3.1 EVALUATING LEARNING AND PERFORMANCE IN THE WARNING DECISION TRAINING BRANCH'S DISTANCE LEARNING OPERATION COURSE

Brandon Albert Miller*
School of Earth and Atmospheric Science, Georgia Institute of Technology
Atlanta, GA

Bradford N. Grant
NOAA/NWS/Warning Decision Training Branch
Norman, OK

Wilson J. Gonzalez-Espada
Arkansas Tech University
Russellville, AR

1. INTRODUCTION

One of the primary goals for National Oceanic and Atmospheric Administration's (NOAA's) National Weather Service (NWS) Warning Decision Training Branch (WDTB) is to increase expertise among NWS forecasters in order to better serve the public in warning situations. WDTB delivers training on the integrated elements of the warning process in the form of distance learning courses such as the Distance Learning Operations Course (DLOC), and the Advanced Warning Operations Course (AWOC). The AWOC is a course designed to provide every NWS forecaster advanced training on warning decision making knowledge, skills, and abilities (KSAs). These KSAs deal with aspects of science, technology, and human factors in warning decision making. The prerequisite for enrolling in the AWOC is that students must have completed either DLOC or the in-residence WSR-88D Operations Course (taught from 1991-1997). DLOC is designed to improve a forecaster's ability to effectively use radar data in forecasts and warnings. See www.wdtb.noaa.gov for details of the curriculum in DLOC (WDTB, 2005).

A specific goal of both WDTB's training courses is to improve performance by modifying behaviors to fit a desired standard. In order to modify a behavior, however, learning of the new behavior and the skills to apply it are first required (Kirkpatrick, 2005). How does one know if the behaviors and/or skills have been learned?

Assessment, or evaluation of the learning, can provide insight into the amount of knowledge gained and/or retained as a result of a training intervention. Common ways of assessing learning (also referred to as level 2 evaluation) include formal testing of knowledge or performance, demonstrations of learning that has been accomplished, surveys, interviews, or any combination of these (Kirkpatrick, 2005). Once learning can be determined, the extent to which the learners are using the training on the job can more effectively be assessed. This type of training evaluation, which attempts to measure behavior changes, or the transference of learning to job performance, is often called level 3 evaluation. This study, which was accomplished in the summer of 2005 as part of the Research Experiences of Undergraduates (REU) program at the National Weather Center in Norman, Oklahoma, helps determine the extent of level 2 learning from DLOC, and tries to relate the extent that DLOC learning has been applied in operations (level 3 evaluation). In addition, the authors will show the degree of correlation between the amount of usage and retention of DLOC material. Finally, we will investigate any statistically significant differences between the 2004 and 2005 DLOC students as a whole.

2. EVALUATION THEORY

This study is based on the four level training evaluation model presented by Kirkpatrick (1994, 2005). The four levels are:

- Level 1: *reaction* (how do trainees react to the training)
- Level 2: *learning* (to what extent has learning occurred)

* *Corresponding author address:* Brandon Miller, School of Earth and Atmospheric Science, Georgia Institute of Technology, Atlanta, GA 30313. gtg240h@mail.gatech.edu

- Level 3: *behavior* (how much has on-the-job
- Level 4: *results* (what impacts has the training had on the organization either monetarily or intrinsically)

Training is typically measured either quantitatively or qualitatively, to evaluate the effectiveness of the learning event. It is appropriate to evaluate first, at levels 1 and 2 before assessing level 3 (behavior). This way the previous levels can be understood as they relate to the training being evaluated. Determining the chain of impact of various components within the four levels is an important analysis within training organizations. For example, instruction will likely not be retained as effectively if the trainees are not accepting or responsive to the training (i.e. level 1 to level 2).

