranged from 0 participants to all the participants included in any of the quadrants and the sum of the number of answers in each of the pre/post condition totals 100 participants answering each question. Table 4-2 is an excerpt from the of the values of CC, IC, CI and II for the simulation and illustrates the result of the training effectiveness metrics for each question.

Table 4-2: Excerpt of the values for the simulation model and the calculated training effectiveness metrics

	CC	IC	CI	II	Total	TPC	PPPC	LAC	NTIC
Question 1	37	8	45	10	100	45%	-37%	0.31	-0.10
Question 2	39	1	19	41	100	40%	-18%	0.02	-0.30
Question 3	12	22	18	48	100	34%	4%	0.48	-0.29
Question 4	4	60	9	27	100	64%	51%	0.81	0.00
Question 5	7	5	21	67	100	12%	-16%	0.24	-0.68
Question 6	52	10	4	34	100	62%	6%	0.19	0.16

4.3 Results

Results of the simulations are presented with an emphasis on comparing the traditional and newly proposed assessment metrics, and the relationship between the two new metrics.

4.3.1 Scenario Results

Table 4-3 illustrates the metrics calculated for each of the twelve scenarios detailed in Table 4-1. In scenario 1, where all the participants have pre-knowledge of the concept taught, the TPC reports the score as 100% implying that all the participants learned the concept, which is an incorrect interpretation of the training effectiveness. The PPPC reports the score as 0% implying that none of the participants learned the concept. Although this is a correct interpretation of training effectiveness, it is not distinguishable from scenario 4 and it would not be possible to distinguish concepts in which the participants had all pre-knowledge or zero learning. Looking at the two new coefficients for scenario 1, the LAC is 0 implying that 100% of the participants had prior knowledge and none learned the topic during training, and an NTIC of 0 implying that there is equal amount of positive training impact and negative training impact. The two introduced coefficients must be examined together to clearly understand the performance of the participants for each scenario.

Table 4-3: Metrics calculated for each scenario

Scenario	Baseline	Positive	Negative	TPC	PPPC	LAC	NTIC
1	C	Z	Z	100%	0%	0	0
2	Z	C	Z	100%	100%	1	1
3	Z	Z	C	0%	-100%	0.5	-1
4	Z	Z	Z	0%	0%	0.5	0
5	M	H	Z	100%	70%	0.85	1
6	H	M	Z	100%	30%	0.65	1
7	L	H	L	80%	50%	0.80	0.52
8	H	L	L	80%	10%	0.46	0.52
9	L	L	H	15%	-55%	0.25	-0.69
10	L	L	M	15%	-20%	0.21	-0.64
11	L	L	H	15%	-50%	0.48	-0.64
12	L	L	M	15%	-15%	0.40	-0.69

To visualize the implication of each scenario, the TEM is provided in Figure 4-3 and includes each scenario labeled by its number. From the matrix we can easily see that scenarios 2, 5, 6, 7, and 8 show a positive training impact on the participants, to varying degrees, and it is easy to visualize the magnitude of the impact based on how the points lie in the upper right quadrant. We can also see that scenario 8, along with scenario 1, consists of more prior knowledge than learning, representing cases in which the training was perhaps unnecessary. Scenario 4 shows zero learning impact, and participants had equal learning and prior knowledge. Finally, scenarios 3, 9, 10, 11 & 12 show more negative training impact that positive impact. Similar interpretations for most, but not all, scenarios can be made by looking at the PPPC. However, it is not possible to make that same determination using the TPC. Thus, the LAC and NTIC provide a finer resolution on the PPPC and TPC. This additional information will help trainers and organizations better understand whether the concept needs to be taught and ensure that the participants experience more positive than negative learning due to the content presented or method by which it was delivered.

Figure 4-3: Training Effectiveness Matrix for the 12 scenarios

4.3.2 Simulation Results

Figure 4-4 illustrates the LAC and NTIC values of the simulated data, calculated for each of the 1000 simulated data points (i.e., test questions or concepts), plotted on the TEM. The data points are observed to range from (0,-1) to (1,1) as we would expect in participant answers. Larger values of LAC result from either high PTI or low PK. In either case, with a large LAC the NTI cannot be small, so the lower right corner of the TEM does not contain any data points. Similarly, for low LAC there can be little positive learning, so the upper right corner of the TEM does not contain any data points.

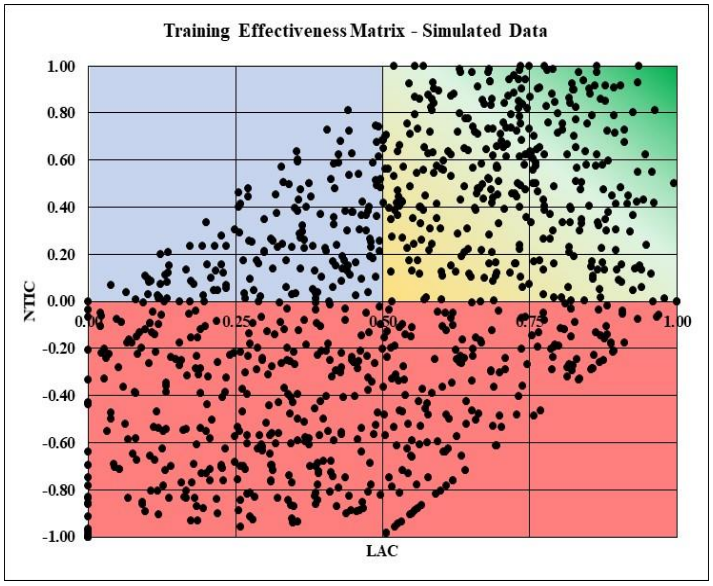

Figure 4-4: Training Effectiveness Matrix for the 1000 simulated data points.

The simulated data was used to provide a large number of different cases and allows for examining the sensitivity of the TPC, PPPC, LAC and the NTIC to changes in percentage of prior knowledge and negative training impact. Starting with prior knowledge (PK), Figure 4-5 presents the values for the simulated cases of (a) TPC, (b) PPPC, (c) LAC, and (d) NTIC on the y-axis, and PK on the x-axis. TPC is observed to be insensitive to the changes in prior knowledge with a slope of -0.072. The striations of data points observed at the bottom left and right of the scatter plot are related to the results for very low values of CC. PPPC has a negative correlation of -0.55 indicating that as the percentage of prior knowledge increases from 0% to 100%, the PPPC decreases, although the total knowledge is not decreasing. The plot also illustrates that data does not occur above a line extending from (0,1) to (1,0) as both PK and PPPC are related to changes in IC. As IC approaches 100%, PK approaches 0% and PPPC can assume any value. Conversely, as IC nears 0%, PK approaches 100% and PPPC is limited, with a maximum of 0.0 when PK equals 1.0.

We observe that LAC has a very strong negative correlation of -0.82 with PK, indicating that it is very sensitive, much more so than PPPC, to changes in prior knowledge. The plot for LAC

also exhibits less scatter than the plots for the other measures, demonstrating a stronger linear relationship with PK. The empty quadrants are due to PK being one component of LAC. As PK increases the maximum value of LAC is limited, and vice versa for low values of PK. Finally, as expected, the NTIC appears insensitive to prior knowledge.

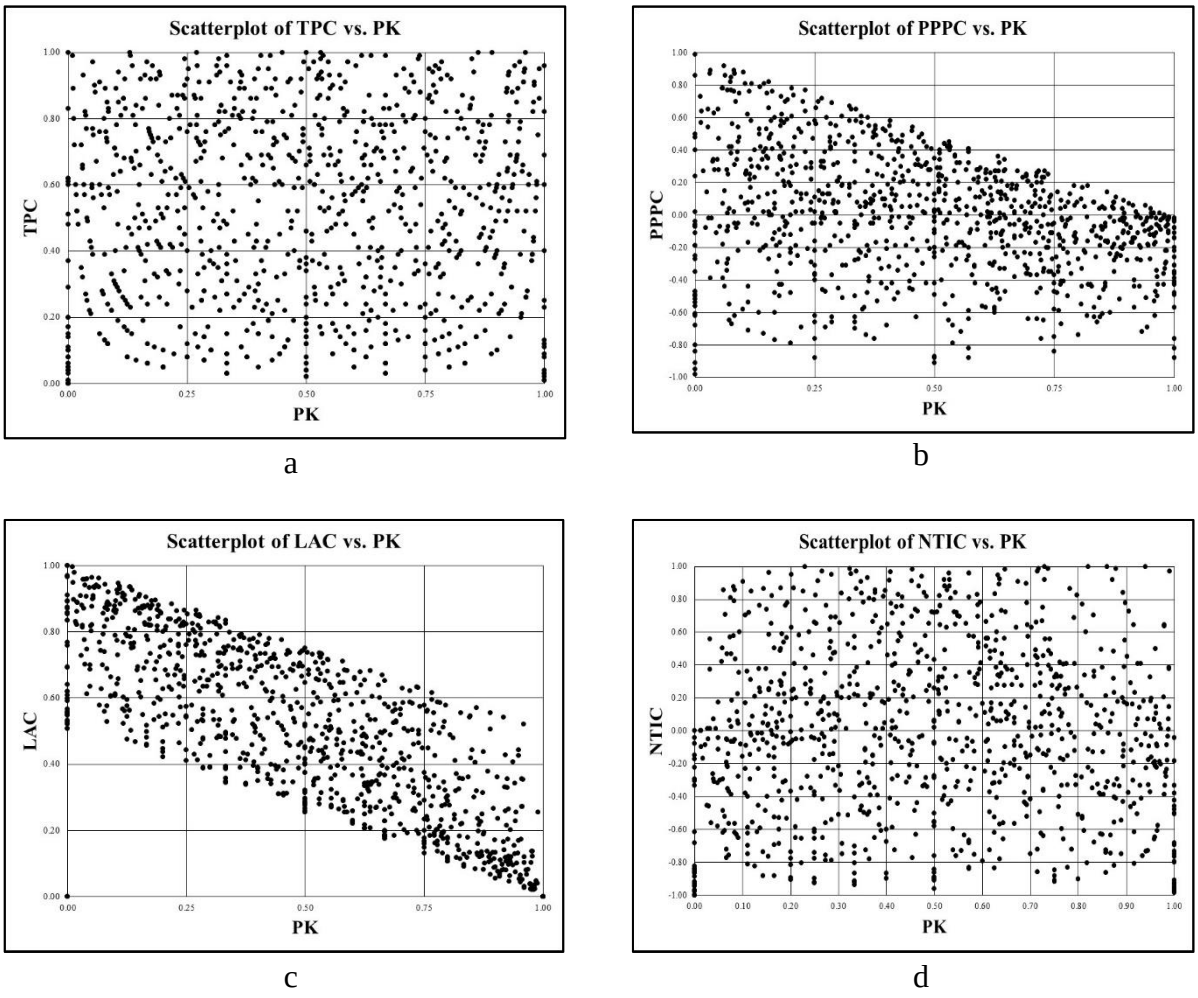

Figure 4-5: Sensitivity analysis of the simulation values of (a) TPC, (b) PPPC, (c) LAC, and (d) NTIC (y-axis) with increasing prior knowledge (PK, x-axis)

Next, the changes in the four metrics are investigated as the negative training impact varies from 0% to 100%. Figure 4-6 illustrates the changes in (a) TPC, (b) PPPC, (c) LAC and (d) NTIC with respect to NTI. TPC and PPPC are observed to have a negative correlation of -0.77 and -0.62, respectively, indicating that as the percentage of negative training impact increases,

both the TPC and PPPC decrease. PPPC has a lower limit for a given value of NTI since IC always ranges from 0 to 100 while CI is directly related to NTI. LAC, as expected, is observed to be insensitive to NTI. Finally, we observe that NTIC has a strong negative correlation of -0.82 with NTI. This is expected as the NTIC is directly dependent on the negative training impact and is the most sensitive of all the metrics to NTI. As NTI approaches zero we observe that NTIC ranges from 0-1 as participants can only experience PTI when there is no NTI.

