School Confidential
Program X Evaluation Report

Executive Summary

Seven Program X Training programs from seven different delivery locations (N = 358) are evaluated in this report. The locations include [redacted].

Data from the assessment instrument were analyzed to provide participant performance results for the pretest and post-test with an emphasis on gains in performance. Psychometric analyses were conducted on the items of the assessment instrument as well so that specific recommendations could be made regarding the modification of the instrument. Qualitative and quantitative analyses were conducted on the course evaluation data in order to evaluate the effectiveness of the program to deliver pertinent information, as defined by the learning objectives. Levels of participant satisfaction were also evaluated. An emphasis was placed on the course content, course facilitator and course materials.

The results from the Program X program evaluation indicate that the training significantly increased participant knowledge and competency. Participants' performance improved statistically significantly from the pretest to the post-test ($p < .01$) across all seven delivery sites with an average post-test performance of 84%. Furthermore, participants with lower pretest scores evidenced the most growth indicating that the program was effective in fostering growth most in those that required it. Finally, participants indicated that the knowledge gained will help them in their future roles.

The course evaluation data indicates that participants, in general, provided positive ratings with regard to the course content, course facilitator and course material.

Also, participants tended to perceive the course delivery as effective and found that the program equipped them with valuable skills that will help them be competent Field Investigators.

Based on the psychometric analyses, it is evident that there is a wide range of difficulty associated with the assessment instrument as well as a wide range in the ability of the items to discriminate between low and high performers. Question 19 and 27 appear to be poor questions and therefore should be re-written or eliminated from the exam due to the fact that participant performance was very low on those two items and they do not have a high ability to discriminate between low and high performers. Also, Questions 3, 8 and 29 due not allow much room for growth and therefore may be too easy, which suggests that they should be eliminated as well. However, the remaining questions appear to be appropriate and well written.

According to the confirmatory factor analysis, most of the items appear to be internally consistent due to the fact that the model fit the data well and 27 of the 30 questions had significant loadings on their corresponding factor. However, more data needs to be collected to ensure the reliability and validity of the factor analysis results.

School Confidential

Program X Program Evaluation Report

The goal of this paper is to evaluate the Program X with regard to student performance over time, course delivery, and the quality of the course evaluation and the assessment instrument.

Student performance will be analyzed so that knowledge gained over time may be assessed. This analysis will provide information regarding the effectiveness of the program to deliver relevant course information to its participants. Therefore, student performance will be one of the inputs to the process of evaluating the quality of the program.

In addition to student performance, course evaluations will also be analyzed with regard to the structure and content of the form as well as the information obtained regarding the course. Therefore, course evaluations will be a second input for evaluating the quality of the program.

Finally, the psychometric characteristics of the assessment instrument will be analyzed and discussed. Once all of the data are analyzed and presented, a summary section with specific recommendations will follow in the attempt to improve the program delivery, the course evaluation form, the assessment instrument and participant performance (see Figure 1).

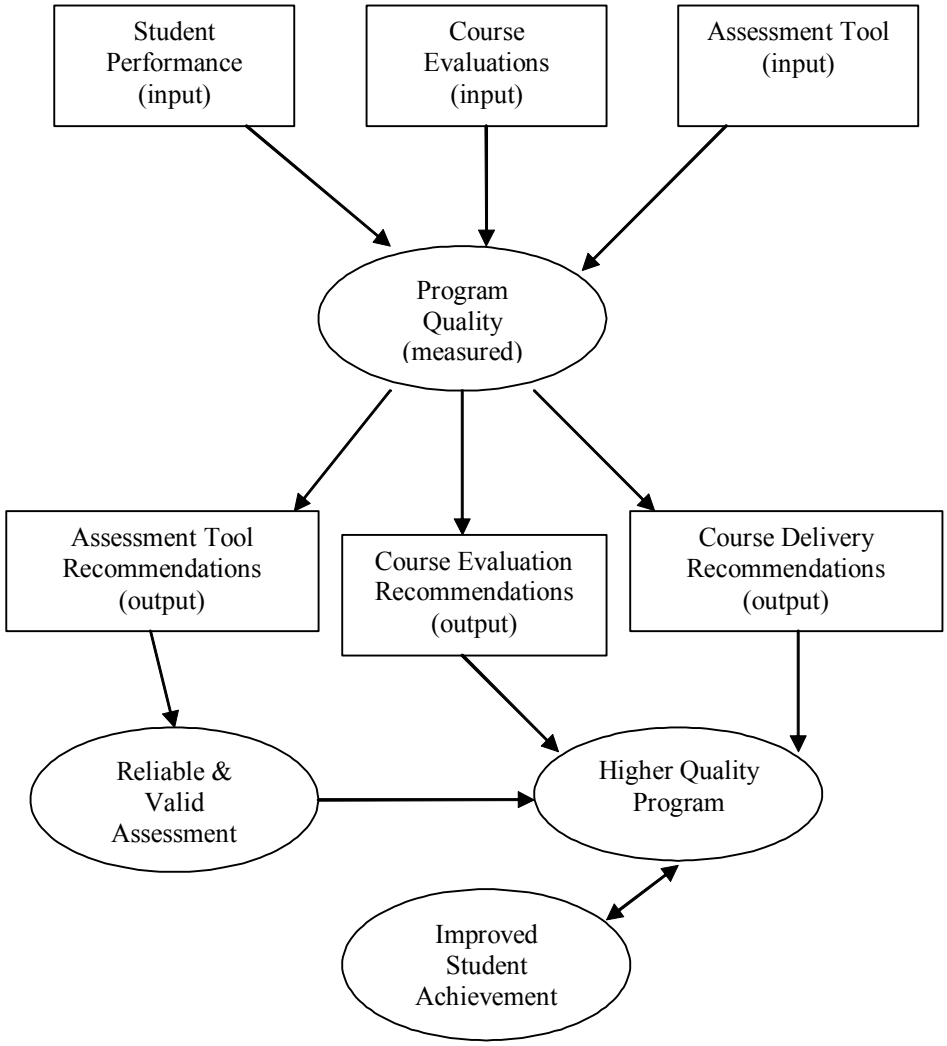

Figure 1. Program Evaluation Inputs & Outputs.

The first part of this paper pertains to student performance. Pretest and post-test data will be analyzed and presented with an emphasis on change in performance over time. The second section of this paper pertains to the course evaluation data. The third section contains the psychometric properties of the assessment instrument. The fourth and final section covers the integration of the three sources of data in an attempt to

address the current program strengths and provide specific recommendations for improvements where weaknesses exist. Therefore, the final product will be a working model for delivering a high quality program.

Background Information

Program X was provided to 358 participants via seven different delivery locations. [redacted] All results will be presented aggregating across the delivery sites due to sample size and reporting efficiency issues. However, specific mention of differences due to delivery site will be presented and discussed when appropriate.

Limitations

Some of the limitations associated with this program evaluation include the fact that many of the participants did not complete both the pretest and the post-test which limits the amount of people that can be included in the knowledge gained analysis, the small sample size relative to the number of items to be assessed which affects the ability to conduct robust psychometric analyses, and the lack of demographic information on the participants. Having the demographic information would help determine whether or not particular background characteristics are associated with knowledge gained and perceptions of the program.

Data Collection and Instrumentation

Pretest and post-test data were collected via the Program X assessment instrument, which consists of 30 multiple choice questions (see Appendix). The test was designed to measure five learning objectives which fall into three distinct modules. The

pretest was designed to serve as a baseline measure in which the post-test could be compared. Gain scores were then calculated for those participants that had both pretest and post-test data.

Course evaluation data was also collected from participants at the completion of the program. The data obtained from the course evaluations were both quantitative and qualitative. The course evaluations were intended to measure students' perceptions of the course delivery, materials and content as well as the program facilitator.

Data Analysis and Results

Student Performance

Three hundred and fifty-eight people participated in this program. However, 295 people took the pretest, 261 took the post-test and only 199 people took both the pretest and the post-test. The table below provides the descriptive statistics for the pretest, post-test and change scores.

