Revisiting Quantitative Methods for Evaluating Training Programs for Systems Undergoing Operational Test (OT)

Dr. Christopher D. Hekimian
Sr. Systems Engineer and Policy Analyst
SAIC
4501 Ford Ave., Suite 330
Alexandria, VA 22302
Tel: 703 499-0518
E-mail: hekimianc@saic.com

Ms. Laura Chan
Sr. Engineer/Analyst
SAIC
4501 Ford Ave., Suite 330
Alexandria, VA 22302
Tel: 703 296-0057
E-mail: laura.w.chan@saic.com

About the Presentation...

- We shall revisit a method that can be used to quantitatively assess the efficacy of training programs
- The purpose is to provide a basis to state, within a prescribed degree of confidence, whether a change in fielded system performance can be attributed to technical issues or to training
- A classical, hypothesis-based approach, where operator “expert” or “trainee” -status is the independent variable, is described

Contents

The Problem

- *A system seems to perform well during Developmental Testing (DT) but when made operational, there is a notable decline in performance...* ***Is the problem with the system or with the training?***

Common Methods for Assessing Training

- Surveys, Questionnaires
- Interviews
- Focus Groups

These provide qualitative assessments that are not only based upon the respondents' experiences during OT, but also upon the aggregated experience, knowledge, feelings and attitudes of the respondents.

Drawbacks of Common Methods for Assessing Training

- Subjectivity
  - What one person's assessment of what is good/bad or acceptable/unacceptable is likely to vary based upon numerous things, including the background of the individual responding to the question
    - Compounding the problem is that survey question responses ranges are seldom "anchored" to something objective
- Internal Validity
  - Due to ambiguity or diverse definitions, interviewer or survey questions may not be measuring what the evaluator is interested in
    - Scientifically valid surveys are pre-tested and aligned in order to ensure validity.
- External Validity
  - The results of the sample survey may not be extendable to the larger population of all potential users of the system
- Bias
  - Some respondents may have a conscious or subconscious bias towards a given question response
- Apathy
  - Some respondents may resent the additional demand on their time of one more surveys. They may not use care in responding to the survey questions

Advantages of Common Methods

- Requires little or no additional testing because responses are based upon results obtained while collecting other measures
- Provides the opportunity for other relevant insights regarding the system or related DOTMLPF to be collected

DOTMLPF: Doctrine, Organizations, Training, Materiel, Leadership And Education, Personnel, And Facilities

Experiment Design Method for Determining the Efficacy of Training Programs

Hypothesis testing allows one to make an authoritative statement like :

"the training did not have a significant impact on the OT performance results"

- The ability gained through training is assigned as the independent variable
  - This is done using two test methods:
    - *Realistic Scenario Testing (RST)* performed by experts under conditions that replicate operational conditions to the greatest degree possible - - establishing a performance baseline
    - *Operational Testing (OT)* performed by actual end users in operational conditions or in a realistic operational exercise
- Experiment involves the test results of two groups, the experts performing RST and the end users performing OT
- Are the performance results of the experts during RST that much better than the real operators during OT?
- Validity of the results depends on the faithfulness with which the RST replicates the OT

Defining the Performance Metric

- The data elements that would normally be used to assess effectiveness of a system are used to assess training efficacy
- All variable types can be used
  - Nominal
    - Not associated with a value, just a label or categorization
  - Ordinal
    - Associated with a range of integer values such as with a Likert scale (i.e., 1 to 5 scale)
  - Ratio
    - Value can be any real number
- Performance metrics that are returned in averaged amounts need to be partitioned
  - i.e., instead of considering one test of 100 trials, ten averages of ten different trials are computed
    - The need to do this will soon be made plain

Governing Assumptions

- Systems used during OT and RST are essentially the same
- Experts are sufficiently experienced with the system
- RST environment and test conditions are sufficiently similar to OT condition
- Test subjects (if any) are similar between RST and OT
- Effects of operational stress during OT are negligible or somehow duplicated in RST

Statistical Significance

- Randomness effects all experimental results. Simply looking at the overall number of successes and failures in a test is an insufficient basis for conclusions. One must determine if the test results show *statistical significance*.

