Journal of Education and Learning; Vol. 9, No. 5; 2020
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

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

Thomas Samuel[1], Razia Azen[2] & Naira Campbell-Kyureghyan[1,3]

[1] Department of Industrial and Manufacturing Engineering, University of Wisconsin, Milwaukee, Milwaukee, WI, USA

[2] Department of Educational Psychology, University of Wisconsin, Milwaukee, Milwaukee, WI, USA

[3] Department of Mechanical Engineering, Merrimack College, North Andover, MA, USA

Correspondence: Naira Campbell-Kyureghyan, School of Science and Engineering, Merrimack College, North Andover, MA, 01845, USA. E-mail: campbellnk@merrimack.edu

Received: July 8, 2020 Accepted: August 17, 2020 Online Published: September 11, 2020

doi:10.5539/jel.v9n5p121 URL: https://doi.org/10.5539/jel.v9n5p121

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

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 and 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 assess 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, and 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:

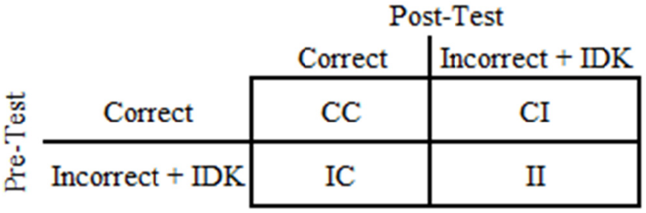
- Quickly, accurately & repeatably
- Easily interpreted, understood and acted upon to improve outcomes
- Visually impactful to communicate easily to industry stakeholders
- Usable in Multiple Choice Question (MCQ) and True/False (T/F) instances when an I Don't Know (IDK) option is present

The aim of this paper is to introduce 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.

2. Method

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 1. Additionally, Figure 1 summarizes the terminology that will be used in this paper. A similar method to break down pre-/post-test assessment results into was quadrants was conducted by Walstad and Wagner (2016). 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	Incorrect + IDK
Pre-Test	Correct	CC	CI
	Incorrect + IDK	IC	II

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

In Figure 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

2.2 Traditional Assessment Metrics

Assessment 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. It is important to note that learning is the measure of success, and that training is one technique that can be used to impart the desired knowledge. In this manuscript training impact is the impact of the training on learning of the concepts that were trained.

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 (Campbell-Kyureghyan et al., 2013). It can be used both in a pre-/post-training assessment model or in a post-training only assessment model. The formula (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:

$$\text{Total Percent Correct (TPC)} = \frac{CC+IC}{CC+IC+CI+II} \tag{1}$$

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 the use of this metric to make improvements to the training programs is problematic. 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 (Bonate, 2000). The formula (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 1, both the IDK and an incorrect answer are treated identically.

$$\text{Post – Pre-Training Percent Correct (PPPC)} = \frac{CC+IC}{CC+IC+CI+II} - \frac{CC+CI}{CC+IC+CI+II} = \frac{IC-CI}{CC+IC+CI+II} \tag{2}$$

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.

2.3 Assessment of Training Effectiveness Adjusted for Learning (ATEAL)

In an effort to overcome the deficiencies of the metrics detailed above, this paper introduces and validates the ATEAL method that compensates the assessment scores for trainee prior knowledge and guessing. A description of this method 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.

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 (3) below.

$$\text{Prior Knowledge (PK)} = \frac{CC}{CC+IC} \tag{3}$$

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.

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).

$$\text{Positive Training Impact (PTI)} = \frac{IC}{IC+II} \tag{4}$$

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.

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 (5).

$$\text{Negative Training Impact (NTI)} = \frac{CI}{CC+CI} \tag{5}$$

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.

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 (6) as:

$$\text{PTI} - \text{PK} = \frac{IC}{IC+II} - \frac{CC}{CC+IC} \tag{6}$$

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 (7):

$$\text{LAC} = \frac{1+(\frac{IC}{IC+II} - \frac{CC}{CC+IC})}{2} \tag{7}$$

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.

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:

$$\text{NTIC} = \text{PTI} - \text{NTI} = \frac{IC}{IC+II} - \frac{CI}{CC+CI} \tag{8}$$

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.

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 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.