Table 1

Student Performance Descriptive Statistics: All Participants

	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Pretest	295	.33	.93	.75	.10	-1.02	1.56
Post-test	261	.47	1.00	.84	.09	-1.26	2.18
Change Amount	199	-.40	.60	.09	.12	-.12	3.44

The table above indicates that participants scored an average of 75% on the pretest and 84% on the post-test. The average improvement from the pretest to the post-test was nine percentage points. The skewness and kurtosis statistics indicate that the distributions of scores are negatively skewed and highly peaked with many people

achieving similar scores, especially with regard to their change scores. The histograms below provide a visual account of the participants' performances.

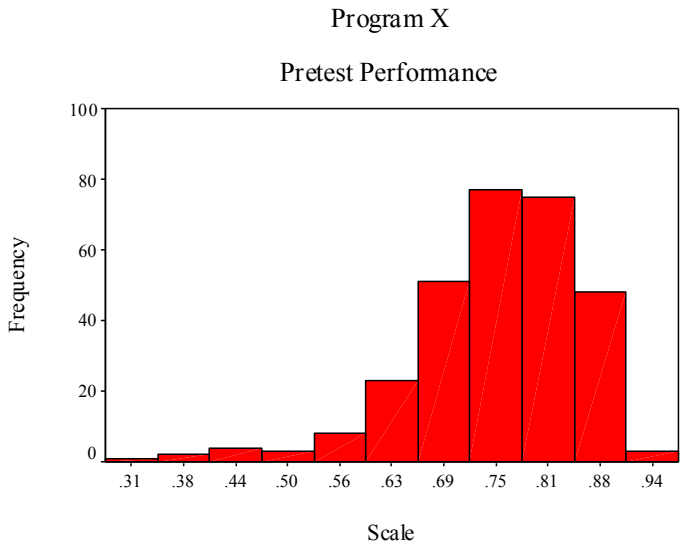

Figure 2. Program X Pretest Performance.

The figure above indicates that most of the participants scored above 50% on the pretest with the bulk of scores ranging between 67% and 91%.

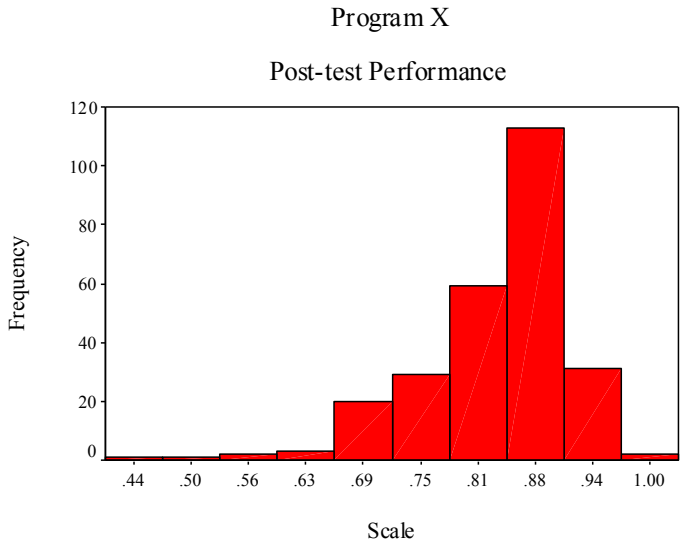

Figure 3. Program X Post-test Performance.

The figure above indicates that although the distribution is still negatively skewed for the post-test, the scores have shifted upward on the scale. Also, the distribution is very peaked with several people between 84% and 91%.

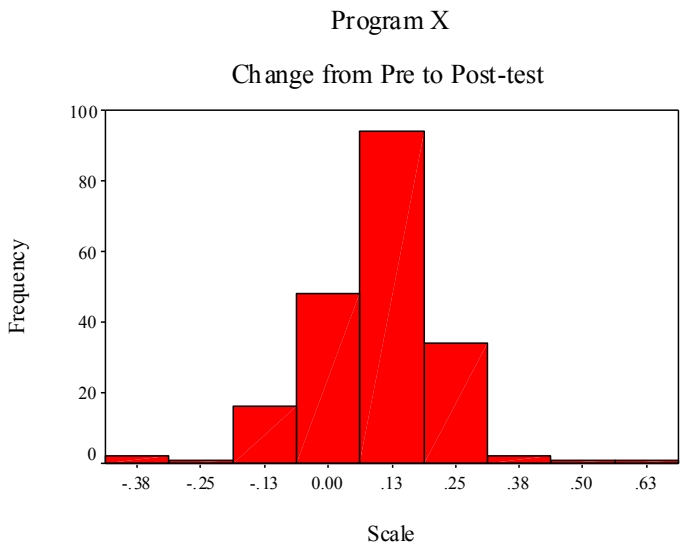

Figure 4. Program X Performance Change from Pretest to Post-test.

The histogram above is fairly normal with a few extreme values on both sides of the distribution. The largest decrease in performance was 40% while the largest increase was 60%. The majority of the sample improved from the pretest to the post-test with many participants improving between 6% and 19%.

In order to determine how the participants performed on average for each question on the pretest versus the post-test, a chart comparing the mean performance for each question by test (pretest versus post-test) was constructed. Please refer to Figure 5.

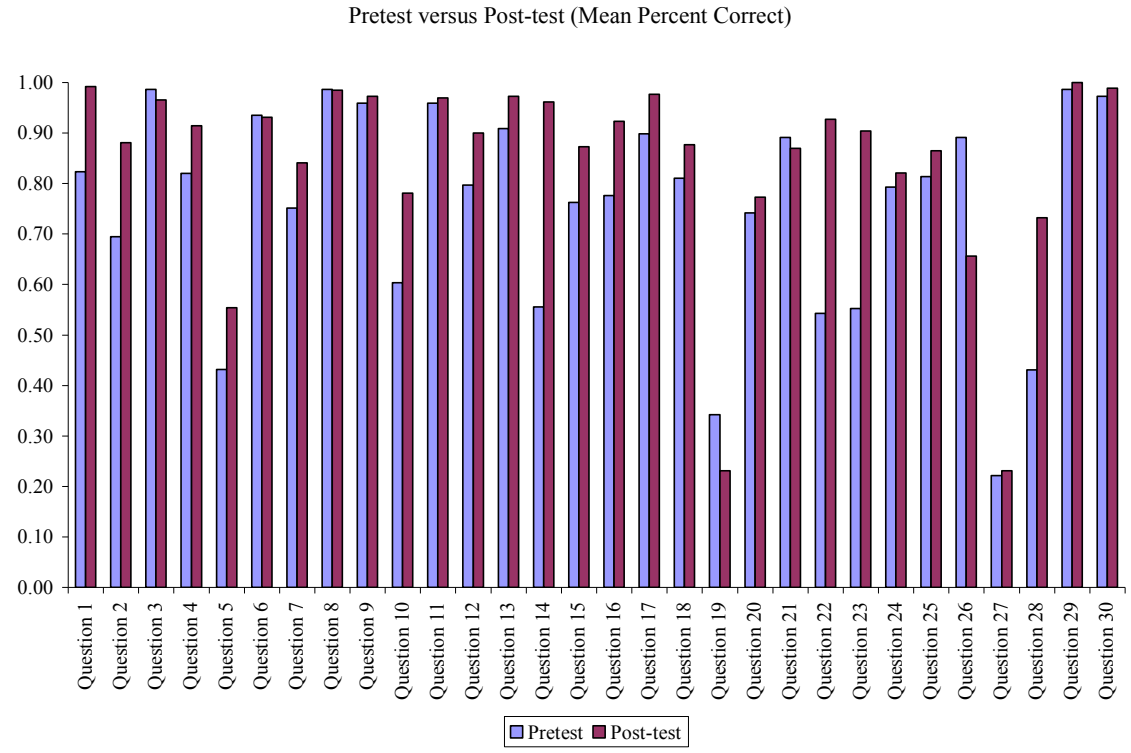

Figure 5. Pretest versus Post-test by Question: Includes all Test Takers.