3. DATA COLLECTION PROCESS

In this study, level 2 data was collected in the form of an online test from two samples of former DLOC students, one group which completed the course in January 2004, and the other in January 2005. The sample of former students was representative of students from all over the country. To facilitate gathering the largest and most representative data set possible, a post-test was provided to every student participating in the 2004 and 2005 DLOC (~ 160 people). Sixty-six former students (28 in the 2004 class and 39 in the 2005 class) accepted to take the voluntary assessment and were included in this study. This gives a response rate of just over 40%. Although the sample was self-selected, it represents a significant proportion of our population of interest and sample bias was not expected. The former DLOC students were separated into one of 6 regions; Eastern, Central, Southern, Western, Pacific/Alaskan, and Other (which contained forecasters in national centers such as the National Hurricane Center).

Within the study, the quasi-independent variables were the exam results gathered during the DLOC course, while the quasi-dependent variables consisted of the new test results and the 7 reported levels of application recorded in the survey. These variables were measured with a content test and an applicability survey. In the post-test given to DLOC students, the scoring was made completely objective by making the test a multiple-choice test, in which one and only one answer for each question is correct. The test, which is a 25 item test, was of significant length to properly gauge whether learning of the course objectives had taken place. Since DLOC is a distance learning course, each trainee completed the tests online. To ensure reliability, the test was implemented by a similar online testing system that the students completed during the actual DLOC. The performance domains for each of the tests in DLOC originate from the learning objectives in the course. These objectives are measured directly and objectively, helping to ensure both reliability and content validity. One of the 5 exams in DLOC covers topics in the Instructional Component (IC) entitled, “Convective Storm Structure and Evolution.” This IC contains material which has been tested over in previous years of DLOC. Therefore, the exam on Convective Storm Structure and Evolution was a logical choice for retesting former students for course objective retention.

In addition to the test, a survey was included in the data collection process (see Appendix A), which provided information on the extent to which the employees had utilized DLOC instruction. This process enables us to determine some relationships between the amount of learning and application of the training material. In the survey, each student ranked their applicability of eight different aspects of DLOC, one through nine (with one constituting no application and nine as full application). The median of the eight aspects was taken to be the student’s application of DLOC. The median was chosen since it is less sensitive to outliers than the mean. Frequently, a student may not be able to use one or two aspects of DLOC due to their job assignment. So taking the mean to represent application could skew the results. The survey also collected geographical information, which helps to ensure the sample represents all regions of the NWS Weather Forecast Offices (WFOs).

The post-test, along with the accompanying survey, was placed on a designated website. An email was then sent to each of the former students explaining the nature of the research and the request for their participation. When taking the exam, the students were instructed to take no

more than 60 minutes, and to take the test "closed book", with no help from outside sources. Directions were clear that it was imperative to the research that students take the test with only the knowledge retained from DLOC. They were also assured that their scores from the test would in no way affect their job, and individual scores would never be released.

Exam scores from the original test and the post-training test were analyzed to determine if learning has been retained over time. "Acceptable retention" was set at 70%, which is the criterion established for a "passing" score in the DLOC. So, if after a period of time, the mean score on the post-test is still above a passing 70%, it can be inferred that the knowledge is sufficient enough to be applied to changing workplace behaviors.

4. STATISTICAL ANALYSIS

To compare the average scores on both tests, a Student's "T" test was used. This test determines whether two means are significantly different from each other. To measure the strength of the correlation between post-test scores and applicability, Pearson correlation coefficients were calculated. An additional test was used to determine whether the correlation coefficient was significantly different from zero. From these tests, we can draw inferences into how much long term learning has taken place, as well as whether or not we can attribute this learning solely to DLOC, or from continued use of course objectives in the workplace.

To determine if different NWS regions report different levels of applications of the concepts learned through the DLOC course, a Kruskal-Wallis test, a nonparametric version an Analysis of Variance (ANOVA) was used. In addition, Chi-Square tests were performed to determine what specific items were answered significantly different in both test administrations in order to possibly identify weaknesses in DLOC training, as well as questions with a recurring incorrect response, which could point to ambiguity in a certain test question. Making these extra observations will help the WDTB evaluate the effectiveness of DLOC, as well as providing some possible directions in improving it. Throughout the research, an alpha level (or significance level) of 0.1 was chosen, meaning that only p-values of less than 0.1 would be considered significant. An alpha value of 0.1 was adopted instead of the traditional 0.05 due to the fact that the sample size is rather small and not normally distributed. An alpha value above 0.1, however, becomes too lenient and not strict enough to weed out potential chance errors when using a smaller sample size.