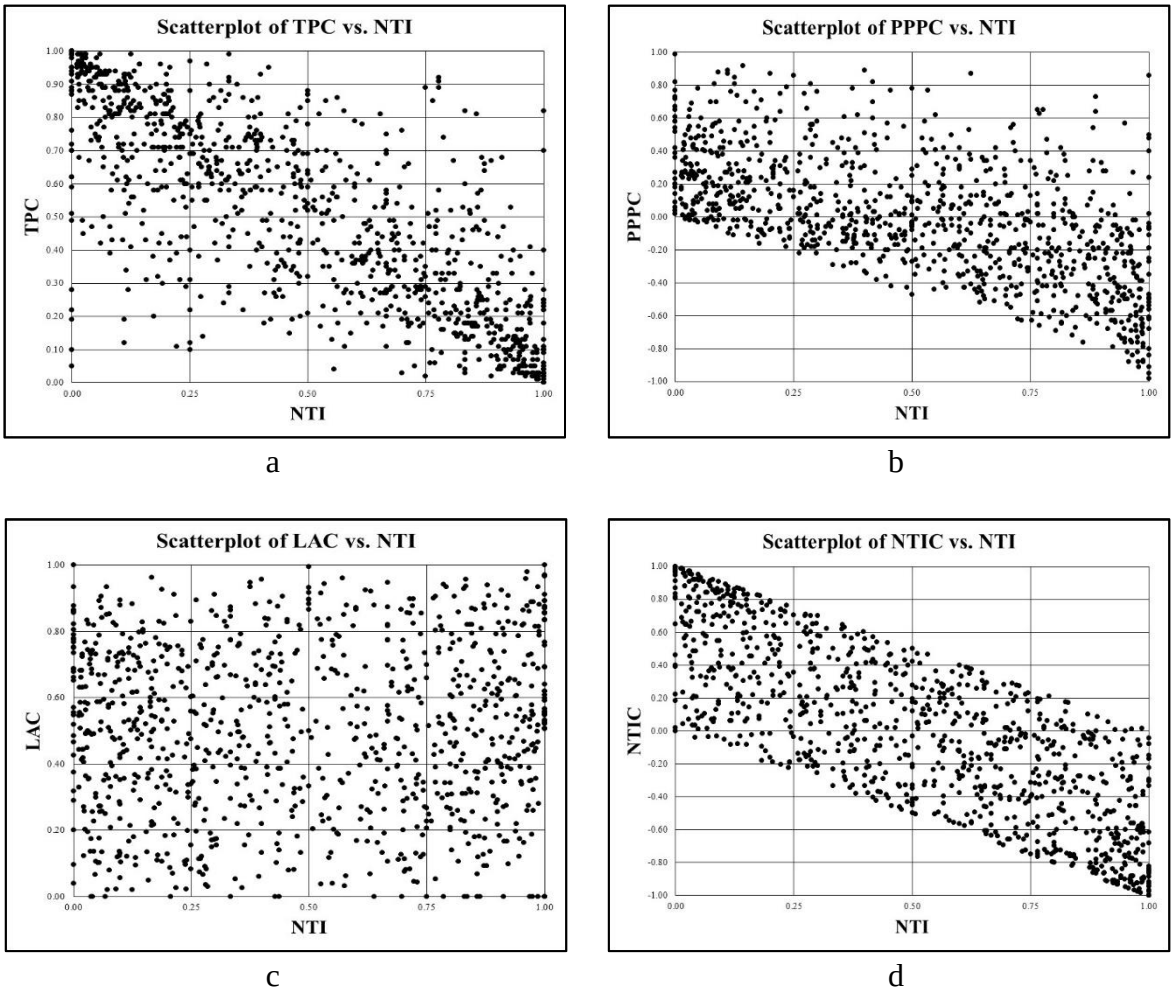

Figure 4-6: Sensitivity analysis of the values of (a) TPC, (b) PPPC, (c) LAC, and (d) NTIC (y-axis) with increasing negative training impact (NTI, x-axis)

4.4 Discussion

The analysis conducted illustrates some interesting behaviors of the metrics used to report training effectiveness. In this section we will discuss the merits and drawbacks of each of the metrics and their behavior in the various scenarios and simulation results.

4.4.1 Scenario Comparisons:

In applying the ATEAL method and plotting the scenarios on the Training Effectiveness Matrix (TEM), we see that Quad 1 (Figure 4-3) ranges from (0.5,0), which illustrates zero learning as in Scenario 4, to (1,1), which illustrates perfect learning as in Scenario 2. Scenarios 5, 6, and 7 lie in Quad 1 as all of these scenarios illustrate a higher percentage of participants experienced positive learning than those that had prior knowledge and/or experienced negative training. When there is a higher percentage of participants having prior knowledge than experiencing positive training (Scenarios 1 & 8) we observe that they lie in Quad 2. Thus, it is easy and quick to determine the concepts for which there are a larger percentage of participants that had a higher level of prior knowledge. In these cases it would be advisable for the trainer to spend minimal time reviewing the concept and not test on it as it is redundant; that is, valuable training time could be better spent on concepts unknown to the participants. Scenarios 3, 9, 10, 11, and 12 lie in Quad 3 and these scenarios represent cases when there is a larger percentage of participants experiencing more negative learning than positive learning. These are the worst-case scenarios and represent cases where the participants were either guessing or were confused by the training content and/or the delivery method. It is important for the training provider and the organization to determine the number of participants that experienced higher NTI, pay attention to these concepts, and closely analyze and develop corrective actions to prevent this

from future occurrences. Additionally, these results can also be used to determine the amount of supervisor support and reinforcement needed to help support the use of skills (Russ-Eft, 2002).

In analyzing the same scenarios with TPC, we see that this metric is overly optimistic in its interpretation of the participants' performance. As shown in Table 4-4, for Scenarios 1, 5, and 6, TPC reports the performance of participants as 100%. This would imply that all participants learned these concepts; however, in these scenarios, all participants had prior knowledge. In using this metric, we would interpret the training as extremely effective although the participants would feel that the training of the concept was a waste of time because they already knew it. The correct course of action for a concept that behaves like Scenario 1 is to either not train on the concept or do a cursory training without testing on the concept and focus instead on concepts for which the participants have less prior knowledge. In Scenario 3, all the participants exhibited negative learning but the TPC reports the performance as 0%, implying that there was no learning among the participants. In this scenario we know that the participants were, in effect, guessing or losing knowledge due to the training process, which would indicate that there were significant issues with the content or the method of delivery. It is not possible to distinguish between this outcome (Scenario 3) and Scenario 4 in which had all the participants answered incorrectly in both the pre- and post-test assessments. Additionally, when using the TPC metric to measure training effectiveness, it is not possible to distinguish between Scenarios 9, 10, 11, and 12, which all had differing amounts of negative learning and participants answering incorrectly in both the pre- and post-test assessment. This severely limits the understanding of participant performance and the determination of needed training improvements.

When examining the scenario results using PPPC, we observe that this metric performs better than the TPC metric in representing the learning of the participants. In Scenario 1, it reports that

there was no learning by the participants since they had 100% prior knowledge; however, unlike the ATEAL method, it is not possible to easily discern if the low score is due to prior knowledge or a lack of learning or guessing. In Scenario 2 PPPC indicates that the participants experienced 100% learning, same as the ATEAL method. This is distinctly different from the results illustrated by the TPC (100% in both scenarios 1 and 2) and helps the trainers better understand the impact of the training. In Scenario 3, PPPC reports a result of -100% since all the participants experienced negative learning, same as the ATEAL method that plots Scenario 3 at the lowest score in Quad 3. In Scenarios 5, 6 and 7the PPPC reports positive learning based on changes in the number of participants who have prior knowledge and those experiencing positive learning. . When there is more negative learning than positive learning or prior knowledge (Scenarios 9, 10, 11 & 12) PPPC reports a negative value, thereby indicating that there is a significant issue with the training and that the participants are being affected in a negative manner. These negative results are similar to the ATEAL method that plots these scenarios in Quad 3. In Scenario 8, the PPPC reports that the participants experienced positive learning, however, using the ATEAL method, we are very quickly able to diagnose that Scenario 8 had more prior knowledge than positive learning. This is not readily apparent when looking at the PPPC results, and it requires the trainers/assessors to review the raw data to arrive at the conclusion that the ATEAL method readily provides. Additionally in Scenarios 1 and 4, PPPC reports that no participants learned the concept trained, however, when using the ATEAL method, we observed that in Scenario 1 all the participants had prior knowledge of the concept taught and did not need to learn the concept and in Scenario 4, none of the participants exhibited any learning.

4.4.2 Simulation Results:

In interpreting the results of the simulation using the ATEAL methodology, we observe that the LAC is the most sensitive (slope of -0.82) of all the metrics to prior knowledge of the participants. This implies that as the prior knowledge among the participants increases, for a certain question or concept taught, the value of the LAC decreases. Similarly, the NTIC is the most sensitive (slope of -0.82) of all the metrics to negative training impact. As in the case of the LAC, this implies that as the participants experience more negative training for a certain question or concept, the value of NTIC decreases, and when 100% of the participants experience negative training, all associated NTIC values are negative. Thus, the use of these two coefficients to develop the TEM, enables the matrix to be more sensitive for the effects of prior knowledge and negative training when reporting the training effectiveness for the concepts taught.

It is also important to note that the NTIC can be sensitive to the number of trainees with prior knowledge. If a small number of trainees have prior knowledge the NTI can be large, even if only one or two trainees experienced negative learning. Conversely, if most trainees have prior knowledge the PTI is greatly impacted by even a small number of trainees who learn the concept. Thus, either very high or very low values of NTIC must be further examined to determine the cause, since either extreme case may indicate problems with the training related to prior knowledge rather than the training quality.

The TPC metric is completely insensitive to participant prior knowledge and treats it as learning, which is troublesome as it does not give feedback to the trainers or the organization that would help improve the training and better focus on the needs to the participants' knowledge gaps. It paints an overly optimistic picture of the training when, in effect, the participants' and

organizations' time might be wasted by the training. Additionally, the participants could be getting bored during the training, causing them to lose focus and pay less attention to the concepts that they actually do not know and need to learn. The TPC does illustrate a negative trend when the participants experience negative training. This is due to the fact that participants experiencing negative learning answer incorrectly in the post-test assessment, thus reducing the TPC score. The score, however, does not clearly show that this is due to negative learning and it can be interpreted to mean that the participants did not learn the content being trained, which is a completely different scenario.