Figure 5 shows that none of the items were answered correctly by all of the participants on the pretest, which indicates that there was at least some room for growth on all items. However, performance on Questions 3, 8 and 29 was very high suggesting that the items may be too easy. Also, Question 29 was the only item that everyone answered correctly on the post-test when the entire sample is included.

Questions 19 and 27 yielded the worst performance on the post-test with only 23% answering the question correctly. However, participants did poorly (under 70%) on Question 5 (55%), and Question 26 (66%) as well. Due to the fact that so many participants answered these four questions incorrectly, the responses were examined to determine if a pattern exists.

Tables 2 through 5 provide the response frequencies. Responses categorized as "Other" indicate that the participant chose more than one option, which was scored as an incorrect response. Responses coded as missing indicate that the participant did not answer the question.

Question 5 asks participants to fill in the blank and states, "The occurrence of a disease or of all disease in a population is called". The correct answer is "Prevalence" which is option "D". Fifty-five percent of the participants correctly chose option "D". However, almost 30% incorrectly chose "Morbidity", which is option "B". None of the participants incorrectly chose "Surveillance" which is option "E". Table 2 provides the response frequencies.

Table 2

Frequency of Responses by Option: Question 5

		Frequency	Percent	Valid Percent	Cumulative Percent
Response	Other	1	.38	.39	.39
	A	22	8.43	8.53	8.91
	B	77	29.50	29.84	38.76
	C	15	5.75	5.81	44.57
	D	143	54.79	55.43	100.00
	Total	258	98.85	100.00	
	Missing	3	1.15		
Total		261	100.00		

Question 19 asks, "Which of the following is an example of the use of probing to obtain more information without influencing the interviewee's response?" The correct answer is option "A" which asks "Could you describe how you are feeling?" Only 23% of the participants correctly chose option "A"; however, 33% percent incorrectly chose

option “D” which asks, “What specifically about your symptoms concern you?” Table 3 provides the response frequencies.

Table 3

Frequency of Responses by Option: Question 19

		Frequency	Percent	Valid Percent	Cumulative Percent
Response	A	60	22.99	23.08	23.08
	B	54	20.69	20.77	43.85
	C	28	10.73	10.77	54.62
	D	79	30.27	30.38	85.00
	E	39	14.94	15.00	100.00
	Total	260	99.62	100.00	
	Missing	1	.38		
Total		261	100.00		

Question 26 states, “One of the people you interview is an older male who does not provide clear answers to your questions. Instead of responding directly to your questions, this person goes ‘off-track’ and gives answers that don’t specifically address your question. For the strategies listed below, which one would most effectively get the responses you need to conduct your interview?” The correct response is “Repeat the question” which is option “D”. Sixty-six percent of those who answered the question correctly chose option “D”. Twenty-seven percent incorrectly chose option “E” which states, “All of the above would be effective.” No one incorrectly chose option “C” which states, “Increase the distance between yourself and the interviewee.” Table 4 provides the response frequencies.

Table 4

Frequency of Responses by Option: Question 26

		Frequency	Percent	Valid Percent	Cumulative Percent
Response	A	3	1.15	1.16	1.16
	B	15	5.75	5.79	6.95
	D	170	65.13	65.64	72.59
	E	71	27.20	27.41	100.00
	Total	259	99.23	100.00	
	Missing	2	.77		
Total		261	100.00		

Question 27 states, "You are asked to conduct an interview with a person who is in the hospital. When you arrive, the patient's room is filled with friends and family. In order to minimize the effects of the environment and avoid interview bias, you decide to:". The correct response is option "A" which states, "Request that everyone leave so you can conduct the interview without any outside interference." Only 23% of the participants correctly chose option "A". However, 64% incorrectly chose option "B" which states, "Reschedule the interview when 'Visiting Hours' is over." No one incorrectly chose option "E" which states, "Skip questions whose answers might embarrass the person in front of their friends and family." Table 5 provides the response frequencies.

Table 5

Frequency of Responses by Option: Question 27

		Frequency	Percent	Valid Percent	Cumulative Percent
Response	A	60	22.99	23.08	23.08
	B	166	63.60	63.85	86.92
	C	33	12.64	12.69	99.62
	D	1	.38	.38	100.00
	Total	260	99.62	100.00	
	Missing	1	.38		
Total		261	100.00		

In addition to the summary statistics conducted based on participant performance for the entire sample, summary statistics were also conducted for only those who had both pretest and post-test data. The results for this sample therefore contain a data point for each participant at both points in time (pretest and post-test). The descriptive statistics are presented in Table 6.

Table 6

Student Performance Descriptive Statistics: Restricted Sample

	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Pretest	199	.33	.93	.76	.10	-.97	2.04
Post-test	199	.50	1.00	.85	.08	-1.31	2.63
Change Amount	199	-.40	.60	.09	.12	-.12	3.44

The table above indicates that the average performance is not much different for the restricted sample as it was for the entire sample. Both the pretest and post-test average performance is one percentage point higher than when the entire sample is included (76% versus 75% and 85% versus 84%). The distributions are also similar with a negative skew and high peak.

Histograms representing the distributions for the pretest and post-test, based on the restricted sample, are provided below (see Figures 6 and 7). Please refer back to Figure 4 for the histogram that reflects the change in performance from the pretest to the post-test.

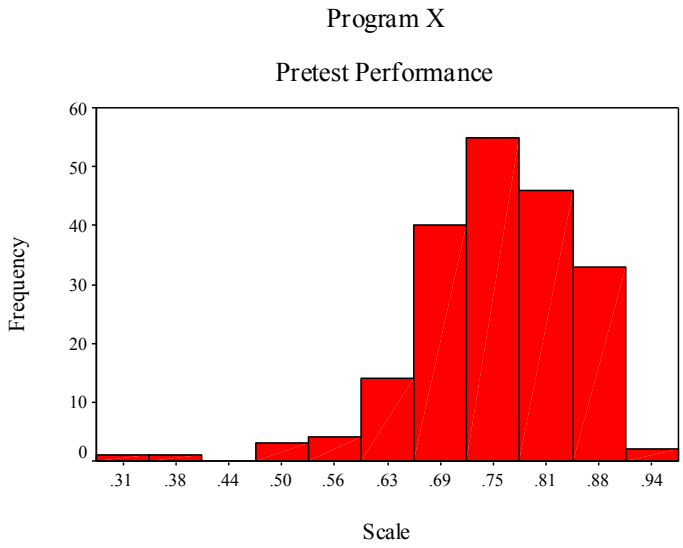

Figure 6. Program X Pretest Performance: Restricted Sample.

The figure above indicates that there are a few extreme scores on the lower end of the distribution although most of the participants scored above 70%. Many of the participants yielded similar scores with the most common range being between 72% and 78%.

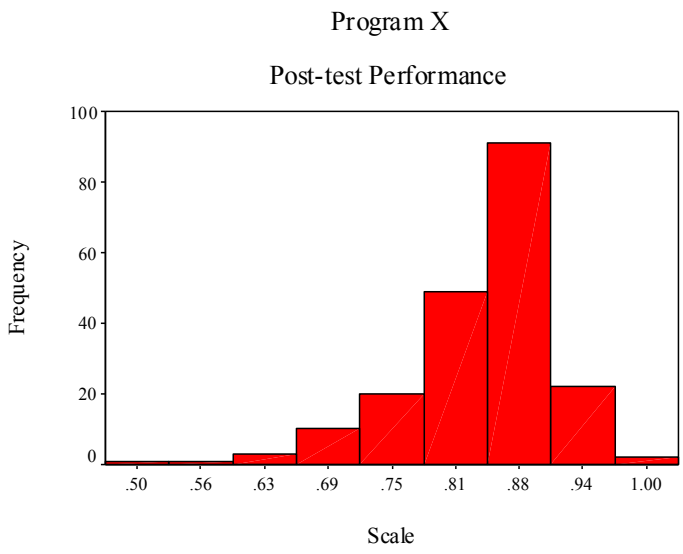

Figure 7. Program X Post-test Performance: Restricted Sample.