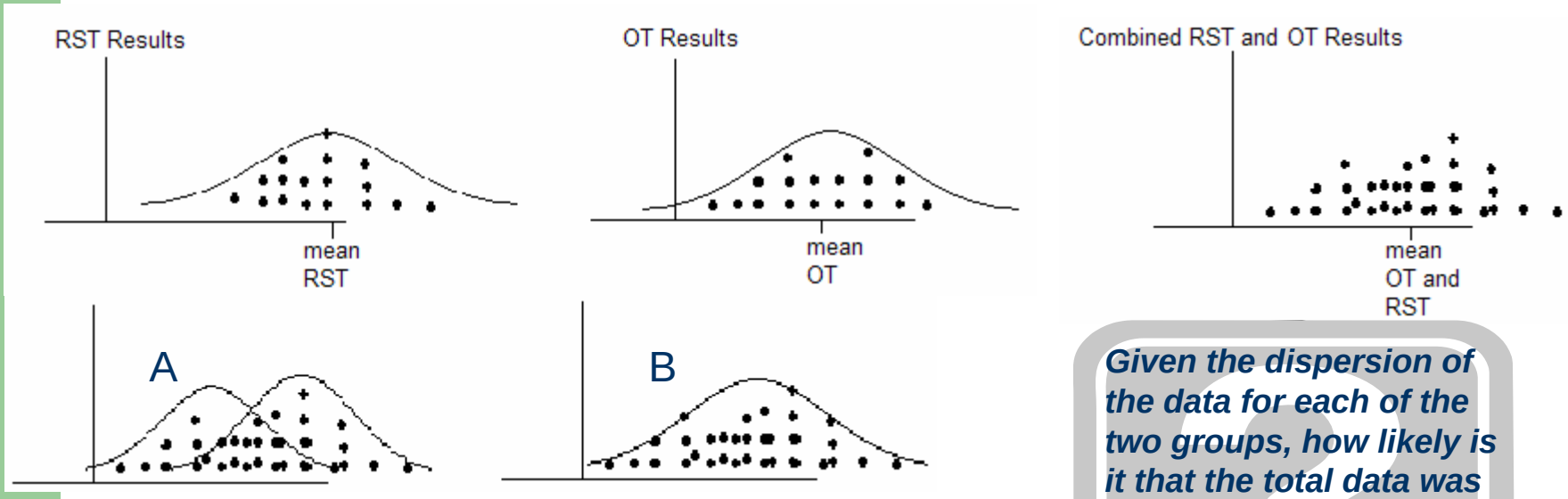
- Significance tests account for the possibility of a test result occurring due to chance alone. Various tests for significance have been developed and the correct test must be applied based upon the type of variables and the data distribution.

- Two of the most common tests for significance are the t-test, which is used for parametric (normally distributed) data; and the Chi-square test, which is used for nominal (categorical) data.

- If a significance test indicates that there is less than a 5% chance that the observed difference in results could be due only to chance, then usually a significant effect of the independent variable upon the results is noted.

Significance Calculations

- The statistical tests of Chi-Square (for nominal variables) and t-test (for ordinal, ratio, or averaged data) are available to test the following hypothesis:
  - "Training did not have a significant impact on the performance of the system during OT"
- The t-test will provide a measure of the likelihood that the performance averages taken from two test groups were taken from distributions with the same mean



Given the dispersion of the data for each of the two groups, how likely is it that the total data was drawn from two separate distributions (A) or from a single distribution (B)?