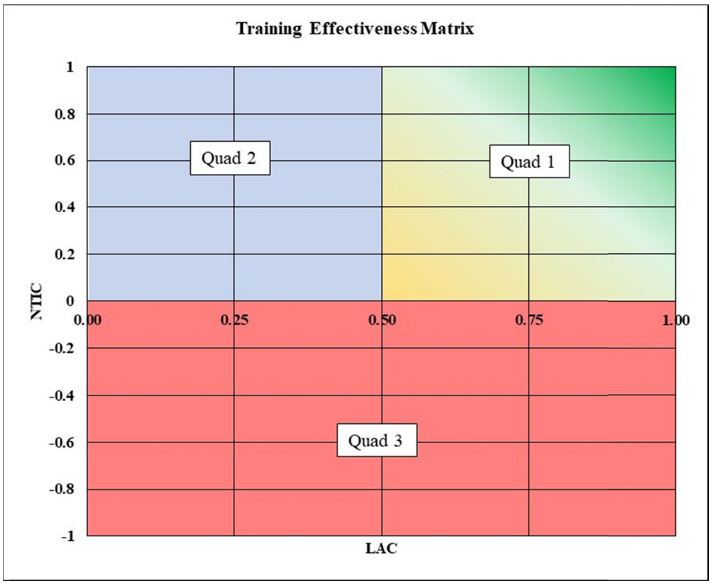


Figure 2. Training Effectiveness Matrix with the quadrant layout

Quad 1 contains the questions/concepts for which the percentage of participants had more positive learning than either prior knowledge or negative learning. 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 and no prior knowledge or negative learning, then the question would score as (1, 1) on the axes in Figure 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 learning; however, the question did not experience more negative learning than positive learning. 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 learning and it outweighs any positive learning. 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.

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.

2.4.1 Hypothetical Scenarios

The scenarios, detailed in Table 1, were developed to represent the responses (using the categories from Figure 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 1 included various combinations of complete (C), high (H), moderate (M), and zero (Z) levels of Baseline knowledge, Positive learning, and Negative learning.

Table 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 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 3 in the Results section).

2.4.2 Data Simulation

To further expand on the scenarios modelled and examine a larger population of questions and trainees, 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 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 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

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.

3.1 Scenario Results

Table 3 illustrates the metrics calculated for each of the twelve scenarios detailed in Table 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 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

Note. TPC – Total Percent Correct; LAC – Learning Adjustment Coefficient; PPPC – Post – Pre-Training Percent Correct; NTIC – Net Training Impact Coefficient.

To visualize the implication of each scenario, the TEM is provided in Figure 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.

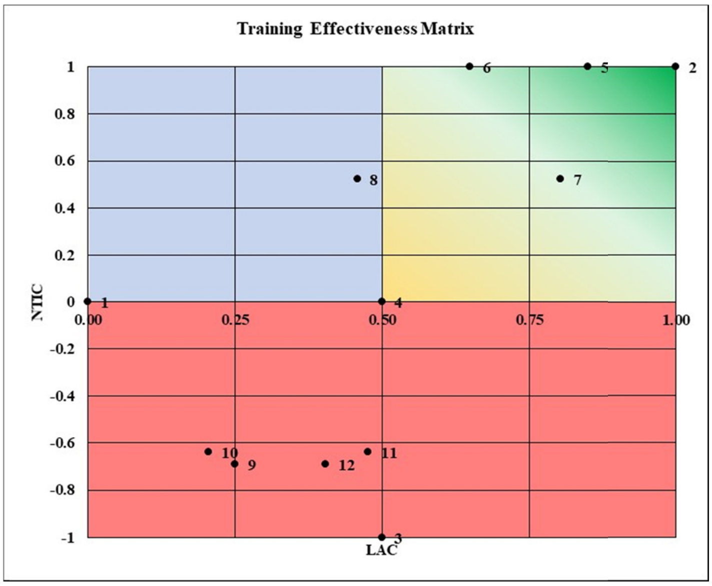


Figure 3. Training Effectiveness Matrix for the 12 scenarios

3.2 Simulation Results

Figure 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 left corner of the TEM does not contain any data points.