Finally, the PPPC metric is sensitive to prior knowledge as it decreases with an increase in prior knowledge as noted by several authors (e.g., Bonate, 2000; Dimitrov & Rumrill Jr., 2003; Tannebaum & Yukl, 1992). The PPPC also has similar sensitivity towards Negative Training Impact, in that it decreases with an increase in negative training impact. However, unlike the NTIC, when the negative training impact is close to a 100%, a small percentage of the data points are greater than zero. This makes interpretation of the PPPC metric slightly more challenging than the NTIC in which all the values are negative when 100% of the participants experience negative training. Additionally, it is difficult to discern participant performance when there is a low positive score; that is, we are not able to easily determine whether the low score was due to high prior knowledge or due to negative learning. Hence, it makes it difficult to quickly determine the countermeasures that are needed to improve the effectiveness of the training.

The comparisons of the scenario and simulation results using these metrics and associated discussions in this section allow us to observe the following benefits of the newly introduced ATEAL:

- It is much more effective in helping determine the true performance of the participants in a training session for each concept taught.
- The metrics involved are easy to calculate and provide visual guidelines for the training providers and the organizations on the best and worst learned concepts.
- It is much more specific than the other two metrics and helps to quickly diagnose issues with participant performance by identifying whether the training should be improved (by making the content taught more challenging, to get around prior knowledge) or if the training is causing confusion among the participants and thus reducing their learning.

4.5 Conclusion / Future Direction

Metrics to quantify the amount of learning that training participants exhibit for a particular training course, or concepts within the course, are critical to understanding and quantifying the effectiveness of the training. The Assessment of Training Effectiveness Adjusted for Learning (ATEAL) method is introduced in this paper and defines new metrics to measure the level of prior knowledge, as well as positive and negative training impacts experienced by the participants. Additionally, it introduces two coefficients, Learning Adjustment Coefficient (LAC) and Net Training Impact Coefficient (NTIC), that are plotted in a novel method to create the Training Effectiveness Matrix (TEM). This matrix helps visually assess the performance of the participants for each question/concept introduced in the training. The method proves effective in quickly identifying the training gaps that the participants experienced and providing direction on the countermeasures that should be taken for each concept trained.

Validation of this new method and comparison of its performance to the traditional metrics of TPC and PPPC was conducted using scenario modelling and a simulation. Some recommendations that can be derived from this study are:

- Using only the TPC in the post-test assessment to assess training effectiveness (i.e., how much the participants learned) may give a highly inaccurate impression and does not provide clear guidance on areas of improvement.
- The PPPC is a much better metric than the TPC to assess training effectiveness, but it lacks the ability to quickly provide guidance on changes to be made to the training content or training delivery to improve training effectiveness.
- The use of the ATEAL method in calculation of the Learning Adjustment Coefficient and the Net Training Impact Coefficient is extremely easy and interpretation using the Training Effectiveness Matrix is intuitive and visual.

4.6 References

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review, 3*(4), 385-416.

Arthur Jr, W., Bennett Jr, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society, 4*(1), 3-12.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Chapman and Hall/CRC.

Campbell-Kyureghyan, N., Ahmed, M., Beschorner, K. (2013, May). *Measuring training impact 1-5*. Paper presented at the US DOL Trainer Exchange Meeting, Washington DC, March 12-13, 2013.

Dimitrov, D. M., & Rumrill Jr, P. D. (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159-165.

Freifeld, L. (2018). *2018 Training Industry Report*. Retrieved October 10, 2019 from

Glaveski S. (2019). *Where Companies Go Wrong with Learning and Development*. Retrieved October 2, 2019 from https://hbr.org/2019/10/where-companies-go-wrong-with-learning-and-development

Kirkpatrick, D. L. (1967). Evaluation of training. *In R. L. Craig & L. R. Bittel (Eds.), Training and Development Handbook (pp. 40–60)*. New York: McGraw Hill.

Russ-Eft, D. (2002). A typology of training design and work environment factors affecting workplace learning and transfer. *Human Resource Development Review, 1*(1), 45-65.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*(1), 471-499.

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2019). Evaluation of learning outcomes through multiple choice pre-and post-training assessments. *Journal of Education and Learning, 8*(3).

Simkins, S., & Allen, S. (2000). Pretesting students to improve teaching and learning. *International Advances in Economic Research, 6*(1), 100-112.

Tai, W. T. (2006). Effects of training framing, general self-efficacy and training motivation on trainees' training effectiveness. *Personnel Review, 35*(1), 51-65.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*(1), 399-441.

Walstad, W. B., & Wagner, J. (2016). The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education, 47*(2), 121-131.

Chapter 5: Assessment of Training Effectiveness Adjusted for Learning (ATEAL). Part II: Practical Application

Thomas Samuel, Razia Azen & Naira Campbell-Kyureghyan

Abstract

Safety training programs are a popular method, in industry globally, to increase awareness of risks to employees and employers and plays a critical part in reducing safety incidents. The most frequently used method to assess the effectiveness of the training is to have the participants answer Multiple Choice Question (MCQ) and True/False (T/F) questions after the training. The metrics used to report the outcome of the assessments have drawbacks that make it difficult for the trainer and organization to easily identify the concepts that need more focus and those that do not. The goal of this research study is to compare how the methods used to measure training effectiveness of concepts in Level 2 post training assessment differ in how they assess training effectiveness using actual training results. Pre- and Post-training assessments were administered to the participants in 3 different utility industries and were analyzed for training effectiveness using the traditional metrics as well as using ATEAL method. The results were then compared and detailed recommendations of the best and least learned concepts by industry are presented based on these comparative analyses. The ATEAL method is further used to quantify the opportunities for improvement in the training programs based on the participant prior knowledge and any negative training impact observed. Results of the comparison of the various methods show that the proposed ATEAL method provides a quick, accurate and easy way to assesses the effectiveness of the training of concepts and the method identified that for 40% of the concepts trained a higher percentage of participants exhibited more prior knowledge than positive learning and for 6% of the concepts a higher percentage exhibited negative training. These results also provide a directional guide on the improvements that can be made to improve the training

effectiveness of the programs. Additionally, it also shows that the ATEAL method can be used in any learning environment where there is a pre-/post-test evaluation of the change and is not limited in application to MCQ and T/F questions.

Keywords: adult learning, training effectiveness, control question, prior knowledge, concepts trained

5.1 Introduction

Workplace training, globally, is an important way for organizations to increase the knowledge of their employees and it has been reported that organizations invest approximately $55.3 billion to $200 billion annually (Salas & Cannon-Bowers, 2001) on employee training. Brunello and Medio (2001) observed that different countries invest differing amounts in employee training based on tenure, and there is an overall approach globally to increase the knowledge of employees in an organization using formal training methods. With this level of fiscal and time investment being made in training it is important to ensure that the training is effective and will result in the expected changes in behavior among the participants.

Of the various topics that employees are trained on, safety training is particularly important due to the impact of poor safety practices (Campbell-Kyureghyan & Cooper, 2012). According to the Bureau of Labor Statistics, the number of fatal work injuries in the US for 2018 was 5,250, an increase of 2% (5,147) from 2017. Similar statistics have been reported by Ho and Dzeng (2010) on occupational disasters in Taiwan. This impact to human life and societies worldwide has necessitated a number of legislative acts and organizations being instituted to reduce occupational injuries and mandate workers to undergo safety education through training. This is a sound approach as training is a proven method to improve the safety conditions for workers worldwide with proven reduction in safety incidents on the worksite (Bahn & Barratt-Pugh,

2012; Ho & Dzeng, 2010; Burke et al., 2006; Becker & Morawetz, 2004; Demirkesen & Arditi, 2015; Campbell-Kyureghyan, Hernandez & Ahmed, 2013). The importance of training is particularly more so in dynamic work environments such as construction which was noted by Campbell-Kyureghyan, Ahmed & Beschorner (2013) as traditional approaches to implement safety protocols with workstation redesigns are ineffective or not practical.

Blume, Ford, Baldwin & Huang (2010) and Tai (2006) noted that effective training can increase the knowledge, skills and abilities (KSA's) of the employees for organizational benefit. In the case of safety training this is particularly important as there is significant human and societal impact to the employee's application of their safety KSAs in the work environment. Alvarez, Salas & Garofano (2004) stated that training experts typically study training effectiveness through evaluation and, although training evaluation and training effectiveness are distinct concepts, they are related and models that integrate both concepts provide a better overall picture. The importance of effective safety training was also stated by Demirkesen and Arditi (2015) who observed that safety improvements may not be achieved unless special attention is paid to the effectiveness of learning during the training session.

The methods used to measure training effectiveness typically involve assessing the overall performance of the participants and no easy methodology exists to help organizations and trainers determine the learning gaps and to determine the best and least learned concepts while compensating for prior knowledge and guessing. Additionally, the current methods do not provide easy directional guidance on the countermeasures that need to be taken to improve the effectiveness of the training for each concept trained. The improvements that can be made to training on concepts related to safety is specifically impactful due to the human and societal benefits that changes in safety behavior have on participants and organizations.

In the companion paper (Part 1), we describe the Assessment of Training Effectiveness Adjusted for Learning (ATEAL) methodology that is able to assess the training effectiveness of each concept taught in a training session by adjusting for negative training impacts and prior knowledge of the participants. This research study presents the results of the different training effectiveness assessment methods of concepts for a pre-/ post-test assessment model and determines how the models differ from each other on the concepts they report as best and least learned.

5.2 Method

5.2.1 Assessment Metrics

A complete description of the assessment metrics is contained in the companion paper and a brief summary is presented here. To align on nomenclature, the possible outcomes of answers in a pre- and post-test assessment are detailed below in Figure 5-1.