Statistical Significance- Chi-Square [1]

- Applicable to nominal test result data
- Microsoft Excel and other spreadsheet programs can accommodate Chi-square calculations
- “Help files” provide usage and application instructions

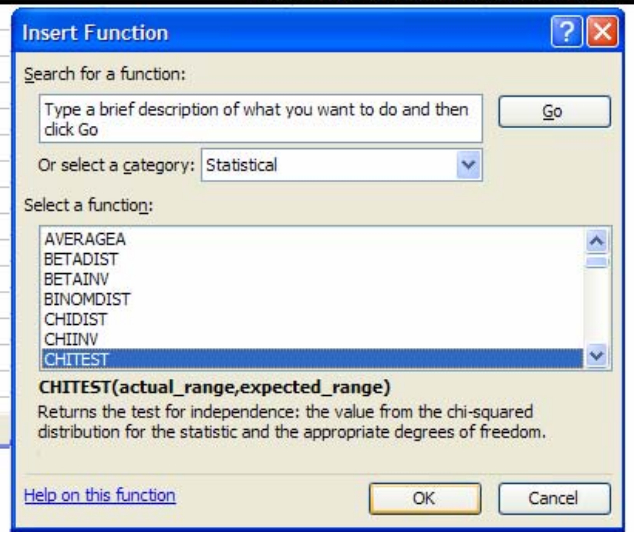
Statistical Significance- Chi-Square [2]

Book1

	A	B	C
1	RST Totals by Experts	OT Totals by End Users	Result
2	A2	B2	Able to assemble system < 5 minutes
3	A3	B3	Unable to assemble system < 5 minutes
4	RST Expected Values	OT Expected Values	Result
5	(A2+B2)(A2+A3)/(A2+B2+A3+B3)	(A2+B2)(B2+B3)/(A2+B2+A3+B3)	Able to assemble system < 5 minutes
6	(A3+B3)(A2+A3)/(A2+B2+A3+B3)	(A3+B3)(B2+B3)/(A2+B2+A3+B3)	Unable to assemble system < 5 minutes

Book1

	A	B	C
1	RST Totals by Experts	OT Totals by End Users	Result
2	6	11	Able to assemble system < 5 minutes
3	2	5	Unable to assemble system < 5 minutes
4	RST Expected Values	OT Expected Values	Result
5	5.67	11.33	Able to assemble system < 5 minutes
6	2.33	4.67	Unable to assemble system < 5 minutes
7			
8			
9	=		
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			



1. Populate a table as shown. Note that columns A and B of rows 2 and 3 are the results of tests. The same columns in rows 5 and 6 are derived as shown from the test results above them
2. Use INSERT => FUNCTION and then select the STATISTICAL category. Select CHITEST
3. The wizard will allow you to select the cells of the test results (4 cells, A2,A3,B2,B3 in our example), and then the expected values that you have inserted, each, respectively
4. Clicking OK will return the result of the Chi-Square Test
5. The function returns the probability that the difference in performance between the RST and OT tests could have occurred due to chance as opposed to due to training effects (i.e. CHITEST value of 0.05 indicates a 5% probability that the difference in test results was not due to training issues [1,2])

1.) Based upon all governing assumptions

2.) Typically, 5% is the threshold (CHITEST >.05) where one would assume that the effects of training were not a significant factor in the difference in performance

Statistical Significance- t-Test [1]

- Applicable to normally distributed data such as those returned from survey results or laboratory measurements
- Microsoft Excel and other spreadsheet programs can also compute t-test results

Statistical Significance- t-Test [2]

Book1

	A	B	C
1		Kill Rates using new Targeting System	
2	Trial	RST Results by Experts	OT Results by End Users
3	1	0.8	0.8
4	2	0.85	0.83
5	3	0.82	0.84
6	4	0.9	0.9
7	5	0.92	0.88
8	6		0.9
9	7		0.92
10	8		0.91
11	9		0.8
12	10		0.78
13	11		0.8
14	12		0.9
15	13		0.92
16	14		0.76
17	15		0.77