According to the histogram above there are still a few extreme scores, although they are not as low as that seen on the pretest. The most common score range was between 85% and 91%

In order to determine the pattern of growth from the pretest to the post-test, a growth chart was constructed. A growth chart is a scatter plot that correlates the participants' standard scores (*z* scores) on the pretest with their change scores. This scatter plot helps to determine whether or not those with the most room for growth actually yielded the most growth. This is especially important for cases in which some of the participants scored high on the pretest. Due to the fact that they already started out high, the expected growth is low for those participants. An effective program is one that fosters growth most in those that require it.

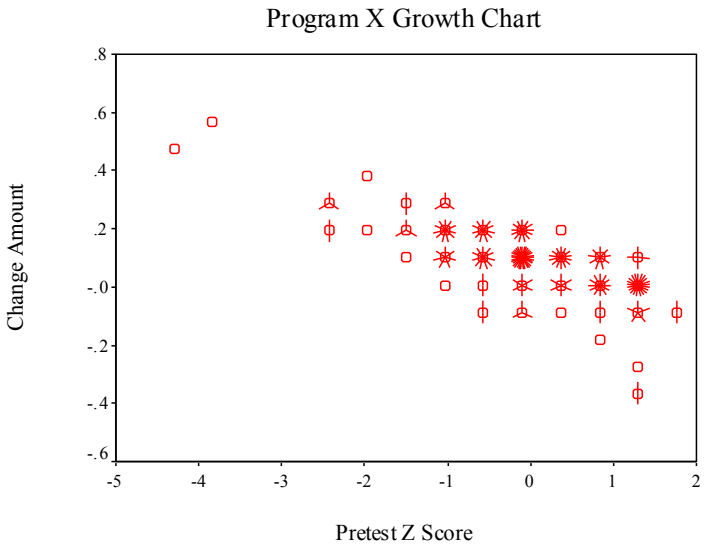

Figure 8. Program X Growth Chart.

The growth chart above indicates that those with the lowest pretest scores evidenced the most growth, on average. In fact, the two participants that scored more than four standard deviations below the average improved by more than 40 percentage points.

Finally, the average percent correct for each question on the pretest versus the post-test was determined for the restricted sample containing both data points. Please refer to Figure 9.

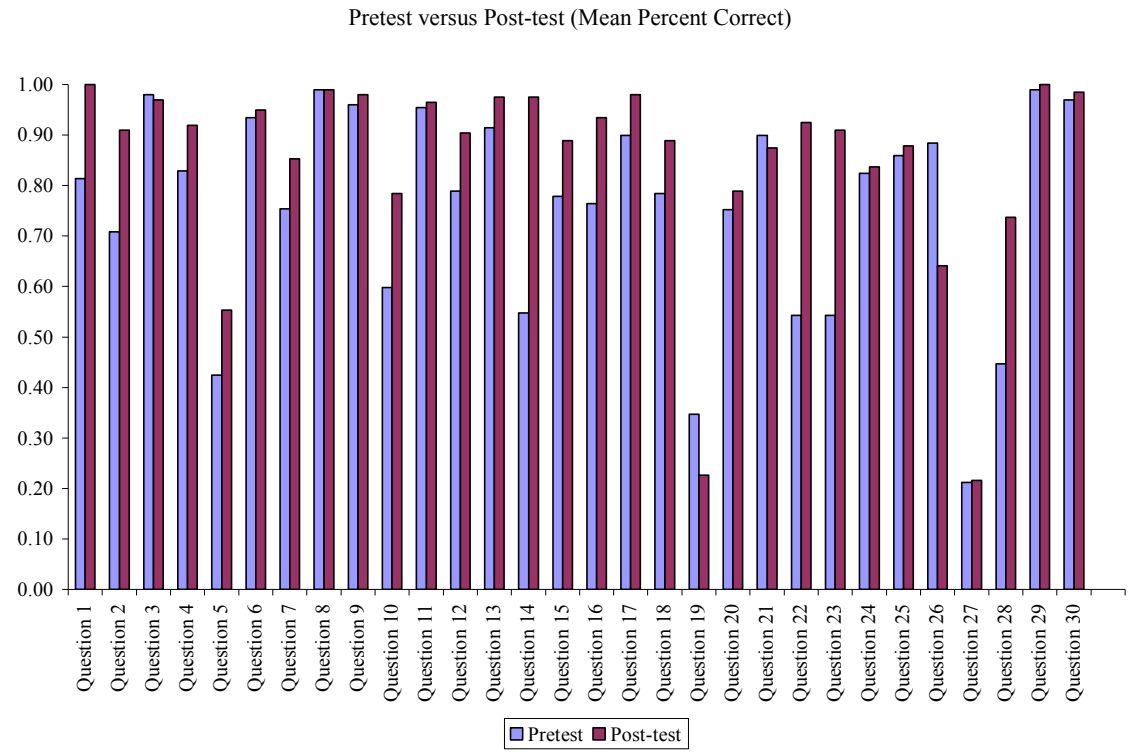

Figure 9. Pretest versus Post-test by Question: Restricted Sample.

The figure above indicates that participants went down on Question 3 (98% versus 97%), Question 19 (35% versus 23), Question 21 (90% versus 87%) and Question 26 (88% versus 64%). The largest gains were seen for Question 14 (55% versus 97%), Question 22 (54% versus 92%), Question 23 (54% versus 91%) and Question 28 (45% versus 74%). Several items yielded growth in the double digits.

In addition to the descriptive statistics, a repeated measures analysis of variance (ANOVA) was conducted to test the significance level of the participants' change in performance over time. The results indicate that the participants' change in performance is statistically significant (see Table 7 below).

Table 7

Repeated Measures Analysis of Variance

Measure: TIME

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
TEST	.79	1.00	.79	110.01	.00
Error(TEST)	1.43	198.00	.01		

According to Table 7, on average, the students improved significantly from the pretest to the post-test [$F(1,198) = 110.01$, $p < .01$]. Figure 10 illustrates this change.

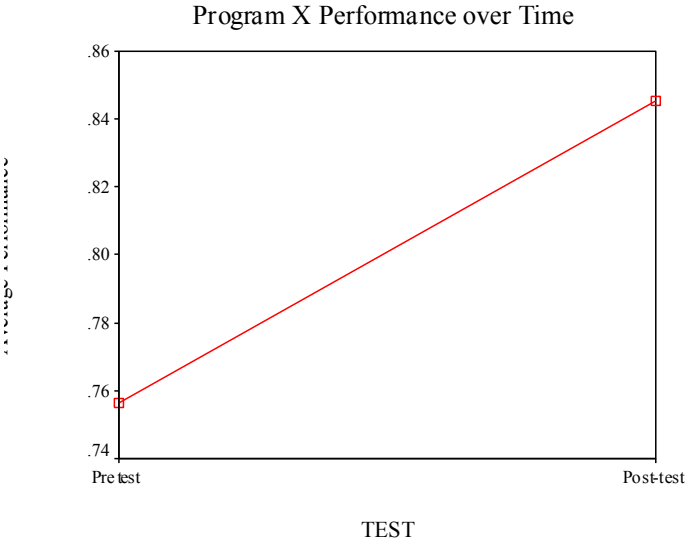

Figure 10. Participant Performance over Time.

In addition to the whole group aggregated across the delivery sites, Table 8 provides the descriptive statistics by delivery site and Figure 11 shows the average change in performance by delivery site. However, due to small cell sizes, the repeated measures ANOVA is not presented.