5. RESULTS

5.1 Participants vs. Non-Participants

Figures 1a and 1b show the distribution of scores for each year, with each looking similar to a normal, bell-shaped curve, which is assumed when using parametric tests. A t-test performed to gauge how different the participants in the post-test were from the non-participants yielded non-significant results ($t = 0.5048$, $p = 0.61$). This test suggests that participant self-selection did not create two significantly different groups from within our population of interest.

5.2 Prolonged Learning

Several t-tests were performed to determine the amount of information retained from DLOC as it compared to the amount that was present at the termination of the course. The pre-test mean for 2004 was 20.48 (out of 25), and for 2005, was 20.59; while the post-test means for 2004 and 2005 were 17.93 and 18.23 respectively. These differences are statistically significant ($t2004 = 4.277$, $p = 0.003$; $t2005 = 5.517$, $p < 0.001$). See Figures 2a and 2b for all students' initial scores and post-test scores for 2004 and 2005 DLOC classes, respectively.

Chi-Square tests determined that 4 items were answered significantly worst in the post-test compared to the pre-test for the 2004 group, and that 7 items were answered significantly worst in the post-test compared to the pre-test for the 2005 group. An examination of the items revealed that some were not as efficient in discriminating between high achievers and low achievers on the test. For others, no specific reason for the difference was found. A possible reason for the disproportionate amount of significantly different responses in 2005 is that the sample size was larger, making it easier to become statistically different.

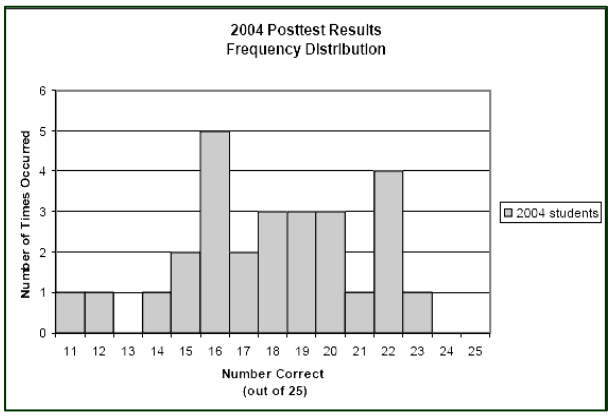

Figure 1a. Frequency distribution of combined 2004 student's post-test scores. Notice the near normal distribution (bell-shaped curve) centered on the mean, which is 17.93.

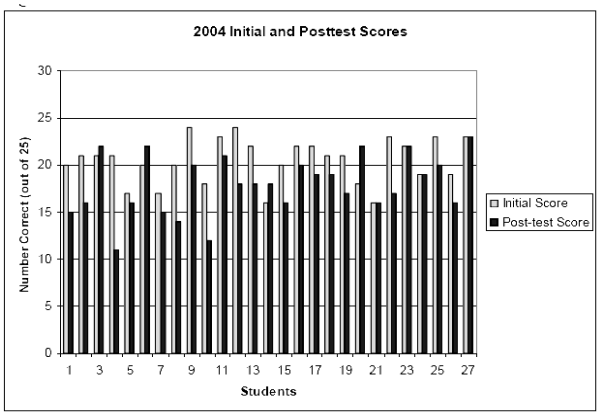

Figure 2a. Initial and post-test scores for all students in 2004 DLOC class. The initial mean score was 20.48 and the post-test mean score was 17.93 (out of a possible 25).