		Post-Test: Correct	Post-Test: Incorrect + IDK
Pre-Test	Correct	CC	CI
Pre-Test	Incorrect + IDK	IC	II

Figure 5-1: Terminology describing pattern of responses in a pre-/ post-test assessment model

Each quadrant in Figure 5-1 contains the frequency or percentage of respondents that answered in a certain manner and can be interpreted as:

CC: The question is answered correctly in both pre- and post-tests

CI: The question is answered correctly in the pre-test and incorrectly or IDK in the post-test

IC: The question is answered incorrectly or as IDK in the pre-test and correctly in the post-test

II: The question is answered incorrectly or as IDK in both pre- and post-test assessments

5.2.1.1 Total Percent Correct (TPC): the TPC measures the number of questions that the participants answered correctly in the post-training assessment or the number of participants who answered a certain question correctly and it is shown below in formula (5.1).

$$\text{Total Percent Correct (TPC)} = \frac{CC+IC}{CC+IC+CI+II} \quad \ldots\ldots(5.1)$$

5.2.1.2 Post – Pre-Training Percent Correct (PPPC): the PPPC measures the difference between the pre-/post-training scores, and can only be used when the same questions are administered before and after the training. It is computed as shown below in formula (5.2).

$$\text{Post – Pre-Training Percent Correct (PPPC)} = \frac{CC+IC}{CC+IC+CI+II} - \frac{CC+CI}{CC+IC+CI+II} = \frac{IC-CI}{CC+IC+CI+II} \quad \ldots..(5.2)$$

5.2.1.3 Prior Knowledge (PK): the PK measures the proportion of all participants who answered a question correctly in the post-training assessment who also answered correctly in the pre-training assessment, as is shown in formula (5.3).

$$\text{Prior Knowledge (PK)} = \frac{CC}{CC+IC} \quad \ldots\ldots..\ldots..(5.3)$$

5.2.1.4 Positive Training Impact (PTI): the PTI, shown in formula (5.4), measures the proportion of all the participants who needed to learn the concept (responded incorrectly or IDK in the pre-test assessment) who actually did learn the concept as indicated by their response changing to correct in the post-test.

$$\text{Positive Training Impact (PTI)} = \frac{IC}{IC+II} \quad \ldots\ldots(5.4)$$

5.2.1.5 Negative Training Impact (NTI): the NTI, shown in formula (5.5), measures the proportion of participants who presumably knew the concept prior to training (answered correctly in the pre-training assessment) who answered incorrectly or IDK in the post-test assessment.

$$\text{Negative Training Impact (NTI)} = \frac{CI}{CC+CI} \quad \ldots\ldots(5.5)$$

5.2.1.6 Learning Adjustment Coefficient (LAC): the LAC measures the necessity of the training by comparing the positive impacts of the training (PTI) to the prior knowledge (PK) of the participants, and it is calculated as shown in formula (5.6).

$$LAC = \frac{1+(\frac{IC}{IC+II}-\frac{CC}{CC+IC})}{2} \quad \text{.....(5.6)}$$

5.2.1.7 Net Training Impact Coefficient (NTIC): the NTIC measures the negative impact of the training session by comparing the positive impacts of the training (PTI) to the negative impact of training (NTI) of the respondents, and it is calculated as shown in formula (5.7).

$$NTIC = PTI - NTI = \frac{IC}{IC+II} - \frac{CI}{CC+CI} \quad \text{.....(5.7)}$$

5.2.1.8 Training Effectiveness Matrix (TEM): The LAC and the NTIC can be summarized in a Training Effectiveness Matrix (TEM) that allows for visual identification of the training effectiveness for a concept/question, as shown in Figure 5-2. The quadrants of the matrix are described below.

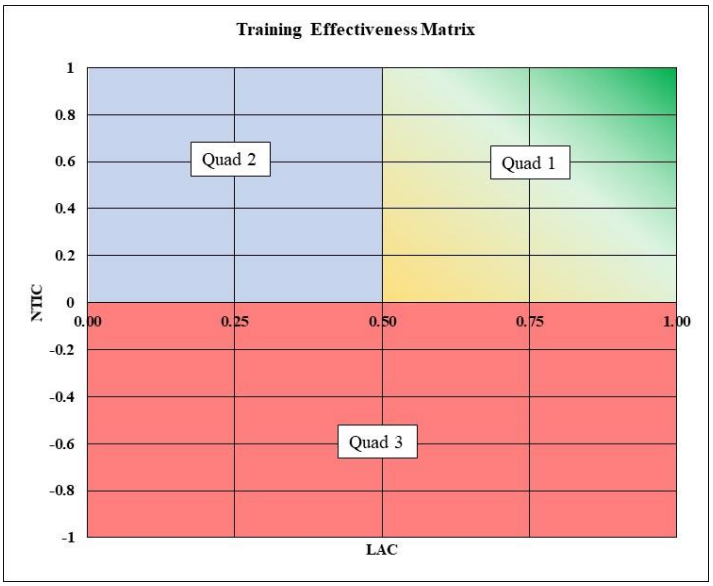

Figure 5-2: Training Effectiveness Matrix with the quadrant layout

Quad 1: Contains questions/concepts for which the participants experienced more positive training impact than either prior knowledge or negative learning impact. The color gradient

ranges from yellow to green which indicates increasing levels of positive training impact for the participants.

Quad 2: Contains questions/concepts for which the participants had more prior knowledge than positive training impact but did not experience more negative training than positive training.

Quad 3: Contains the questions/concepts for which the participants had higher negative training impact and it outweighs any positive training impact.

5.2.2 Industry Application

Workplace safety and ergonomic training was developed and deployed for multiple sectors of the utility industry by a team of researchers at the University of Wisconsin-Milwaukee under a DOL Susan Harwood Training Grant. Table 5-1 illustrates the number of participants, their roles, number of questions based on types and the usage of Control Question (CQ) and "I Don't Know" (IDK) option in the three energy sectors. The results from these training sessions will be used to evaluate the performance of the assessment metrics by comparing and contrasting how each of metrics illustrates participant performance for the concepts taught.

Table 5-1: List of the number of training participants, assessment questions, and usage of CQ and IDK option in each industry

Utility Sector	Participant Role	# of Participants	# of MCQ Assessments	# of T/F Assessments	MCQ Assessments	
					CQ	IDK
Natural Gas	Employee – Tier 1	414	7	8	X	
	Employee – Tier 2	375	7	8	X	
	Manager – Tier 1	86	7	8	X	X
Electric Transmission	Employee – Tier 1	54	9	5		X
	Manager – Tier 1	7	9	5		X
	Employee – Tier 2	359	10	5	X	X

Power Generation	Employee – Tier 2	157	10	5	X	X
	Manager – Tier 1	14	13	9	X	X
Total=1,466	Managers =	107				
	Employees =	1,359				

Further details of the training methods, the content, and knowledge testing, are detailed in a prior paper written by the same authors (Samuel, Azen & Campbell-Kyureghyan, 2019). The training sessions were all face-to-face and instructor-led with the number of training participants ranging from 6-40 per class. The pre-test and post-training assessments contained Multiple Choice Questions (MCQs) and True or False (T/F) items to determine the knowledge of the content for each participant. The pre-training assessment was completed just prior to the training session and collected on completion. The training session typically lasted from 1-3 hours and the same assessment was administered as the post-training assessment. The number of MCQ and T/F questions for each of the utility sectors, based on the role of the participants, is summarized in Table 5-1. In the MCQ assessment, one question, in both the pre- and post-training assessment, was a question contextually similar to the content being trained but was not specifically covered in the training class. This is referred to as the Control Question (CQ) and further details are provided in Samuel et al., (2019) and Caston, Cooper, and Campbell-Kyureghyan (2009). Additionally, for the pre- and post-training assessments for the Electric Transmission and Power Generation utility sectors an additional "I Don't Know" (IDK) option was added, as indicated in Table 5-1.

Training content and concepts were based on research that specifically targeted the areas of safety and ergonomics in non-repetitive work environments (Ahmed & Campbell-Kyureghyan, 2014). To define the ergonomic risks onsite visits were conducted, and data gathered from interviews with managers and employees and direct observations using

videotaping methods. Due to the differences in the types of utilities and the work performed concepts were changed to best cater to each industry and combined with information from nationwide industry and fatality statistics for utility industries (Campbell-Kyureghyan & Cooper, 2012). Table 5-2 details the concepts trained and the number of questions in the assessments by concept for the various training groups in each utility sector. Both employees and mid-level management were trained as it has been reported that management's commitment to safety results in lowering injury rates and improving the company safety culture (Demirkesen, 2015).

Table 5-2: Concepts trained and number of assessment questions for each utility industry sector

	Natural Gas		Electric Transmission		Power Generation	
	Employee - Tier 1 & 2	Manager – Tier 1	Employee - Tier 1	Employee - Tier 2	Manager – Tier 1	Employee Tier 2
Confined Space	0	0	0	0	1	1
Control Question	1	1	0	1	1	1
Electric Safety	0	0	2	2	0	0
Employee Rights & Responsibilities	1	1	1	1	0	1
Environment	2	1	1	1	1	1
General	2	2	1	1	1	2
Hearing Loss	1	1	1	1	2	1
Overexertion	3	1	3	3	5	5
PPE	0	0	2	2	0	0
Program Implementation	0	2	0	0	3	0
Root Cause Analysis	0	1	0	0	1	0
Slips, Trips & Falls	3	3	2	2	2	2
Struck by/caught between	0	0	0	0	1	1
Vehicle Safety	1	1	1	1	0	0
Vibration	1	1	0	0	0	0
Workplace Assessment	0	1	0	0	4	0

The assessment metrics were calculated for each of the training groups and are compared and contrasted to identify the metrics that best help determine the performance of the participants and the direction of training improvements required.

5.3 Results

The pre- and post-training assessment results for the participants from the various utilities are calculated using the TPC, PPPC, and ATEAL measures to help identify the concepts which were best learned, the concepts for which the participants had the most prior knowledge, and the concepts for which the participants experienced higher negative impact. Additionally, the responses of the participants on the Control Question and its representation by the various metrics is examined. Ideally, in all cases, we would expect the CQs to be at (0.5,0) in the TEM when using the ATEAL method, and zero when using the PPPC or the TPC as this would indicate zero learning. However, if there was some prior knowledge on the CQ, we would expect the TPC to be greater than zero and the CQ to lie in either Quad 1 or 2 when using the ATEAL method. The following sections present the results by each utility as the concepts trained varied by the industry.

5.3.1 Natural Gas Utility

Table 5-3 illustrates the training performance metrics calculated for the Tier 1 Employee training group in the Natural Gas Utility sector. A total of 405 participants answered each question/concept in this training group. If one used TPC to measure training effectiveness, the conclusion would be that Vehicle Safety, Employee Rights & Responsibilities, and Slips, Trips & Falls are the best learned concepts by this group. However, the PPPC indicates that Hearing Loss is the best learned concept by a large margin.

Table 5-3: Natural Gas Utility – Tier 1 Employees (n=405) assessment result metrics.

	Concept	TPC*	PPPC*	LAC	NTIC
1	Control Question	11%	4%	0.41	-0.47
2	Employee Rights & Responsibilities	99%	9%	0.50	0.90
3	Environment	88%	9%	0.44	0.64
4	General	82%	27%	0.56	0.58
5	Hearing Loss	84%	67%	0.83	0.72
6	Overexertion	82%	31%	0.62	0.71
7	Slips, Trips & Falls	99%	22%	0.61	0.99
8	Vehicle Safety	100%	4%	0.52	1.00
9	Vibration	87%	-2%	0.34	0.50

In applying the ATEAL method and plotting these 9 concepts on the Training Effectiveness Matrix, as shown in Figure 5-3, it is clear that Hearing Loss is the best learned concept, and that the prior knowledge level was low. Employee Rights & Responsibilities had about equal number of participants who had prior knowledge as participants who learned the concept. The participants all experienced positive learning for the concepts of General, Hearing Loss, Overexertion, Slips Trips & Falls, and Vehicle Safety.