Table 8

Descriptive Statistics by Delivery Site: Restricted Sample

		N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Delivery Site		Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
XXX	Pretest	37	.53	.87	.73	.08	-.09	-.24
	Post-test	37	.60	1.00	.83	.09	-.85	.68
	Change Amount	37	-.10	.37	.09	.11	.08	-.13
XXX	Pretest	31	.53	.90	.79	.11	-.79	-.52
	Post-test	31	.50	.93	.84	.10	-2.25	5.22
	Change Amount	31	-.40	.33	.05	.16	-.94	2.28
XXX	Pretest	31	.57	.87	.76	.07	-.83	.52
	Post-test	31	.67	.97	.85	.07	-.95	.53
	Change Amount	31	-.17	.20	.08	.09	-.73	.32
XXX	Pretest	31	.50	.93	.73	.10	-.79	.74
	Post-test	31	.70	.93	.84	.07	-.72	-.48
	Change Amount	31	-.13	.27	.10	.11	-.27	-.36
XXX	Pretest	25	.60	.90	.78	.08	-.41	-.16
	Post-test	25	.60	.97	.84	.08	-.95	1.76
	Change Amount	25	-.10	.17	.06	.08	-.33	-1.07
XXX	Pretest	21	.60	.87	.77	.08	-.38	-.93
	Post-test	21	.80	1.00	.89	.05	.46	.27
	Change Amount	21	.00	.30	.13	.08	.44	-.10
XXX	Pretest	23	.33	.93	.74	.14	-1.77	3.89
	Post-test	23	.67	.97	.86	.08	-.65	.41
	Change Amount	23	-.23	.60	.12	.18	.92	2.59

The table above indicates that all of the delivery sites had average increases in performance from the pretest to the post-test. XXX yielded the most growth (13 percentage points) while XXX yielded the least growth (five percentage points). Overall, XXX had the best performance on the post-test (89%) and XXX had the lowest performance (83%). However, all delivery sites had average scores above 80%, which indicates that the participants did fairly well regardless of the delivery site. See Figure 11.

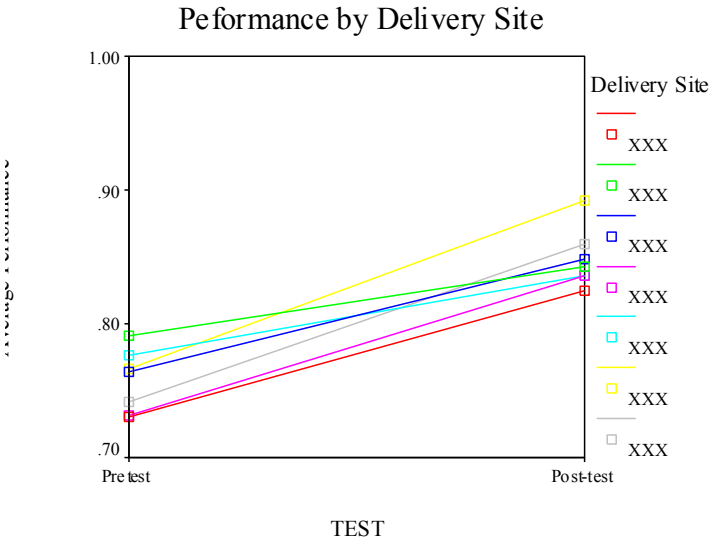

Figure 11. Participant Performance over Time by Cohort.

Figure 11 provides the same information as Table 8 but in a more reader friendly format. It provides a clear picture of the differences and similarities between the delivery sites with regard to average improvement over time.

In summary, the performance data indicate that the participants did fairly well, although they had trouble with some of the questions. Participant performance significantly increased from the pretest to the post-test and all seven delivery sites showed improvement. Furthermore, those that scored lowest on the pretest showed the most growth, on average. Finally, the results of the participants' performance indicate that attention should be paid to some of the questions on the assessment instrument due to extreme item difficulty, decreases in performance and lack of room for growth. This will be discussed in further detail in the assessment instrument section and well as in the summary and recommendations section.

Course Evaluation Form

Two hundred and sixty four participants completed the course evaluation. The course evaluation was structured so that the first 19 questions were based on a four-point likert scale and were intended to evaluate specific aspects of the course. The responses were coded such that "strongly agree" received a value of one, "agree" received a value of two, "disagree" received a value of three and "strongly disagree" received a value of four. All questions were phrased positively so that lower values are consistently associated with positive responses.

Questions 20 through 23 related to the overall quality of the course and were based on a five-point scale ranging from very poor to excellent. However, for these questions the lower the value the more negative the response. The responses were coded such that a response of "very poor" was given a value of one, "poor" was given a value of two, "satisfactory" was given a value of three, "good" was given a value of four and "excellent" was given a value of five.

The 23 quantitative questions mentioned above were broken down into three dimensions based on the nature of the question. Therefore, the three dimensions that were considered when assessing the course evaluations include:

- Dimension 1 - Course Content (Questions 1-12,17-19)
- Dimension 2 - Course Facilitators (Questions 13-16)
- Dimension 3 - Course Materials (Questions 20-23)

The following sections will discus the above-mentioned dimensions and their scores. The scores will indicate the programs performance and specific objectives within each

dimension will be discussed. Table 9 below will be referenced throughout the section to complement the histograms.

Table 9

Course Evaluation Descriptive Statistics by Dimension

	N	Minimum	Maximum	Mean	SD
Course Content	264	1.00	4.00	1.52	.46
Course Facilitators	264	1.00	4.00	1.33	.49
Course Materials	264	2.75	5.00	4.26	.55

The Course Content dimension yielded a mean score of 1.52, which indicates that participants agree to strongly agree that the course content was presented in a manner consistent with their expectations. The histogram below shows the distribution of scores on the Course Content dimension. The distribution has a positive skew and is heavily weighted on the left side or the low scoring side of the Likert scale. Also, there were a few extreme values on the right side of the distribution.

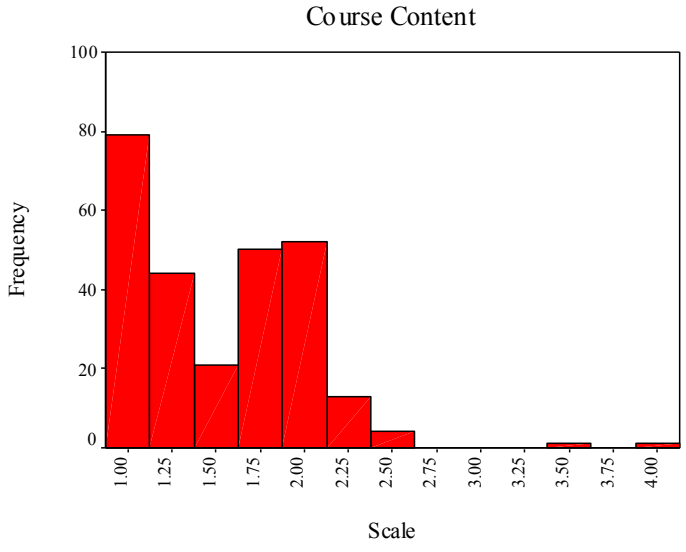

Figure 12. Course Content Dimension Histogram.

In addition to the quantitative data, the participants' responses were analyzed for themes. The following themes emerged for the participants who responded either "agree" or "strongly agree" to the Course Content dimension questions:

- Course objectives were clearly stated.
- The course met their needs.
- They are satisfied with the quality of the course.
- The knowledge gained will help them in their future roles.
- The activities and exercises were relevant to the course.
- They would recommend the course to a colleague.

Furthermore, some of the participants responded that as a result of this course they have a better understanding of epidemiology and the investigation process. Respondents learned and will implement the 10 steps of the epidemiology process when investigating an outbreak. Respondents also commented that as a result of the course they will "Show more diligence when evaluating regarding food inspections as far as food prep and storage." There were only a few respondents that rated the course poorly. However, these respondents did not provide qualitative information regarding what they found unsatisfactory about the program. In conclusion, based on the quantitative and qualitative results, overall Program X has effectively provided pertinent information to the participants and has therefore provided them with the skills needed to become effective Field Investigators.

The Course Facilitator dimension yielded a mean score of 1.33, which indicates that participants agree to strongly agree that the course facilitator was effective in delivering the course content and relating to the participants.