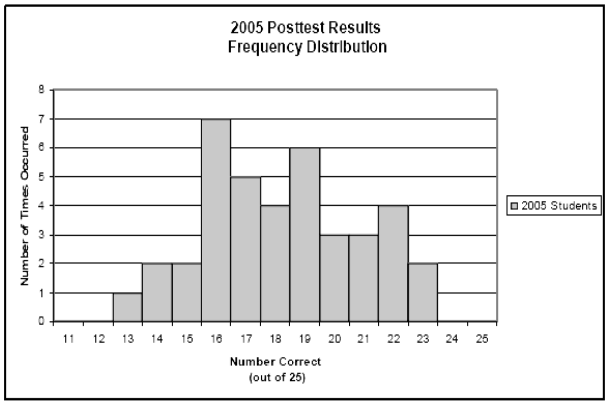

Figure 1b. Frequency distribution of combined 2005 student's post-test scores. Notice the near normal distribution (bell-shaped curve) centered on the mean, which is 18.23.

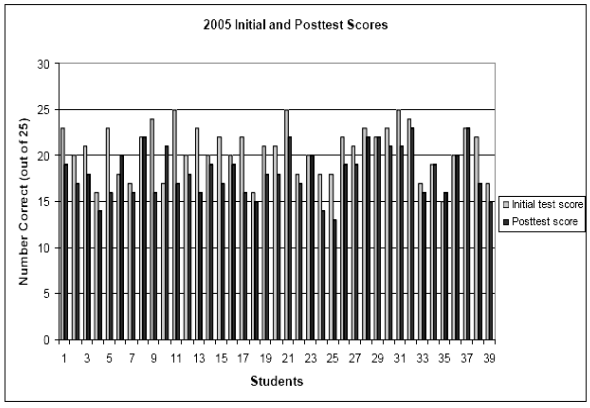

Figure 2b. Initial and post-test scores for all students in 2005 DLOC class. Mean initial score was 20.51 and the post-test mean score was 18.23 (out of a possible 25).

5.3 Level of Applicability

Students were asked to rate, on a scale from 1 to 9, with 1 being no application and 9 equating to "full" application, the level of application of DLOC instructional components. The average application of DLOC ICs for the 2004 class was 7.60, with a range of 1.5 to 9. Similarly, the 2005 class reported an average application of 7.18 with a range of 2 to 9.

The top 3 instructional components for reported application in both classes were (in order):

- Velocity Interpretation
- Base and Derived Products
- Convective Storm Structure and Evolution

There were also 2 specific questions on training application asked to DLOC graduates at the end of survey. The answers supported the ratings of reported instructional component applicability and offered several instances of specific applications.

Reported applications were divided into one of two categories, “low application” (1-6), and “high application” (7-9), and grouped into categories by NWS regions (figure 3). Although inspection of the application data reveals that the mean application rank for the region titled “Other” looks different from the other means, a Kruskal-Wallis test found no statistically significant difference between the applicability level and the participant’s region of residence ($p = 0.15$). It is important to note that the Kruskal-Wallis test is best used when there are at least 5 items per cell, which is not the case in this scenario. Therefore the results of the test must be interpreted cautiously. If the Kruskal-Wallis test is performed again without the “Other” group, the p value becomes 0.96, suggesting that there is no difference between regions and their reported applicability.

Region	Low Applicability (1-6)	High Applicability (7-9)
Eastern (1)	1	7
Central (2)	4	13
Southern (3)	3	11
Western (4)	4	11
Other (NHC, WDTB, etc.) (5)	4	1
Pac/AK (6)	1	3

(group number)

Krustal-Wallis: **p value = 0.146**
Group 1 n=8 Mean Rank=36.5625
Group 2 n=17 Mean Rank=33.0882
Group 3 n=14 Mean Rank=33.7500
Group 4 n=15 Mean Rank=32.1000
Group 5 n=5 Mean Rank=15.3000
Group 6 n=4 Mean Rank=32.6250

Figure 3. Distribution of reported application levels (“low” and “high”) of DLOC instructional components by region. Results of Kruskal-Wallis test for each regional group.

5.4 Correlations between Applicability and Long-term Learning

Correlations between the post-test scores and the reported application for each student are shown in figures 4a and 4b. Figure 4a shows a strong positive correlation between reported application and the resulting post-test score for 2004 DLOC class, with a statistically significant Pearson correlation coefficient ($r2004 = 0.503$, $p = 0.007$).

Interestingly, an apparent outlier was noticed among the data. To avoid reporting a false positive result due to the outlier, the correlation analysis was performed (graph not shown) without that person’s responses, and a statistically significant correlation was still found ($p = 0.016$).