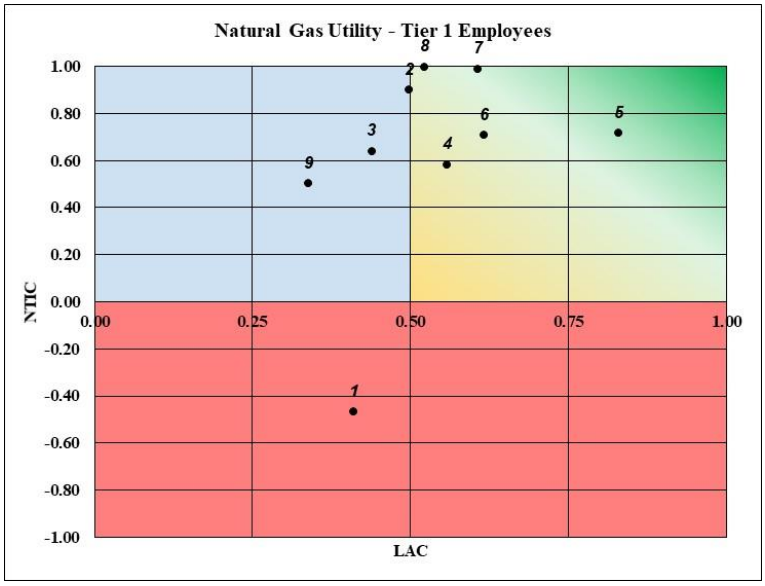

Figure 5-3: Training Effectiveness Matrix for the Natural Gas Utility – Tier 1 Employees

The participants exhibited considerable prior knowledge for the concepts of Environment and Vibration (numbered 3 & 9 in Figure 5-3). A total of three questions were associated with these two concepts and, since there was such a high amount of prior knowledge among the participants, it would potentially have been a better use of participant time to reduce the number of questions and amount of training on these concepts and instead focus on the other concepts that needed to be learned. Finally, the ATEAL method does an excellent job in identifying the CQ (numbered 1 in Figure 5-3) among the concepts taught. As indicated previously, the CQ is a concept that was not taught in the training, but was thematically similar to the rest of the content tested, and was used to estimate the amount of guessing by the participants. The results show that there was more negative training impact than positive training on the CQ and that the participants were having difficulty answering the question. This is the only question for which the NTIC is less than zero. By having the CQ and using it along with the other assessment results, we can clearly see that the ATEAL method helps provide considerably higher resolution in understanding the effectiveness of the training of each concept compared to the PPPC metric.

Table 5-4 illustrates the training performance metrics calculated for the Tier 2 Employee training group in the Natural Gas Utility sector. A total of 347 participants answered each question/concept in this training group. Similar to the Tier 1 Employee group, the TPC metric does not indicate that Hearing Loss is the best learned concept as it includes the prior knowledge in the final assessment results reported. However, the PPPC identifies Hearing Loss as the best learned concept.

Table 5-4: Natural Gas Utility – Tier 2 Employees (n=347) assessment result metrics

	Concept	TPC	PPPC	LAC	NTIC
1	Control Question	42%	27%	0.55	0.10
2	Employee Rights & Responsibilities	96%	8%	0.50	0.85
3	Environment	87%	6%	0.42	0.60
4	General	77%	23%	0.52	0.49
5	Hearing Loss	81%	56%	0.78	0.63
6	Overexertion	78%	21%	0.53	0.58
7	Slips, Trips & Falls	97%	16%	0.53	0.88
8	Vehicle Safety	100%	7%	0.51	0.96
9	Vibration	83%	-7%	0.30	0.41

In applying the ATEAL method, the Training Effectiveness Matrix for these 9 concepts, shown in Figure 5-4, clearly identifies Hearing Loss (numbered 5 in Figure 5-4) as the best-learned concept. The rest of the concepts have very similar results to those observed with the Tier 1 Employee training group, with the participants having higher prior knowledge for the Environment and Vibration concepts (numbered 3 & 9 in Figure 5-4). The CQ (numbered 1 in Figure 5-4) for Tier 2 trainees lands in Quad 1, whereas for the Tier 1 training group it was in Quad 3. This indicates that there was more positive learning on the CQ than both prior

knowledge and negative training. However, its magnitude is very low (close to 0.5,0) indicating that the net learning was almost zero. This could be explained by the fact that more participants in the Tier 2 Employee group guessed correctly on the CQ compared to the Tier 1 Employee group.

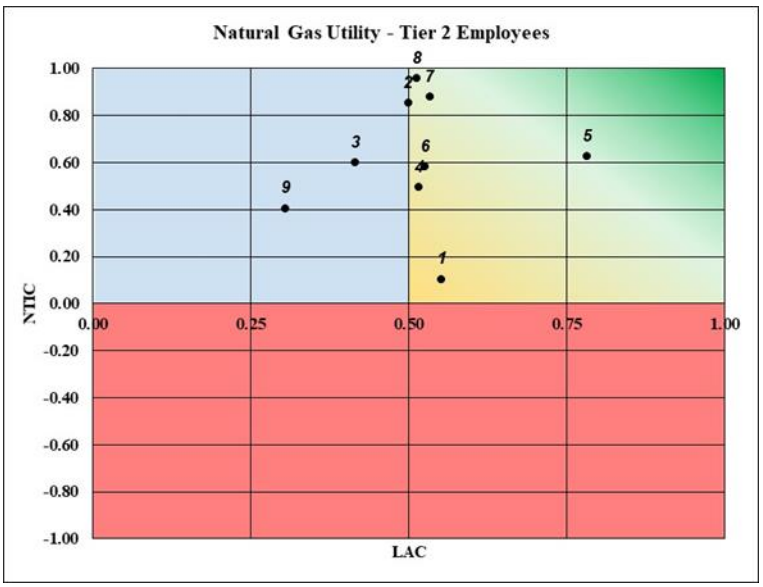

Figure 5-4: Training Effectiveness Matrix for the Gas Utility – Tier 2 Employees

Table 5-5 illustrates the training performance metrics calculated for the Manager training group in the Natural Gas Utility sector. A total of 78 participants answered each question/concept in this training group. Similar to the Tier 1 and Tier 2 Employee groups, the TPC metric does not indicate that Hearing Loss is the best learned concept by the training participants as it includes the prior knowledge in the final assessment results reported; however, the PPPC identifies Hearing Loss as the best learned concept.

Table 5-5: Natural Gas Utility – Manager (n=78) assessment result metrics

	Concept	TPC	PPPC	LAC	NTIC
1	Control Question	14%	-15%	0.22	-0.62
2	Employee Rights & Responsibilities	96%	9%	0.51	0.87
3	Environment	91%	10%	0.48	0.74

4	General	88%	29%	0.62	0.77
5	Hearing Loss	87%	62%	0.80	0.76
6	Overexertion	78%	50%	0.75	0.56
7	Program Implementation	97%	5%	0.48	0.87
8	Root Cause Analysis	71%	26%	0.54	0.43
9	Slips, Trips & Falls	98%	18%	0.58	0.95
10	Vehicle Safety	100%	9%	0.54	1.00
11	Vibration	92%	1%	0.31	0.53
12	Workplace Assessment	56%	21%	0.58	0.15

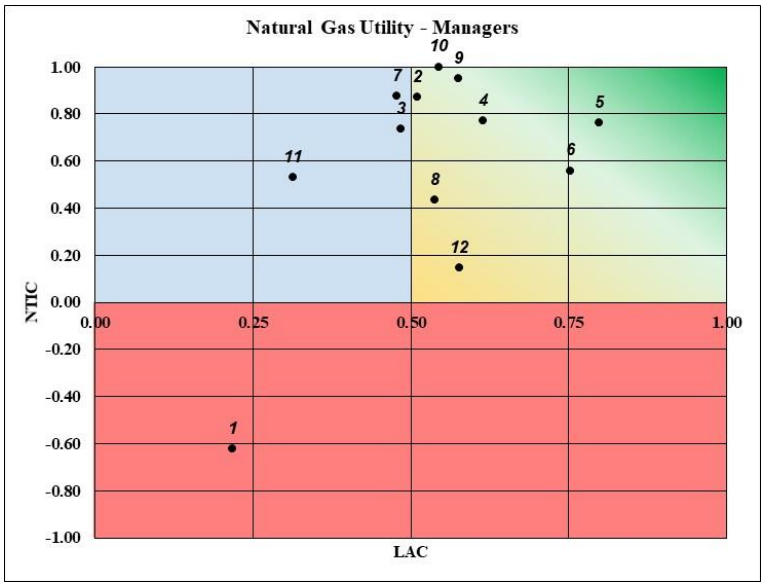

Figure 5-5: Training Effectiveness Matrix for the Gas Utility – Managers

In applying the ATEAL method, Figure 5-5 presents the Training Effectiveness Matrix for the 12 concepts for the Manager training group, and clearly shows that the CQ (number 1 in Figure 5-5) is unlike the other questions. The participants had higher prior knowledge for the concepts of Environment, Program Implementation and Vibration (numbered 3, 7 & 11 respectively in

Figure 5-5), for which there were a total of 4 questions. The Manager training group is observed to exhibit positive learning for the other 8 concepts covered in the training program.

5.3.2 Electric Transmission Utility

Table 5-6 illustrates the training performance metrics calculated for the Tier 1 Employee and Manager training group in the Electric Transmission Utility sector, as both groups were administered the same MCQ & T/F assessments. A total of 60 participants answered each question/concept in this training group. If TPC was used to measure training effectiveness, we would have concluded that Vehicle Safety, Slips, Trips & Falls and PPE are the best learned concepts by this group. However, PPPC indicates that General is the best learned concept by a large margin.

Table 5-6: Electric Transmission Utility – Tier 1 Employees (n=60) assessment result metrics.

	Concept	TPC	PPPC	LAC	NTIC
1	Electric Safety	93%	21%	0.53	0.78
2	Employee Rights & Responsibilities	95%	13%	0.49	0.80
3	Environment	88%	12%	0.50	0.70
4	General	63%	43%	0.73	0.23
5	Hearing Loss	87%	27%	0.52	0.68
6	Overexertion	79%	13%	0.48	0.56
7	PPE	96%	29%	0.40	0.43
8	Slips, Trips & Falls	96%	15%	0.60	0.95
9	Vehicle Safety	97%	8%	0.56	0.96

In applying the ATEAL method and plotting the Training Effectiveness Matrix as shown in Figure 5-6, it is clear that this group had more learning than prior knowledge for the General

(numbered 4 in Figure 5-6) concept; however, they did experience more negative training for this concept than the concepts of Slips Trips & Falls and Vehicle Safety (numbered 8 & 9 in Figure 5-6). The participants had considerably higher prior knowledge for the concepts of Employee Rights & Responsibilities, Environment, Over Exertion and PPE (numbered 2, 6 & 7 respectively in Figure 5-6). Thus, despite having PPPC scores of 13%, 12%, 13% and 29% respectively, they still lie in Quad 2, thus indicating that there was a lower need to train on these concepts. There was no CQ for this training group.