The histogram below shows the distribution of scores on the Course Facilitator dimension. The distribution is similar to the Course Content dimension in that it has a positive skew and is heavily weighted on the left side or the low side of the Likert scale. One individual provided a very negative rating; however, the vast majority provided highly positive ratings, which confirms that the facilitators have proficient qualities.

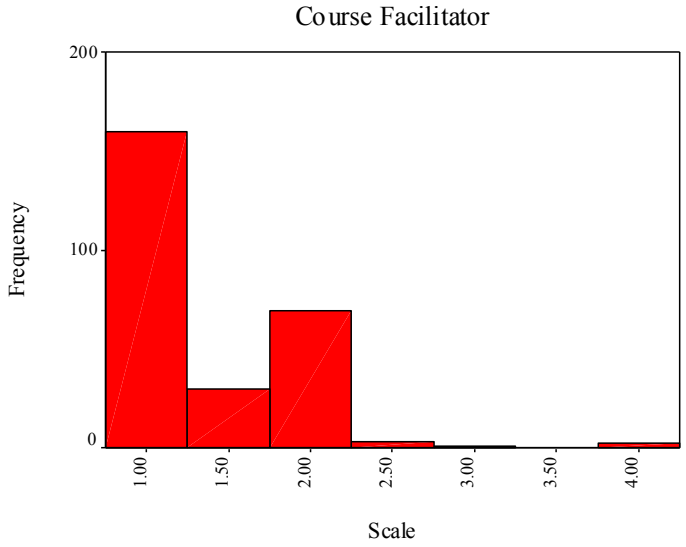

Figure 13. Course Facilitator Dimension Histogram.

The qualitative data indicated that respondents thought that the facilitators were professional, and that they have an excellent knowledge base which was evident when they communicated the material. One respondent mentioned that a strong point of the program was the "Instructor knowledge and delivery and the diversity of teaching styles." Therefore, both the quantitative and qualitative data confirm that the facilitators were effective, competent and professional.

The Course Materials dimension was rated quite high on average, with a mean rating of 4.26. The distribution has a negative skew and is heavily weighted on the right

side or the high scoring side of the Likert scale. Also, the entire scale was not used. This indicates that all participants believed that the course materials were satisfactory or better. The distribution of scores is provided in Figure 14 below.

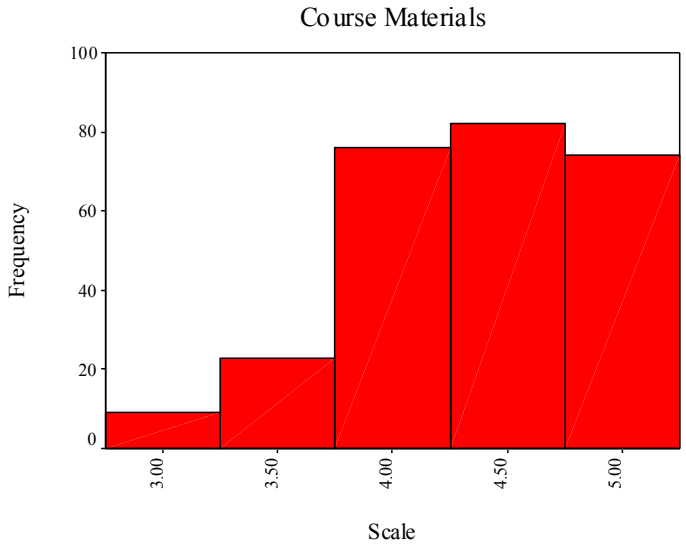

Figure 14. Course Materials Dimension Histogram.

The histogram's negative skew indicates that the course materials were found to be good or excellent in the majority of cases. Therefore, almost every respondent found the meeting space, visual aids, course materials and online training component satisfactory or higher. Also, additional analyses indicated that the bulk of satisfactory responses were in response to question 21 which asked about the meeting space component. The qualitative information suggests that in some cases respondents did not like moving around the room frequently and when movement was necessary, they felt their movement was somewhat inhibited. Therefore, further thought should be given to the room layout as well as providing adequate seating so that participants have enough personal space to feel comfortable.

According to the qualitative data, some of the themes that emerged include the following:

- The workbook is loaded with useful materials.
- Great print materials.
- The Jeopardy and Food Bourne Activity was a strong point.
- Hand-outs for future use and group breakouts were strong points.
- The materials were complete and re-usable.

As with the Course Content dimension and the Course Facilitator dimension, the quantitative and qualitative data are consistent. Both sources of data indicate that the course materials are good to excellent, on average.

In order to determine participants' preferences with regard to delivery format, Question 24 asked participants to determine the type of training delivery format that best fit their needs. Table 10 below provides the results.

Table 10

Response Patterns for Question 24

	Frequency	Percent	Cumulative Percent
On-site	114	44.9	44.9
Distance Learning	16	6.3	51.2
Combination	123	48.4	99.6
None	1	.4	100.0
Total	254	100.0	

The table above indicates that about 45% of the participants felt that the on-site training delivery format best fit their needs. Almost 6% indicated that the distance

learning delivery format was best for them. Finally, almost 49% felt that a combination of the two formats best fit his/her needs.

Question 25 asked participants which distance learning delivery format best fit their needs. The results are presented in Table 11 below.

Table 11

Response Patterns for Question 25

	Frequency	Percent	Cumulative Percent
Satellite Video Teleconference	59	33.7	33.7
Video	9	5.1	38.9
DVD	16	9.1	48.0
CD-Rom	9	5.1	53.1
Print	16	9.1	62.3
Internet	66	37.7	100.0
Total	175	100.0	

The table above indicates that the most popular distance leaning format is the internet (37.7%). Satellite Video Conference was second with 33.7%. Print and DVD are tied for third with 9.1% each and the Video (5.1%) and CD-Rom (5.1%) methods were the least popular.

The qualitative questions (Questions 26-31) were summarized in narrative form as well as in thematic form. These questions were primarily intended to gain information regarding ways in which to improve the course. According to the participants' responses, the course could be improved by:

- Increasing the length of the program so that the first day could be spent on interviewing and general epidemiology and the second day could be spent on specific disease processes.
- Organizing the rooms so that they are more conducive to movement.
- Improving the registration process and the process for submitting the pre-test. A few participants had trouble registering and submitting their pretests.

The thematic analysis indicated that participants would like more time spent discussing the following:

- Module1 at the beginning of the course; provide a brief outline.
- Real life situations.
- Viewing graphs and charts from scenarios.
- Food borne outbreaks.
- Outbreak investigation.
- The analysis process.
- Report writing.
- Specific Epidemiology case studies.

Courses that they would like to see offered by School Confidential include the following:

- Emergency preparedness for hospitals.
- Water contamination.
- Advanced BDLS.
- Nurse specific courses.
- Cultural Competence.
- Risk Communication.

- Bioterrorism, chemical and radiation response.
- WMD, CBRNE.
- Natural disasters (hurricane).
- Advanced Epidemiology with statistics.
- West Nile.
- Advanced Program X.

Additional Comments:

- Ask participants to set phone to vibrate.
- Please provide pre-test scores.
- Put the 10 steps of the epidemiology process, as well as other pertinent information, into a small slick sheet that can be carried into the field.

Assessment instrument

The assessment instrument was evaluated based on the post-test scores for those who took both the pretest and post-test. This method of analysis was chosen due to the following reasons: (1) the assessment should be evaluated based on its ability to assess what students have learned throughout the course and not assessed as a baseline measure and (2) including those that did not take the pretest would confound the results due to the fact that some of the participants would have taken the test twice while some would have been taking the test for the first time. In order to keep previous experience constant, only those who took both exams were included.

In future research, a comparison between those who took the pretest and those who did not should be performed on larger samples. In this way, one could determine student growth as well as the influence of having seen the test previously (Patten, 2004;

Cook & Campbell, 1979). Therefore, it is important to note that the scores on this assessment may be artificially inflated due to pretest sensitization.