- By entering “1” for the test type a paired t-test could be performed
- A paired t-test could be used in order to compare the ordered results of two data sets for one group, one before and one after training
- This would be a direct, single group test on the efficacy of a training program
- Such results could be of interest but would need to be interpreted carefully due to the possibility that the results of user “habituation” with the system is the causal agent and not training (internal validity)

1. Populate a table as shown. Note that the columns B and C might be populated with test results for individual test trials, averaged trial results for a particular test subject, or averaged survey results
2. Use INSERT => FUNCTION and then select the STATISTICAL category. Select TTEST
3. The wizard will allow you to select the cells of the RST test results (B3:B7 in our example), and then the OT results (C3:C17). The number of entries in each column need not be the same for our example
4. Enter a “2” for the number of tails to use for the test. This will account for the possibility that the end users over or under perform the experts
5. Enter “3” for the test type as there is no reason to assume that the variances of the two data samples will be the same
6. The function returns the probability that the difference in performance between the RST and OT tests could have occurred due to chance as opposed to due to training effects (i.e. TTEST value of 0.05 indicates a 5% probability that the difference in test results was not due to training issues *). If TTEST >0.05 assume training was not a performance issue

* Based upon all governing assumptions

Partitioning Technique for Averaged Performance Parameters (1 of 3)

Biometric System Performance Example

- Administrators of a fingerprint access control device are concerned as to whether the high False Accept Rate (FAR) of their system is due to poor system performance or might be a training related issue
- FAR is an aggregated performance parameter that is compiled from a large number of trials... In this case, fingerprint authentication transactions
- The methods described previously are not suitable for aggregated performance parameters because the data are not nominal, and the FAR is already an averaged value and thus, there would be no multiple trials to average and analyze.

Partitioning Technique for Averaged Performance Parameters (2 of 3)

- In order to employ the method, the two FAR tests (RST and OT) of 500 test subjects each are partitioned into twenty FAR tests (ten each RST and OT) of 50 test subjects each.
  - Every effort is made to ensure that the systems, the environment, and the test subject demographics are similar between RST and OT
  - Partitioning is employed in this case so the ten RST and OT FAR tests can be analyzed using a t-test
  - Partitioning can be by order, as in this example, or randomized
  - The number of test subjects and trials between the RST and OT tests need not be the same
  - FAR tests by their nature typically require a large number of test subjects
  - Overall system FAR calculations should be based upon all 500 trials for each test- - with no attempt being made to combine the individual FAR results into one representative one

Partitioning Technique for Averaged Performance Parameters (3 of 3)

- RST outperformed OT. The t-test indicated that there was an 28% likelihood that the RST and OT FAR test results could have been drawn from distributions with identical means
  - Because of uncontrollable factors such as stress during OT, and the relatively large likelihood that the performance difference was due to chance (28%), Analysts do not attribute the lapse in performance to training issues.
  - Technical solutions are emphasized over training ones

- The same data may be analyzed in an identical manner for different aggregated parameters (such as False Reject Rate)
  - The results should not considered to be confirmatory because the two analyses are not based upon independent data
  - If the <5% threshold for significance is met, then training problems would be indicated

Advantages and Drawbacks of Quantitative Methods for Evaluating Training

Advantages:

- The methods are objective in terms of data collection, in terms of analysis and in terms of interpretation of results
- The methods are associated with specific levels of confidence for conclusions

Drawbacks:

- Cost of conducting RST in addition to OT

Conclusions

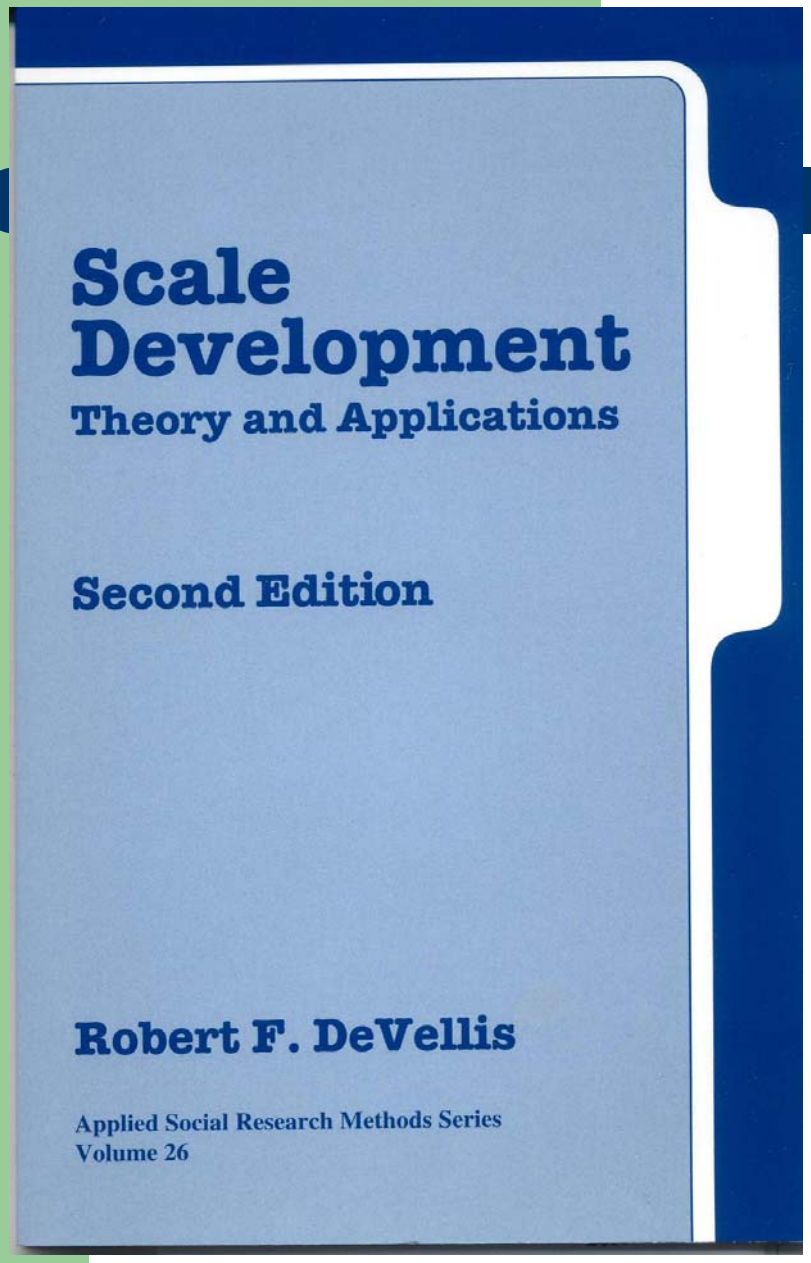
- Survey methods can be supplemented, confirmed, or refuted using quantitative methods
- Method for attributing lapses of performance to either system or training was discussed
  - Experimental design using RST and OT establishes training as the independent variable
  - Nominal performance metric (Chi-square)
  - Ordinal and ratio metrics (t-test)
  - Aggregated performance metric (t-test)
- Other applications for using quantitative methods for analyzing training programs include:
  - Side by side comparison of different training approaches
  - Pre and Post training analyses
  - Analysis of effectiveness of modifications to training programs

Thank you!

Dr. Christopher D. Hekimian
Sr. Systems Engineer and Policy Analyst
SAIC
4501 Ford Ave., Suite 330
Alexandria, VA 22302
Tel: 703 499-0518
E-mail: hekimianc@saic.com

Ms. Laura Chan
Sr. Engineer/Analyst
SAIC
4501 Ford Ave., Suite 330
Alexandria, VA 22302
Tel: 703 296-0057
E-mail: laura.w.chan@saic.com

Back-up



An excellent reference describing how to apply Likert type scale survey techniques correctly