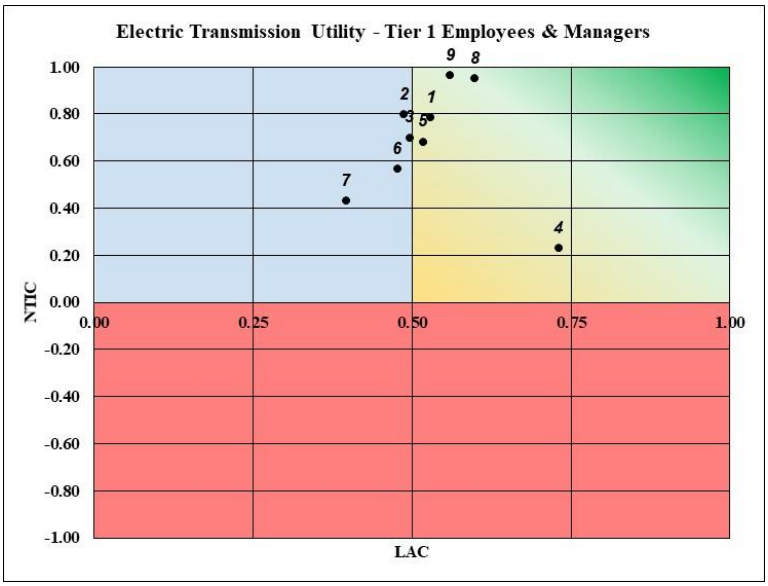

Figure 5-6: Training Effectiveness Matrix for the Electric Transmission Utility – Tier 1 Employees & Managers

Table 5-7 illustrates the training performance metrics calculated for the Tier 2 Employee training group in the Electric Transmission Utility. If we used the TPC to measure training effectiveness, we would have concluded that Vehicle Safety, Hearing Loss and Slips, Trips & Falls are the best learned concept by this group. However, PPPC would have us conclude that the General and Control Question concepts were the best learned by a large margin.

Table 5-7: Electric Transmission Utility – Tier 2 Employees assessment result metrics.

	Concept*	TPC	PPPC	LAC	NTIC
1	Control Question (n=293)	45%	33%	0.60	0.19
2	Electric Safety (n=292)	88%	22%	0.54	0.69
3	Employee Rights & Responsibilities (n=217)	88%	0%	0.32	0.48
4	Environment (n=217)	87%	2%	0.40	0.58
5	General (n=217)	68%	38%	0.59	0.49
6	Hearing Loss (n=287)	93%	21%	0.56	0.81
7	Overexertion (n=293)	87%	7%	0.42	0.62
8	PPE (n=217)	77%	5%	0.31	0.36
9	Slips, Trips & Falls (n=293)	93%	11%	0.43	0.70
10	Vehicle Safety (n=293)	94%	-1%	0.24	0.60

Note. *Where 'n' is the number of participants answering questions on that specific concept

The PPPC indicates that there is negative learning for the concept of Vehicle Safety. This, however, is different from the information we observe when using the ATEAL method. In reviewing the Training Effectiveness Matrix for this training group, shown in Figure 5-7, we observe that none of the concepts exhibited negative learning. The participants have higher prior knowledge on the concepts of Employee Rights & Responsibilities, Environment, Overexertion, PPE, Slips Trips & Falls and Vehicle Safety (numbered 3, 4, 7, 8, 9 & 10 respectively in Figure 5-7). The trainees exhibit positive learning for the concepts of Electric Safety, Hearing Loss and General (numbered 2, 6 & 5 respectively in Figure 5-7). Of these concepts, the General and

Hearing Loss concepts (numbered 5 & 6 in Figure 5-7) are the best learned concepts. The CQ also lies in Quad 1 for this training group, potentially due to correct guessing by the participants.

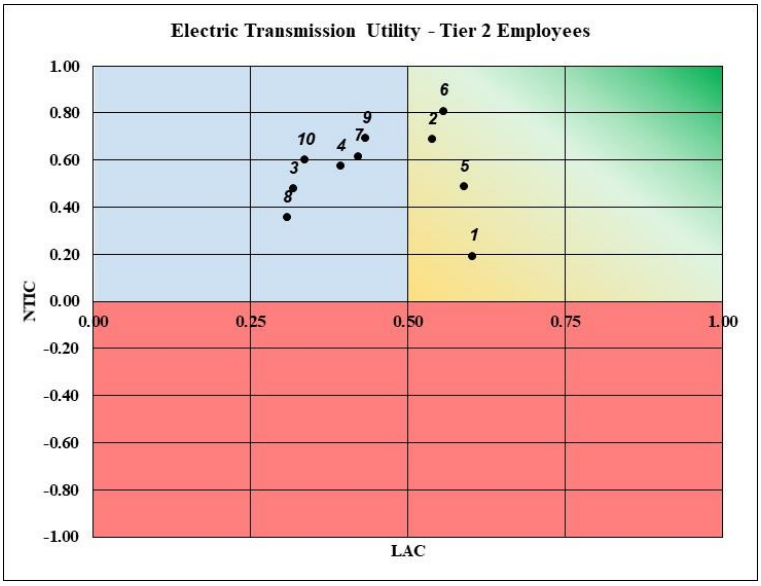

Figure 5-7: Training Effectiveness Matrix for the Electric Transmission Utility – Tier 2 Employees

5.3.3 Power Generation Utility

Table 5-8 illustrates the training performance metrics calculated for the Managers training group in the Power Generation Utility. A total of 12 participants answered each question/concept in this training group. If the TPC was used to measure training effectiveness, we would have concluded that Confined Space and Environment are the best learned concept by this group. However, PPPC would have us conclude that Confined Space, Hearing Loss and Struck by/caught between are the best learned concepts. Additionally, the PPPC illustrates that the Managers had a positive learning experience for the CQ.

Table 5-8: Power Generation Utility – Managers (n=12) assessment result metrics.

	Concept	TPC	PPPC	LAC	NTIC
1	Confined Space	100%	33%	0.67	1.00
2	Control Question	58%	58%	0.79	0.58
3	Environment	100%	8%	0.54	1.00

4	General	92%	17%	0.42	0.67
5	Hearing Loss	92%	33%	0.63	0.88
6	Overexertion	94%	13%	0.60	0.92
7	Program Implementation	97%	17%	0.55	0.92
8	Root Cause Analysis	25%	17%	0.64	-0.73
9	Slips, Trips & Falls	92%	13%	0.51	0.78
10	Struck by/caught between	42%	33%	0.58	0.36
11	Workplace Assessment	27%	27%	0.64	0.27

Using the ATEAL method, the Training Effectiveness Matrix for the Power Generation Utility Managers is shown in Figure 5-8. It shows that more of participants have prior knowledge for the General concept (numbered 4 in Figure 5-8) than those that learned the concept. The matrix shows that the participants experienced significant negative learning for the Root Cause Analysis concept (numbered 8 in Figure 5-8), a detail that could not be discerned by looking at the TPC or the PPPC metrics. The matrix also shows that, other than the two concepts detailed above, the participants experienced positive training for all the other concepts including the CQ.

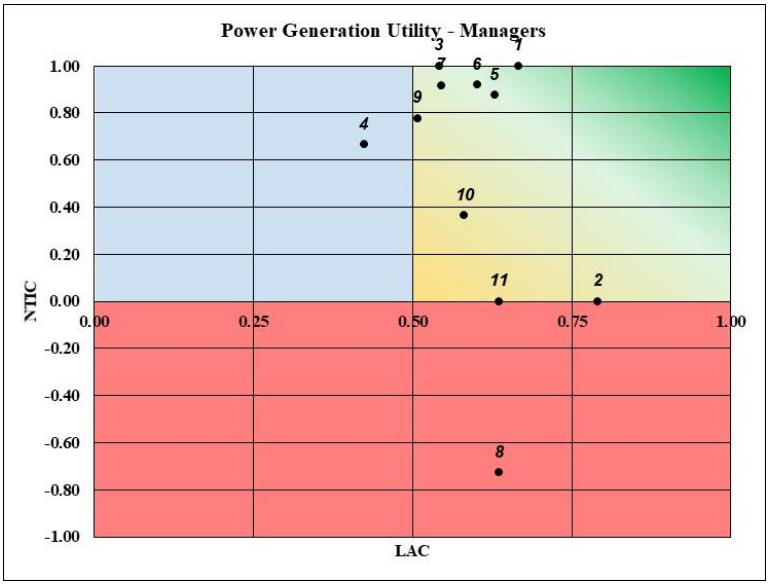

Figure 5-8: Training Effectiveness Matrix for the Power Generation Utility – Managers

Table 5-9 illustrates the training performance metrics calculated for the Employees training group in the Power Generation Utility. A total of 176 participants answered each question/concept in this training group. If we use the TPC to measure training effectiveness, we would conclude that Environment and Slips Trips & Falls are the best learned concept by this group. However, PPPC would have us conclude that Confined Space and Struck by/caught Between are the best learned concepts.

Table 5-9: Power Generation Utility – Employees (n=176) assessment result metrics.

	Concept	TPC	PPPC	LAC	NTIC
1	Confined Space	73%	22%	0.44	0.46
2	Control Question	30%	10%	0.36	-0.09
3	Employee Rights & Responsibilities	90%	9%	0.38	0.59
4	Environment	94%	7%	0.43	0.72
5	General	85%	13%	0.40	0.55
6	Hearing Loss	54%	15%	0.35	0.21
7	Overexertion	91%	14%	0.47	0.70

8	Slips, Trips & Falls	93%	12%	0.47	0.75
9	Struck by/caught between	76%	21%	0.45	0.48

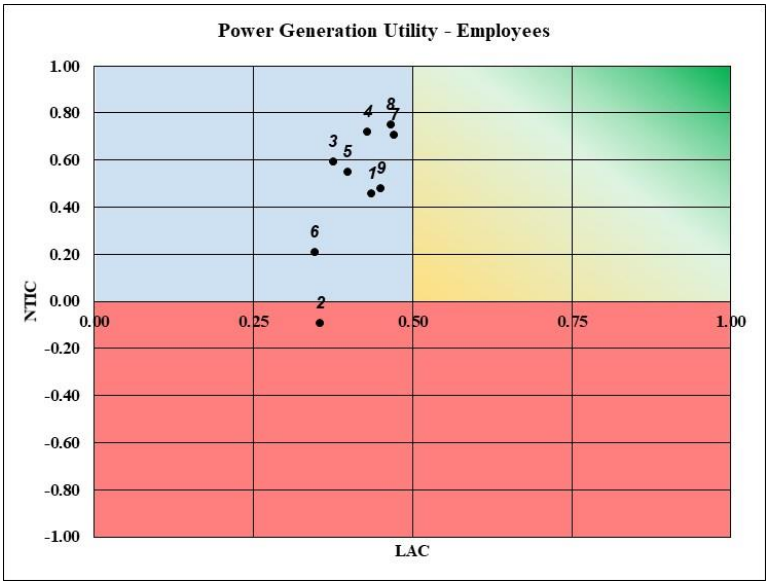

Figure 5-9: Training Effectiveness Matrix for the Power Generation Utility – Employees

Using the ATEAL method and the Training Effectiveness Matrix shown in Figure 5-9 for the Power Generation Utility – Employees training group, we observe that, despite the results reported by the TPC and the PPPC, the a larger number of trainees had prior knowledge on all of the training concepts. The method also clearly isolates the CQ (numbered 2 in Figure 5-9) by illustrating that it lies in Quad 3 and that the participants were having a difficult time answering it. This implies that considerable improvement is needed on the concepts trained for the training to be useful to the participants and the organization.

5.4 Discussion

The results from the different training groups using the various assessment metrics demonstrated that the ATEAL method is easy to interpret and is helpful in determining the steps that need to be taken to improve the training. For consistency and flow, the discussion is presented by utility as was the case in the Results section.