Due to the binary nature of this assessment instrument (correct versus incorrect), non-parametric statistical techniques were used to evaluate the tool. Specifically, item difficulty and discrimination indexes were calculated and a binary confirmatory factor analysis was conducted using LISREL, version 8.7. However, a minimum sample of 300 is recommended for an instrument with 30 items (Kline, 1998). Therefore, once a larger sample is achieved, the analysis should be rerun.

A difficulty index and a discrimination index were calculated to determine the level of difficulty of the items and the ability of the items to discriminate between the lower and higher performers. An assessment that consists of too many easy questions or too many difficult questions may not be adequately assessing student knowledge and capabilities. However, an assessment instrument should have some easy questions and some hard questions to help discriminate between the low and high performers (Mertler, 2003).

The difficulty index can range from zero to one, with lower values indicating more difficulty. For example, a difficulty index of zero indicates that no one answered the question correctly while a difficulty index of one indicates that everyone answered the question correctly. For specific information on the calculation of the difficulty index please refer to Salkind (2003). The difficulty index for each of the 30 questions is provided in the figure below.

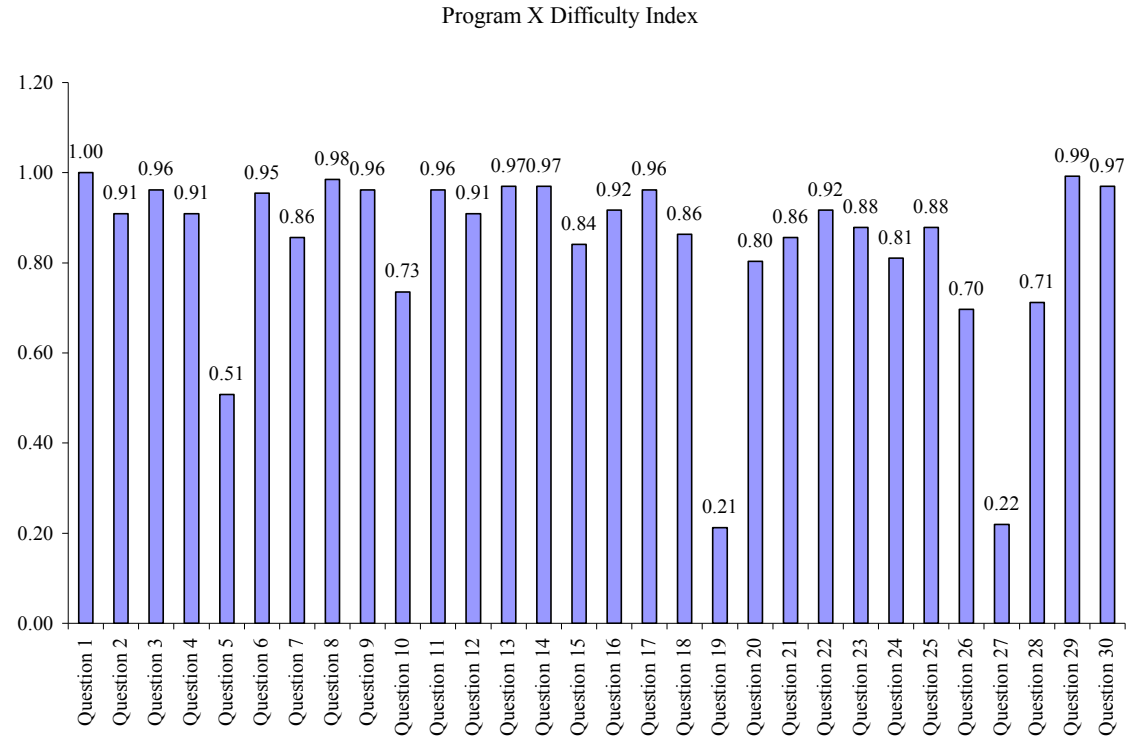

Figure 15. Program X Assessment Difficulty Index by Question.

According to the figure above there is a fairly wide range of item difficultness on the post-test. Question 19 is the most difficult (.19) followed by Question 27 (.22) and Question 5 (.51). The remaining questions are not difficult, although some are not considered to be easy either. In particular, Question 10 (.73), Question 15 (.84), Question 20 (.80), Question 24 (.81), Question 26 (.70) and Question 28 (.71) all have difficulty indexes less than .85 and are therefore not characterized as easy. The remaining questions with difficulty indexes greater than .85 are categorized as easy to very easy.

The discrimination index also ranges from a value of zero to one. Values closer to one indicate a greater ability to discriminate between the low and high performers. For specific information on the calculation of the discrimination index please refer to Salkind

(2003). The discrimination index for each of the 30 questions is provided in the figure below.

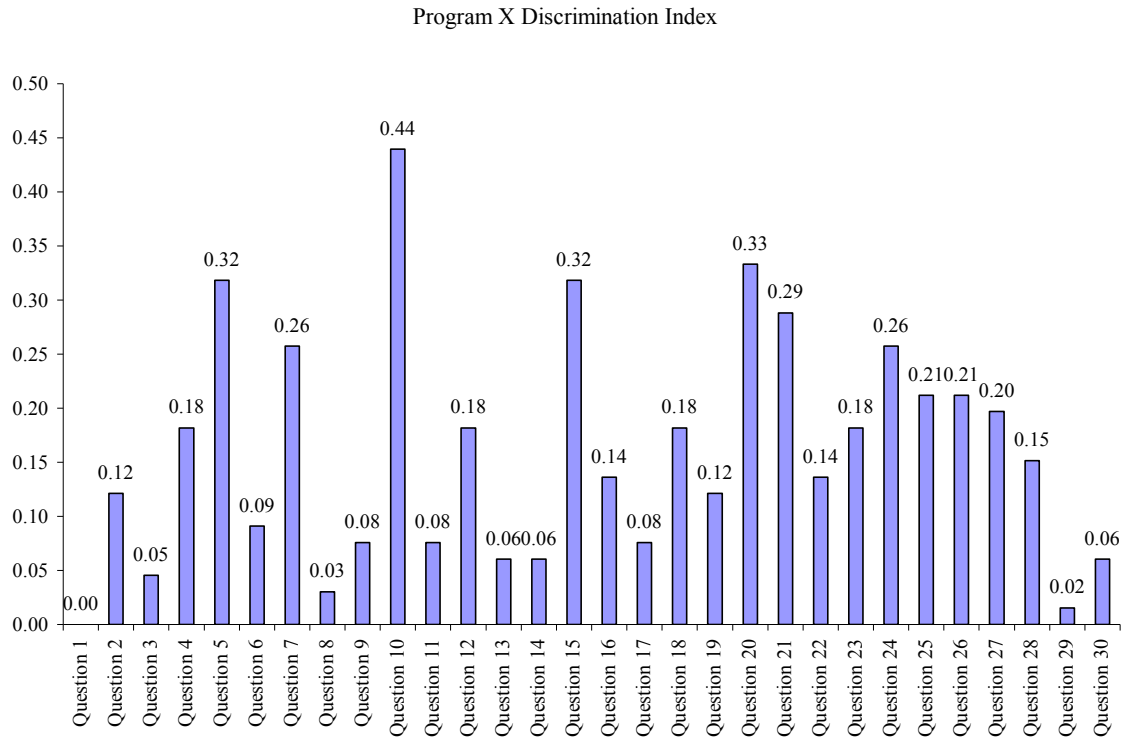

Figure 16. Program X Assessment Discrimination Index by Question.

Figure 16 indicates that there is wide range of discrimination indexes. Question 10 has the best ability to discriminate between the low and high performers. In addition, Question 10 evidenced a good deal of growth, on average (19 percentage points). Question 1 also evidenced a good deal of growth with participants showing an average increase of 19 percentage points on this item as well. In fact, everyone answered Question 1 correctly on the post-test. Due to the fact that everyone answered Question 1 correctly, it has no ability to discriminate between low and high performers.

Although Question 5 was difficult, it has a moderate ability to discriminate between low and high performers (.32). Also, participants did show some improvement on this item with an average increase of 13 percentage points. However, Questions 19 and 27 were also very difficult, but do not have a good ability to discriminate between low and high performers (.12 and .20, respectively). Furthermore, participants decreased their performance on average by 12 percentage points on Question 19 and showed no improvement on Question 27. Therefore, these two questions may be "poor" questions.