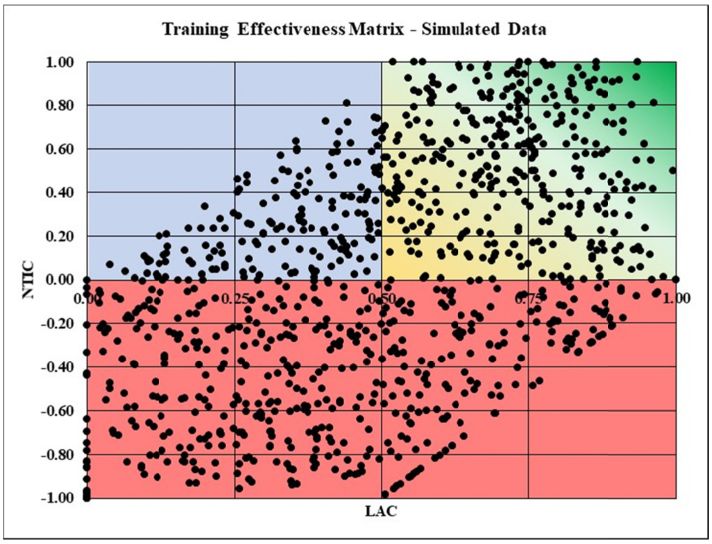
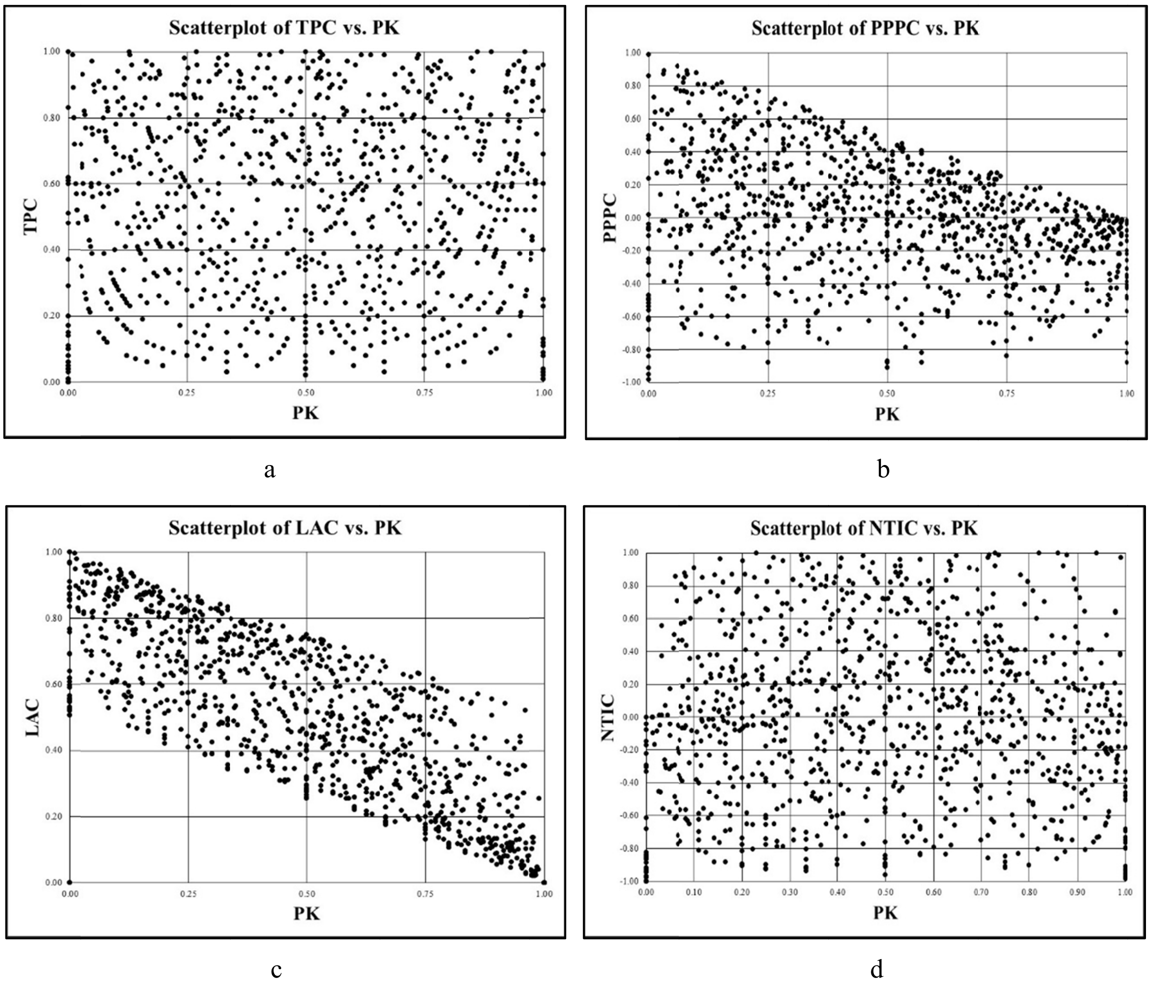


Figure 4. Training Effectiveness Matrix for the 1000 simulated data points.

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 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.