5.4.1 Natural Gas Utility

For the Natural Gas Utility Tier 1 Employees, the ATEAL method identifies 'Hearing Loss' as the best learned concept followed by the concepts of Slips Trips & Falls, Overexertion, General and Vehicle Safety. The participants exhibited higher prior knowledge than positive training for the concepts of Environment and Vibration. This implies that the trainer could either reduce the time spent on training these concepts to the participants, or train content within the concepts that would be more value added to the participants in terms of gaining new knowledge. It is very interesting to note that the ATEAL matrix clearly identifies the CQ as a question the participants had trouble answering and places it in Quad 3. Hence, we can easily conclude that there was more negative learning than positive learning for the CQ. Since this question was not taught in the training session, it is appropriate and correct that the matrix separates it from the rest of the questions taught. Similar trends are observed in the Natural Gas Utility Tier 2 Employees, where Hearing Loss is the best learned concept and the participants had very high prior knowledge for the concepts of Environment and Vibration. In the case of the CQ, however, the matrix shows that there was almost zero learning as it is very close to (0.5, 0). This is the expected outcome in the ATEAL method as the CQ concept was not taught during the training and we would expect no positive or negative learning. Finally, the Natural Gas Utility Managers exhibit the same trends as the Tier 1 Employees. That is, Hearing Loss is the best learned concept and the CQ is located in Quad 3, clearly separated from the rest of the questions.

In the companion paper, we observe that the TPC metric is overly optimistic in its depiction of participant performance in the case of the scenario and simulation analysis. These trends are again seen when analyzing actual training data and the implications are more profound. In the Natural Gas Utility, TPC shows Vehicle Safety to be the best learned concept for the Tier 1

Employees, Tier 2 Employees and Managers. This, however, is due to the participants having very high prior knowledge (over 95%) of this concept; that is, they were able to answer it correctly in both the pre- and post-test assessments. Looking strictly at the TPC metric, the trainer would have interpreted that 'Vehicle Safety' was the best learned concept among the 875 participants in the Natural Gas Utility and that its content and method of delivery was highly effective due to its positive impact with such a large number of participants. However, because of the high level of prior knowledge, there should have only been a cursory overview of this concept and an argument can be made that it did not need to be tested in the post-test assessment.

The results of the PPPC in the scenario and simulation analysis in the companion paper show that it is better at compensating for prior knowledge than TPC. This benefit is further observed when looking at the results of actual training and assessments conducted on the participants from the various utility companies. For the Tier 1 Employees in the Natural Gas Utility, the PPPC identifies the concept of 'Hearing Loss' to be the best learned concept and 'Vibration' to the worst learned concept. In looking at the actual performance of the participants for 'Vibration' (CC=80%; IC=7%; CI=9%) we observe that its negative PPPC value is due to the high prior knowledge among the participants and the small number of participants who experienced negative learning. The PPPC metric also does not isolate the CQ and, although it reports a low performance of the participants for the CQ, it is in line with the results for 'Vehicle Safety' which had a low score due to very high prior knowledge. The same trends for the concepts are observed for Natural Gas Utility Tier 2 Employees. For the Managers, the concept of 'Hearing Loss' is identified as the best learned concept and the CQ receives a negative score.

5.4.2 Electric Transmission Utility

Using the ATEAL method to analyze the data for the Electric Transmission Utility, for Tier 1 Employees there was positive learning on six of the nine concepts taught, with the 'General' concept being the best learned because the participants had the least prior knowledge and comparatively learned the most on this concept. For the Electric Transmission Utility Tier 2 Employees, the concept of 'Hearing Loss' was the best learned concept and the participants exhibited positive learning on four of the ten concepts tested. The CQ, as seen before, exhibited low learning and, although it is in Quad 1, it is the closest of all the concepts taught to (0.5,0). When using the TPC to analyze the data in the Electric Transmission Utility, the concept of 'Vehicle Safety' again seems to be the best learned concept by the Tier 1 & 2 Employees due to the high level of prior knowledge (over 85%) among the participants. In using the PPPC to analyze the data of the Tier 1 Employees, the 'General' concept is identified as the best learned concept and 'Vehicle Safety' as the least learned concept. This is the exact opposite of the results from the TPC metric, and is a more accurate representation of participant knowledge levels as the participants had the highest amount of prior knowledge for 'Vehicle Safety'. Similarly, for the Tier 2 Employees in the Electric Transmission Utility, the 'General' concept is identified as the best learned concept. Due to high prior knowledge and a small number of participants experiencing negative learning, the metric identifies 'Vehicle Safety' and 'Employee Rights & Responsibilities' as the worst learned concept.

5.4.3 Power Generation Utility

Using the ATEAL method, we observe that there was positive learning on eight of the eleven concepts on which Managers were tested. It is extremely interesting to observe that the CQ was the best learned concept, as over 50% of the participants went from incorrect and IDK responses

to the CQ in the pre-test assessment to correct responses in the post-test assessment. This could imply that the concept was inadvertently trained in the class by the trainer or that the participants were able to correctly guess the post-test answer. We observe that the concept of Root Cause Analysis had considerable negative training impact and very low prior knowledge. This is a critical issue as this concept is key for the Managers to diagnose safety issues correctly and implement countermeasures to improve the safety of the employees. In further researching the results, we observe that 58% of the participants exhibited zero learning; thus, it is important for the trainers to revisit this concept with this group to ensure that they understand and learn the concepts. It is not possible to quickly arrive at this conclusion when solely looking at TPC and PPPC metrics. Hence, this shows that using the ATEAL method is better and quicker at helping discern participant learning and helps trainers determine countermeasures in an expeditious manner. For the Power Generation Utility Employees, we observe that the CQ lies in Quad 3 and we observe that, for all the other concepts taught, the participants exhibited considerably higher prior knowledge than learning. This is concerning as it shows that a majority of the participants did not learn anything new and the effective use of their time comes into question.

Using the TPC the concept of 'Environment' is shown to be the best learned concept for both the Employees and the Managers due to high prior knowledge (over 84%). The Managers of the Power Generation Utility are also shown to have high learning for the concept of 'Confined Space' as reported by this metric. For this concept there was considerably less prior knowledge (66%) and 33% of the Managers learned the concept. In using the PPPC to analyze the results for the Managers in the Power Generation Utility, we observe that the CQ is reported as the best learned concept. Although this is counterintuitive, the results are due to the 0% prior knowledge and 50% of the participants who answered correctly in the post-test assessment. The other

concepts ranked lower mainly due to the fact that participants had higher prior knowledge. Finally, for the Employees of the Power General Utility, the concept of 'Confined Space' is reported to be the best learned concept, although 48% of the participants had prior knowledge of this concept, and 'Environment' is the least learned concept due to 84% of the participants having prior knowledge of this concept.

A common observation through the results and discussion across all of the utilities is the level of prior knowledge that the participants possess for the various concepts trained. Using the ATEAL method we can clearly determine when there are more participants exhibiting prior knowledge than learning. This is impossible to determine when using the TPC metric as it does not compensate for prior knowledge and reports it as learning. Using the PPPC, it takes more time to discern if the low (or) negative values are due to high prior knowledge or negative learning. The metric does not separate the elements, so it requires additional detailed review of the raw score that takes time and effort and may not always be conducted.

The limitation of ATEAL is that the method requires the presence of matched pre- and post-training assessment results, as the analysis is based on baseline knowledge and learning and cannot be used when there is only post-training assessment results. The application of the method may also require some basic training for trainers and organizations. This training, however, is minimal, as the calculations are simple and graphics are easily implemented by using widely available software packages such as MS Excel.

One of the generalizable benefits of the ATEAL method is that it can be used for any type of assessment situation where there is a pre- and post-test assessment. For example, suppose assembly workers were being trained to improve assembly practices, and the assessment was made by an assessor observing the assembler for performance in the categories of quality, speed,

efficiency, following standard work, etc. If the assessment is made on the assembler prior to training, and a score obtained for the various categories, the training conducted and the assembler can then be reassessed on their performance post training and the ATEAL method can be used to measure the training effectiveness in this scenario. Thus, the method is more widely applicable than in just the case of MCQ assessments. This may have remarkable implications for the organizations and the participants as the training time can be reduced and the effectiveness improved simultaneously. Additionally, reduction in training time may have fiscal impacts that result in a higher return on investment (ROI) for the training with a higher focus on concepts for which the participants genuinely have knowledge gaps.

5.5 Conclusion / Future Direction

Metrics to quantify the amount of learning that training participants exhibit for a particular training course, or concepts within the course, are critical to understanding the effectiveness of the training, specifically in the context of workplace safety-related concepts. Using the ATEAL method to measure training effectiveness for training conducted with 1,466 participants from a variety of utility industries, and comparing the results to traditional measurement metrics, we observe that the ATEAL method proves very effective in quickly identifying the training gaps that the participants experienced and in giving direction on the countermeasures that should be taken for each concept trained.

Some recommendations that can be derived from this study are:

- Using only the TPC in the post-test assessment to evaluate training effectiveness (or) how much the participants learned is shown to be a highly inaccurate method and does not give clear guidance on areas of improvement.

- The PPPC is shown to be a better metric than the TPC to evaluate training effectiveness; however, it lacks the ability to quickly provide information on the changes needed in the training content or its delivery to improve training effectiveness.
- The ATEAL method uses metrics that are of greater accuracy, are easy to calculate, and provide intuitive output that allows for easy visualization of the training effectiveness results. It provides a great way to illustrate the training effectiveness of each concept taught to the participants and can be used to quickly determine the countermeasures that need to be taken by the trainer with regards to content delivery or development as part of the training program. This then provides information on how to improve training effectiveness in future training sessions on the topic. Organizations can also benefit considerably from this method as it helps them understand the concepts that the participants can be held accountable for as well as the specific concepts that need further reinforcement to ensure the employees have safe work practices in their work environment.
- Using the ATEAL method, the trainers and the organizations are able to quickly identify the concepts for which the participants had considerable prior knowledge. This enables them to focus on concepts for which the participants truly have knowledge gaps and ensure the best return of investment on the training provided and the time used for the training.

Acknowledgments

This study was partially funded by the US DOL Susan Harwood Grants: SH-20840-SH0, #SH-22220-SH1, #SH-23568-SH2, #SH-24880-SH3. The authors also express their gratitude to Karen Cooper, Sruthi Boda and Madiha Ahmed for assisting with the test development and administration. We would additionally like to thank all the companies and employees who participated in this study.

5.6 References

Ahmed, M., Campbell-Kyureghyan, N. (2014, June). Reliability of learning assessment. Proceedings of the XXVIth Annual International Occupational Ergonomics & Safety Conference, El Paso, TX. June 5-6, 2014

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human resource development review, 3*(4), 385-416.

Bahn, S., & Barratt-Pugh, L. (2012). Emerging issues of health and safety training delivery in Australia: Quality and transferability. *Procedia-Social and Behavioral Sciences, 62*, 213-222.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society, 4*(1), 3-12.

Becker, P., & Morawetz, J. (2004). Impacts of health and safety education: Comparison of worker activities before and after training. *American Journal of Industrial Medicine, 46*(1), 63-70.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065-1105.

Brunello, G., & Medio, A. (2001). An explanation of international differences in education and workplace training. *European Economic Review, 45*(2), 307-322.

Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Salvador, R. O., & Islam, G. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health, 96*(2), 315-324.

Campbell-Kyureghyan, N., Cooper, K. (2012). Impact of customized training on learning across demographic groups. *JPIIE, 9*(1): 25-31.

Campbell-Kyureghyan, N., Hernandez, A. P., & Ahmed, M. (2013). *Effectiveness of first and second tier safety and ergonomics training in power utilities*. Proceedings of the XXVth Annual Occupational Ergonomics and Safety Conference, Atlanta, GA, USA.

Campbell-Kyureghyan, N., Ahmed, M., Beschorner, K. (2013, March). *Measuring training impact 1-5*. Paper presented at the US DOL Trainer Exchange Meeting, Washington DC, March 12-13, 2013.

Caston, S., Cooper, K., & Campbell-Kyureghyan, N. (2009). Assessment of ergonomic training in small business foundries. In *Proc. XXIst Annual International Occupational Ergonomics and Safety Conference* (Vol. 9).

Demirkesen, S., & Arditi, D. (2015). Construction safety personnel's perceptions of safety training practices. *International Journal of Project Management, 33*(5), 1160-1169.

Ho, C. L., & Dzeng, R. J. (2010). Construction safety training via e-Learning: Learning effectiveness and user satisfaction. *Computers & Education, 55*(2), 858-867.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*(1), 471-499.

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2019). Evaluation of learning outcomes through multiple choice pre-and post-training assessments. *Journal of Education and Learning, 8*(3).

Tai, W. T. (2006). Effects of training framing, general self-efficacy and training motivation on trainees' training effectiveness. *Personnel Review, 35*(1), 51-65.

https://www.bls.gov/news.release/pdf/cfoi.pdf

Chapter 6: Conclusion

6.1 Summary

This research has introduced a variety of methods for improvement of training effectiveness assessment by quantifying the effect of guessing and accounting for participant prior knowledge of the concepts delivered during training. This was accomplished by addressing research questions that investigated (1) how the addition of the IDK option in the pre-test and post-test Level 2 MCQ assessment changes the proportion of correct and incorrect answers, (2) if the addition of the IDK option truly reduces the amount of guessing in pre-test and post-test assessments and (3) if the participant chooses IDK in the pre-test assessment, is there a difference in how that participant responds on the post-test assessment depending on the type of question (MCQ or a Control Question - CQ). Additional research questions on (4) how learning outcomes for the concepts taught during the training session can be assessed and (5) how the different methods used to measure training effectiveness of concepts in Level 2 assessments in a pre-/ post-test assessment model differ from each other on the concepts they report as best and least learned. This was accomplished by conducting post-hoc analysis on training assessment data collected from 1,474 participants in three major utility industries.

Scenario and simulations were also used to validate the models developed as part of this research. The outcome of this research is a detailed literature review that identified gaps that exist in the area of learning assessments and training effectiveness in addition to three peer-reviewed manuscripts that address the research questions outlined above.

The first manuscript introduced the concept of the Control Question and showed that there was a statistically significant reduction of 28% in the use of the IDK option in the post-test compared to the pre-test for all questions including the CQ. This illustrated that although the

IDK option performs as expected in reducing guessing in the pre-test assessment, it does not reduce guessing in the post-test assessment.

The second manuscript introduces the Assessment of Training Effectiveness Adjusted for Learning (ATEAL) method to measure training effectiveness to adjust learning for participant prior knowledge and poor training they might have experienced. It was found that the coefficients used in the ATEAL method were more sensitive to participant prior knowledge and negative training impact than Total Percent Correct (TPC) or Post-Pre Percent Correct (PPPC). A Training Effectiveness Matrix (TEM) was developed to visually represent the coefficients to enable ease of use of the ATEAL method.

Finally, the third manuscript details the practical application of the ATEAL method by conducting post-hoc analysis on the safety training data from participants of the various utility industries. It was found that the ATEAL method performed better at identifying the concepts that were the best learned while compensating for prior knowledge, guessing and any poor training potentially experienced. Additionally, with the use of the TEM, trainers and organizations can quickly identify gaps and take countermeasures to improve the training for the participants. Importantly, it was also found that the ATEAL method not limited to application to MCQ assessments as it can be used in any situation where there are before and after measurements made on a process where there is a transfer of knowledge.

One of the contributions of this research to the body of knowledge in the field is quantifying the impact of the IDK option on learning outcomes through using MCQs pre-/ post-test assessments. The research quantifies the impact on guessing by participants with the addition of the IDK option by introducing a concept called the Control Question (CQ). The CQ is a concept that is tested but was not trained in the session and is similar to a placebo treatment. It was

observed that the introduction of the IDK options in MCQ assessments which have more than two answer options statistically significantly reduces the number of incorrect answers by 63% (thus reducing guessing) in the pre-test assessment. However, the IDK option does not significantly reduce the amount of guessing by the participants in the post-test assessment as measured by the CQ (Samuel, Azen & Campbell-Kyureghyan, 2019). This implies that participants would rather guess at an answer in the post-test assessment rather than answer IDK, even if they did not know the correct answer. One of the key findings of this research is that the IDK option should not be used with the intention of reducing guessing in post-test MCQ assessments as proposed by various researchers and detailed in Chapter 3.

Another contribution of this research to the body of knowledge in the field is the introduction of a new Assessment of Training Effectiveness Adjusted for Learning (ATEAL) method with several new metrics, such as the Learning Adjustment Coefficient (LAC) and Net Training Coefficient (NTC) that are plotted on the Training Effectiveness Matrix (TEM) to help assess the performance of the participants on the various concepts introduced in the training. The method is effective at quickly identifying training gaps experienced by the participants and at providing direction on the countermeasures that need to be taken to improve the training effectiveness for the concept trained. The ATEAL method was further evaluated against traditionally used metrics of TPC and PPPC by conducting scenario modeling and simulation.

Finally, the ATEAL method was used to measure the training effectiveness for the training conducted with 1,474 participants from a variety of utility industries and the results show that the method proves very accurate and effective in quickly identifying the training gaps of the participants and providing directions on the countermeasures that the instructor or the organization needs to take to improve the training effectiveness. The new method was compared

to the results reported by traditional metrics and it was observed that the TPC in post-test assessments is limited in its ability to measure training effectiveness. The PPPC is a more effective metric, however it does not provide the information needed to make training effectiveness improvements in a consistent manner. Importantly, for industrial applications, the ATEAL method will allow trainers and organizations quickly identify concepts for which the participants truly have knowledge gaps to ensure a better return of investment on the time and resources used to provide the training.

6.2 Future Work

Future directions of this research could include generalization of the results to other modes of knowledge transfer like e-learning environments and teaching environments (e.g. university settings) where there is a larger time gap between the pre-test and post-test assessments. Additionally, further research can be performed to determine if changes in the content (non-safety related) and participants (individuals in a non-work environment) would change the results observed in any way.

Another possibility is to connect the results of the Level 2 assessments of the best and the worst learned concepts, as reported by the ATEAL method, to changes in the behavior of the participants in their work environments following the training and determine if there is a quantifiable impact (positive or negative) in safety incidents based on the results of the concepts learned.

Finally, it would be a valuable comparison to apply the disaggregation model, introduced by Walstad and Wager (2016) to the data sets in this research and to assess the impact of the IDK option on the learning outcome results predicted by their model.

6.3 References

Walstad, W. B., & Wagner, J. (2016). The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education, 47*(2), 121-131.

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2019). Evaluation of learning outcomes through multiple choice pre-and post-training assessments. *Journal of Education and Learning, 8*(3).

Chapter 7: Curriculum Vitae

Thomas Samuel, BE, MS, PhD

Education:

B.E., University of Madras, India, May 1998
Major: Mechanical Engineering

M.S., Wichita State University, Wichita, Kansas, June 2000
Major: Industrial and Manufacturing Engineering

Ph.D., University of Wisconsin – Milwaukee, WI, May 2020
Major: Engineering

Dissertation Title: Novel Approach in Measuring Training Effectiveness

Honors & Awards:

Awarded the following at Engineering Open House 2000 held at Wichita State University for the 'Dry Machining of Aluminum' project

- 'Best Graduate Project' Institute of Industrial Engineers (IIE) Senior Chapter #54
- 'Best Project' Department of Industrial and Manufacturing Engineering, Wichita State University
- 'Best Project' Society of Manufacturing Engineers (SME) Student Chapter #14
- 'Best Project' Society of Manufacturing Engineers (SME) Senior Chapter

Summer 2019: Chancellor's Graduate Student Award (University of Wisconsin Milwaukee)

Spring 2020: Chancellor's Graduate Student Award (University of Wisconsin Milwaukee)

Recipient of the 'Kennametal Inventor's Award' for designing, developing and testing an edge preparation process that allowed large scale applicability of media blasting in the manufacture of edge preparation on carbide drills

Industry Conference Presentations:

Lead Panelist in the Learning Leaders Conference for:

- 2012 - Talent Development Strategies: Applying Lessons Learned from Six Sigma & Lean for Continuous Improvement
- 2013 - Continuous Improvement L&D Strategies Roundtable Panel

Lead Panelist in the 2014 Rockwell Collins Lean Conference in Cedar Rapids, IA

Teaching Experience:

Adjunct Professor – Cardinal Stritch University: MBA program

Adjunct Professor – MSOE: School of Engineering

Adjunct Professor – UW Milwaukee: School of Engineering

Member of UWM Industry Advisory Board

Publications:

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2019). Evaluation of Learning Outcomes Through Multiple Choice Pre-and Post-Training Assessments. Journal of Education and Learning, 8(3).

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2020). Assessment of Training Effectiveness Adjusted for Learning (ATEAL). Part I: Method Development and Validation – In preparation

Samuel, T., Azen, R., & Campbell-Kyureghyan, N. (2020). Assessment of Training Effectiveness Adjusted for Learning (ATEAL). Part II: Practical Application – In preparation

Paper 4 – Critical review – In preparation

Certifications:

Certified Design For Six Sigma Master Black Belt (DFSS MBB)

ASQ certified Six Sigma Black Belt – Certification # 5772

ASQ certified Six Sigma Greenbelt.

Certified ISO 9001/2000 Internal Auditor.

Profile

- '20' years of experience in various leadership roles with diverse experience in Continuous Improvement (CI), Mergers & Acquisitions, Operations Leadership, Product Development, Manufacturing, Quality and SG&A environments.
- Significant experience in leading and deploying Six Sigma, Lean and Project Management Programs to global audiences.

Professional Experience

Vice President – Operational Excellence
BW Forsyth Partners, St. Louis, MO..... September 2019 to Current

Sr. Director – Rexnord Business System (RBS)
Rexnord-CENTA, Germany ..... April 2015 to August 2019

Corporate CI Manager
Harley-Davidson Motor Company, WI, USA .....September 2008 to April 2015

Various Quality & Engineering roles
Harley-Davidson Motor Company, WI, USA ..... March 2005 to August 2008

Quality & Lean Manager
Kennametal IPG, Augusta, GA..... July 2003 to March 2005

Hone Engineer, Manufacturing Systems Dept.,
Kennametal Inc., Latrobe, PA.....September 2000 to August 2003