Questions 3, 8 and 29 yielded very high scores (98% or higher) on the pretest and therefore may actually be too easy. These items may be reflecting "common knowledge."

In addition to computing the difficulty and discrimination index, a confirmatory factor analysis was conducted. Questions 1 – 10 were loaded on Factor 1, Questions 11-20 were loaded on Factor 2 and Questions 21-30 were loaded on Factor 3. These loadings reflect the mapping of the individual questions to the three distinct modules.

The results indicated that the model would not converge. Therefore, the analysis was re-run with the entire post-test sample in order to increase the sample size. Once the additional cases were added (N = 261), the model successfully ran.

According to the model results, the data fit the model structure well according to the goodness of it statistic (.97) and the model converged in 39 iterations. However, the output indicated that the matrix analyzed was not positive definite which means that the matrix to be analyzed contains out of bound values. This could be due to multicollinearity in which two or more variables are highly redundant (Kline, 1998).

Therefore, these results should be used as a guide and not necessarily as a definite solution.

Only three of the questions did not have significant loadings on their corresponding factors. Questions 7, 8 and 12 were not statistically significant at the .05 level. Therefore, they are not explaining a significant amount of variability on their corresponding factors. Although a particular item may not be adding much explanatory value, it does not necessarily mean that the loading is incorrect. For example, it may not account for a significant amount of variance on that factor, but it accounts for more variance on its assigned factor than it does any other factor in the model.

In order to determine the accuracy of the loadings, modification indices were requested and indicate that all of the questions that load on the first factor (Questions 1-10) do in fact go together, however Questions 19 and 25-28 also load best on the first factor. Questions 11-18 and 20 load best on the second factor and Questions 21-24, 29 and 30 load best on the third factor. See Table 12 below.

Table 12

Suggested Factor Loadings

	Item																												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Factor 1	X	X	X	X	X	X	X	X	X	X									X						X	X	X	X		
Factor 2											X	X	X	X	X	X	X	X		X										
Factor 3																					X	X	X	X					X	X

Although modification indices are provided, more data should be collected and the analysis should be re-run with the original specifications, unless the recommended changes can be theoretically supported. This is especially true since the matrix analyzed was not positive definite and therefore these results may not be reliable or valid. Furthermore, modifying an instrument based on modification indices is not suggested

because the results could be sample specific and capitalize on chance. Therefore, prior to modifying the instrument, a validation sample should be obtained to ensure (i.e. confirm) that the modification indices are specifying a reliable and valid model.

Summary and Recommendations

Conclusions

The participant performance data indicate that students, on average, significantly improved from the pretest to the post-test, those who had the lowest scores on the pretest evidenced the most growth and participants did fairly well on the post-test, on average. The average performance on the post-test was 84% and the average growth was nine percentage points. All of the delivery sites showed growth on average with XXX yielding the most growth (13 percentage points) and XXX yielding the least growth (five percentage points). XXX also had the highest performance (89%) and XXX has the lowest (83%). Therefore, based on the participants' performance, the program appears to be effective with regard to delivering the course content to the participants.

The course evaluation data indicates that participants, in general, provided positive ratings with regard to the course content, course facilitator and course material. Also, participants tended to perceive the course delivery as effective and found that the program adequately met their professional needs.

The assessment instrument yielded a large range with regard to its level of difficulty and ability to discriminate between the low and high performers. Questions 3, 8, and 29 did not allow much room for growth due to exceptionally high performance on the pretest. Therefore, these items may be too easy. Conversely, Questions 19 and 27

were exceptionally difficult. Furthermore, these two items did not have a good ability to discriminate which suggests that they may be "poor" questions.

The assessment structure imposed on the data fit the data well (GFI = .97) and the items that were intended to measure the same learning objective significantly loaded on their corresponding factors with the exception of Questions 7, 8 and 12. These results suggest that, on average, the assessment instrument items are internally consistent and therefore reliable. These results also suggest that the instrument is likely to have construct validity. However, the results of the confirmatory factor analysis are questionable due to the fact that the matrix analyzed was not positive definite. Therefore, more data should be collected and the analysis should be re-run with a larger sample so that more confidence may be placed in the results.

Recommendations

Participant Performance and Assessment Evaluation

The evaluation of the participants' performance indicates that the participants did well on the post-test, on average, regardless of the program delivery site. Also, the fact that the participants showed growth from the pretest to the post-test suggests that the assessment instrument has at least some degree of content validity. If the test items were not related to the course content, no improvement would have been made. Furthermore, the fact that the assessment structure imposed on the data fit the data well and 27 out of the 30 questions had significant loadings indicates that the items are likely to be internally consistent and therefore the assessment instrument has construct validity. Based on these findings, no recommendations are made with regard to changing the

structure of the assessment instrument at this time. However, it is recommended that more data be collected and that the matrix be re-analyzed with a larger sample.

Although there are no recommendations to change the structure of the assessment instrument, some specific recommendations regarding the individual assessment items include the following:

- Remove Questions 3, 8, and 29 because they may be too easy and therefore not allow for much growth (i.e. ceiling effect). Furthermore, items with very little variance do not discriminate well and may create spurious relationships that cause multicollinearity problems as well as invalid model results.
- Remove (or rework) Questions 19 and 27 because they are exceptionally difficult and do not discriminate well between low and high performers. Difficult items that do not discriminate well are a sign of poorly written questions and/or questions that do not have content validity.

Due to the fact that the assessment instrument has 30 items and only three factors, the removal of the five above mentioned items should not affect the assessment instrument's ability to effectively measure participant performance. Also, if additional assessments are administered (i.e. role playing and other activities), then those additional items should be used in tandem with the primary assessment instrument. In this way, the data may be triangulated in the attempt to increase the reliability and validity of the assessment process.

Course Evaluation Data

According to the participant feedback, some of the program delivery recommendations are as follows:

- Review the course length; it may need to be extended to adequately cover all material.
- Spend more time on Module 1 and provide a brief overview at the beginning of the course.
- Review and discuss test results with the participants.

In addition to program delivery recommendations based on participant feedback, recommendations are also provided with regard to the style and content of the course evaluation forms. The recommendations focus on the scale length and format

Scale Length

Typically, Likert scales have five or seven degrees of measurement. For this particular course evaluation, a five-degree scale is recommended. For example: 1) Strongly Disagree, 2) Disagree, 3) Not Sure, 4) Agree, 5) Strongly Agree. The current four-degree scale does not provide a sufficiently sensitive measure of the respondent's answers to draw powerful conclusions.

Scale Format

Questions 20-23 follow a different format than questions 1-19. In questions 1-19 the scale follows a negative to positive format while questions 20-23 follow a positive to negative format. To avoid confusion or error, all scales should follow a consistent format when possible.

The qualitative section asked respondents to list any courses related to bioterrorism preparedness, infectious diseases, and other emerging health threats they would like to see offered by School Confidential. This question is useful if departments do not have any specific programs in mind. However, if there are a few courses that School Confidential is considering adding to the course selection, then this may be a good way to get feedback on which courses they should include at a later date.

Cultural Competence, nurse specific, West Nile, natural disaster and water contamination courses were requested by many Program X participants and they have been a popular request in other programs evaluated as well. Therefore, School Confidential may want to incorporate these types of courses into their curriculum.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & analysis issues for field settings.* Geneva, IL: Houghton Mifflin Company.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling.* New York, NY: Guilford Press.

Mertler, C. A. (2003). *Classroom Assessment: A practical guide for educators.* Los Angeles, CA: Pyrczak.

Patten, M. L. (2004). *Understanding Research Methods: An overview of the essentials.* Los Angeles, CA: Pyrczak.

Salkind, N. J. (2003). *Exploring Research* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Appendix

The student assessment instrument was removed from this document to protect the confidentiality of the client and to keep the test itself confidential