Scatterplot of TPC vs. PK

Scatterplot of PPPC vs. PK

a

b

Scatterplot of LAC vs. PK

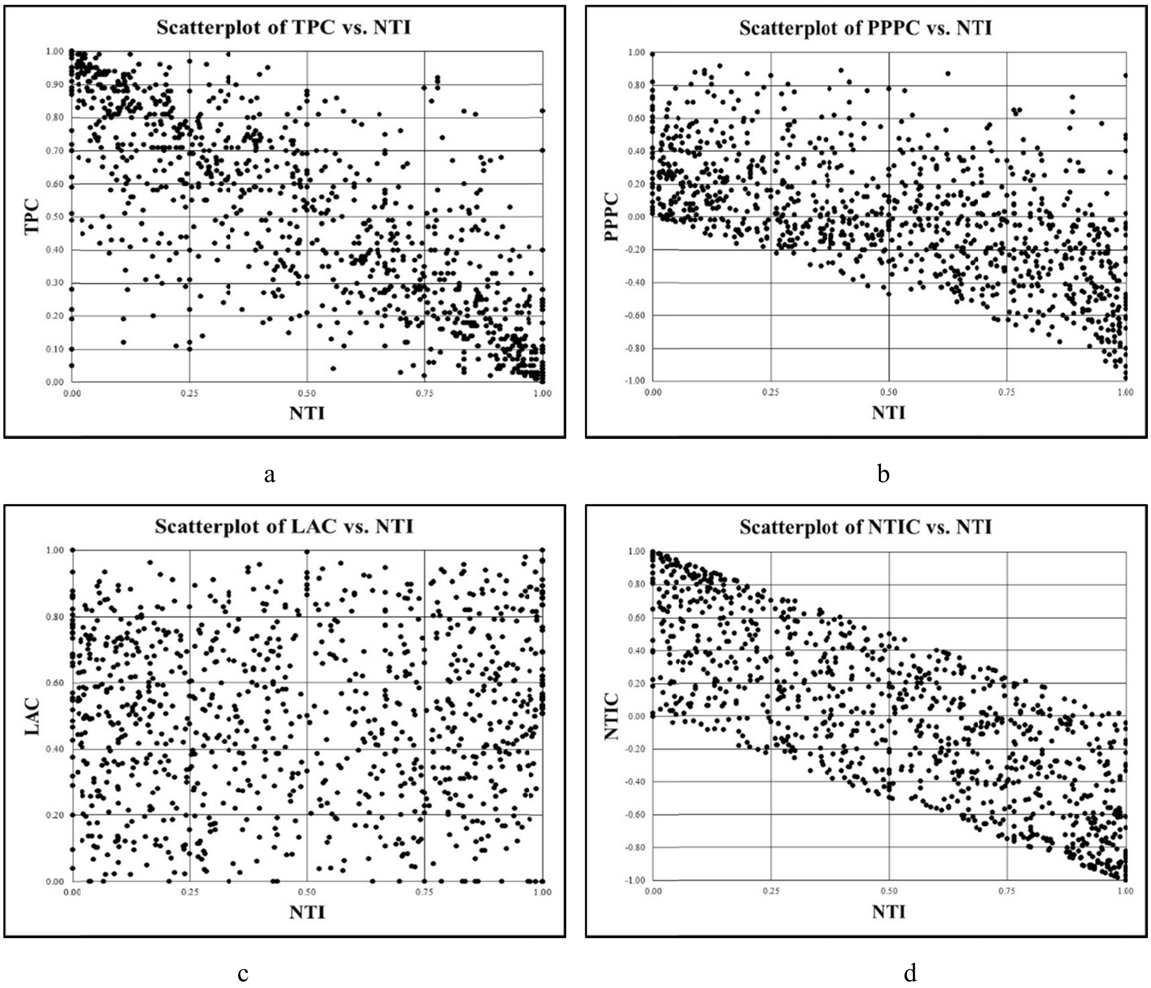
Scatterplot of NTIC vs. PK

c

d

Figure 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 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.



a

b

c

d

Figure 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. Discussion

The analysis conducted illustrates trends in 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.

4.1 Scenario Comparisons

In applying the ATEAL method and plotting the scenarios on the Training Effectiveness Matrix (TEM), we see that Quad 1 (Figure 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 3, 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.

4.2 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.

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.

Acknowledgements

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.

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. https://doi.org/10.1177/1534484304270820

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. https://doi.org/10.1037/0021-9010.88.2.234

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. https://doi.org/10.1007/s11299-005-0001-z

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420035926

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 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

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. https://doi.org/10.1177/1534484302011003

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, *52*(1), 471–499. https://doi.org/10.1146/annurev.psych.52.1.471

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), 122. https://doi.org/10.5539/jel.v8n3p122

Simkins, S., & Allen, S. (2000). Pretesting students to improve teaching and learning. *International Advances in Economic Research*, *6*(1), 100–112. https://doi.org/10.1007/BF02295755

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://doi.org/10.1108/00483480610636786

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology*, *43*(1), 399–441. https://doi.org/10.1146/annurev.ps.43.020192.002151

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. https://doi.org/10.1080/00220485.2016.1146104

Copyrights