In figure 4b, the correlation between the 2005 post-test scores and the reported application is not significantly different from zero ($r2005 = 0.044$, $p = 0.79$). When both classes are examined for correlation between post-test scores and application (figure 4c), the correlation is still slightly positive and significant ($r = 0.233$, $p = 0.06$).

6. DISCUSSION AND CONCLUSIONS

The goal of the first analysis of the data sample was to determine if the students who responded to the survey and took the test were different from those who did not participate. The fact that there is no statistical difference suggests that the two groups are similar and the subsequent results will not be tainted by a biased sample (such as only the “overachievers” participated in the study). In fact, the mean score on the initial test for those participating in the study was actually slightly lower than that for the non-participators (20.48 versus 20.66).

After establishing a representative sample, we can begin to make some conclusions based on the results. The first goal of the study was to determine the amount of information retained from the course. The second round of t-tests performed on the post-test scores of both years compared to the initial tests scores shows a statistical difference. Scores dropped slightly which means that something had changed over the amount of time since DLOC. Likely, some of the information learned during DLOC had been forgotten, which was hypothesized. The mean post-test scores for 2004 and 2005 DLOC were 17.93 and 18.23, respectively (this number represents the average number of correct responses out of 25). However, both classes still averaged > 70% on the post-test,

which is considered a “passing grade”. Thus, there appears that sufficient retention of the material learned in DLOC exists which could successfully change workplace behaviors to fit the desired standard.

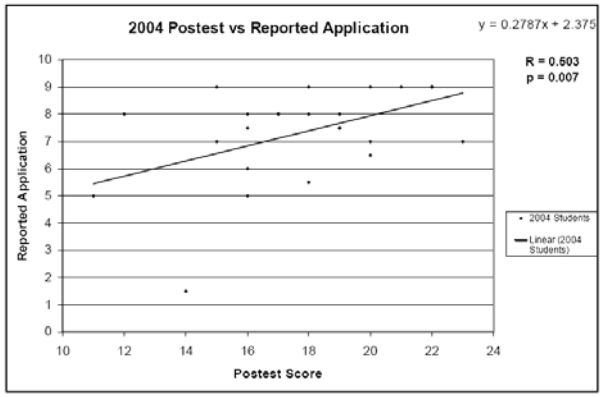

Figure 4a. 2004 DLOC post-test scores vs. reported application showing a positive correlation. P value of 0.007 means the relationship is very significant.

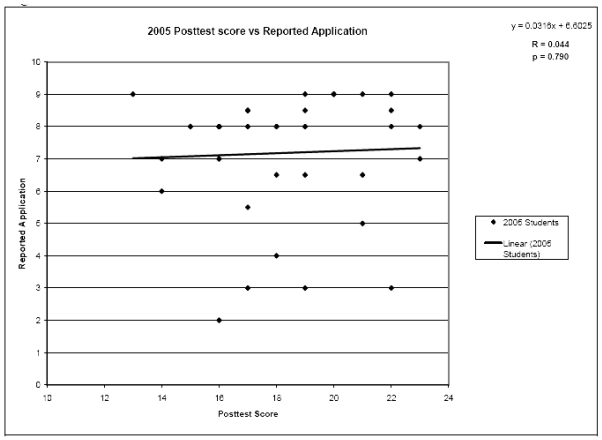

Figure 4b. 2005 DLOC post-test scores vs. reported application showing near zero correlation. P value of 0.790 means the relationship is not significant

Was the information retained due to effective training, workplace application, or a combination of both? The correlations in figure 4 show that there is indeed a statistically significant relationship between application and the score on the post-test.

Breaking the data up into separate classes, however, shows different results for each year. The students who took DLOC in 2004 depend more heavily on application to recall DLOC techniques than do their 2005 counterparts.

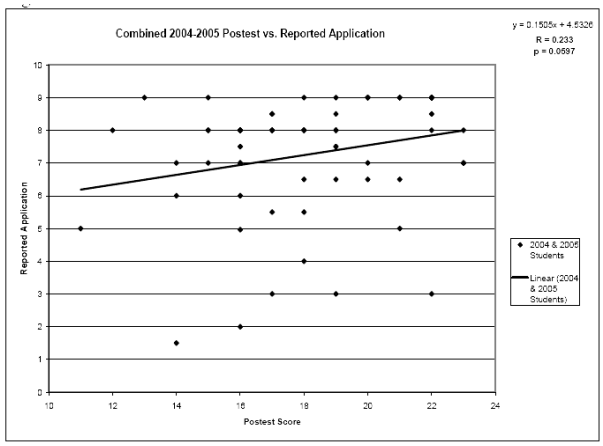

Figure 4c. Combined 2004-2005 DLOC post-test scores vs. reported application. Correlation is slightly positive and the P value of 0.0597 means the relationship is significant.

Some reasons for this include the fact that the 2005 class more recently completed the training and can possibly still recall the instructional concepts clearly, while the 2004 group only recalls what they have been able to apply on a regular basis. Also, the 2004 group has had more of an opportunity to apply the DLOC objectives in the workplace than have the 2005 group, leading to a stronger relationship between application and retention.

From these results it appears that the retention of the material comes from both effective training and regular application. In the short term, before the material has had sufficient time to be applied, it can still be recalled thanks to effective training techniques. In the long term, however, a stronger relationship occurs between retained knowledge and application since the material not applied begins to fade from memory.

There is not, however, a difference in reported application by region as was hypothesized. There is, on the other hand, a difference between the application for those in a region and those who have moved to a national office or some other NWS branch (Group 5 *Other* in figure 3). With the low application for this group, it would make sense for the NWS to limit the participants in DLOC to

forecasters from NWS Forecast Offices (WFOs), and exclude students from national centers, River Forecast Centers and other non-WFOs facilities. This recommendation will help the NWS become more cost-effective and eventually ensure a better return on training investments.

In order to solidify these claims, more research needs to be done. A study which has future implications would be to show whether or not time influences the dependence on application to retain course knowledge. Also, a larger sample size would give clearer, more statistically significant results. Perhaps this research can be repeated when a level 3 evaluation study is conducted on DLOC which would provide a more thorough understanding of the relationship between prolonged learning and application.

7. REFERENCES

Kirkpatrick, D. L, 1994: *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.

_____ and J. Kirkpatrick, 2005: *Transferring learning to behavior: Using the four levels to improve performance.* San Francisco: Barrett-Koehler.

Warning Decision Training Branch, 2005: The home page is www.wdtb.noaa.gov.

APPENDIX A

DLOC Post-Training Evaluation Survey/Quiz

Part I. Logistical questions

1. Last Name, First Name:

2. Current Office:

3. Year enrolled in DLOC: 2004/2005

Part II. Opportunities to apply DLOC objectives

Specific Instructions:
Please indicate the degree to which you have applied the following instructional components of the Distance Learning Operations Course (DLOC) in your current position. In this scale, "1" implies that you have not applied any of the DLOC training. On the other hand, "9" implies full application of the DLOC training in your current position. Numbers in between the scale refers to linear gradations between the two extremes.

DLOC Instructional Component	**Rating Scale**
Radar Applications using AWIPS	1 2 3 4 5 6 7 8 9
Introduction to WSR-88D	1 2 3 4 5 6 7 8 9
Principles of Met. Doppler Radar	1 2 3 4 5 6 7 8 9
Velocity Interpretation	1 2 3 4 5 6 7 8 9
Base and Derived Products	1 2 3 4 5 6 7 8 9
System Operations and Control	1 2 3 4 5 6 7 8 9
Convect.Storm Structure and Evol.	1 2 3 4 5 6 7 8 9
DLOC In-Residence Workshop	1 2 3 4 5 6 7 8 9

Specific Instructions:
Answer each of the following questions as completely and detailed as possible.

1. What specific topic from the DLOC training have you applied the most in your current position? Provide an example to illustrate your point.

2. What specific topic from the DLOC training have you never applied in your current position? Describe the reason why.