A MODEL FOR EVALUATING ONLINE INSTRUCTION

Andrew Hale Feinstein
UNLV Department of Food and Beverage Management
andyf@unlv.nevada.edu

ABSTRACT

This paper describes a step-by-step procedure for evaluating the effectiveness of online instruction. Instrument development and validation are explained. Discussions on experimental design, statistical models, and interpretation of the data are also provided. This procedure differs from the often utilized comparative assessment methodology and instead relies on the Interservice Procedures for Instructional Systems Development and Bloom's Taxonomy of Educational Objectives to develop a model that evaluates cognitive ability increases in learners. Conclusions provide benefits for using this model when evaluating online courses.

Keywords: IPISD; Bloom's Taxonomy of Educational Objectives; online instruction assessment.

Note: This research was made possible by a grant from the National Restaurant EUCATIONAL Foundation (www.nraef.rog)

INTRODUCTION: THE COMPARATIVE STUDY DILEMMA

One of the greatest challenges to instructional system designers is to empirically assess the educational effectiveness of the instructional systems that they create. It seems that the major roadblock to this type of assessment in the inherent difficulty in comparing one form of instruction to another; a common approach that many researchers have attempted – and subsequently failed.

Imagine attempting to empirically assess whether online instruction is educationally superior to the traditional (brick and mortar) form of instruction. This type of comparative assessment has been done so frequently that a book summarizing these studies has been published (Russell, 1999) and a web site (http://teleeducation.nb.ca/nosignificantdifference/) classifying these articles has been created.

The difficulty lies in the fact that it is almost impossible to hold numerous independent variables constant during the comparative assessment – variables that can certainly have an impact on whether a student's outcomes are satisfactory or not. Time is one of these variables. A student who participates in a traditional classroom for 3 hours per week for 16 weeks cannot be meaningfully compared to a student who spends 30 hours online studying the same subject matter. Other variables that are difficult to control include the variability between the competence of the traditional instructor and the online course designer, the time of day students participate in the instruction, the differences in the use of visuals and other supporting materials, the sophistication of the software and hardware being utilized, and connectivity variances. This type of assessment quickly becomes an *apples to oranges comparison*. In essence, online instruction is not being compared to traditional instruction; instead, two unique courses are being compared.

I propose that rather than comparing online instruction to traditional forms, designers focus on evaluating the educational effectiveness of online instruction on its own merits. Designers can accomplish this by primarily focusing on the cognitive implications of their online courses. They can determine the cognitive abilities that are increased in learners who participate in a particular online course and use these results to determine the benefits of the instruction. If subsequent comparisons are need to justify the creation or continued application of an online course, this could then be accomplished by evaluating controllable variables such as costs, availability, a student's technological sophistication, and convenience. It is therefore the purpose of this paper to describe a process by which designers can empirically evaluate the effectiveness of the online courses that they create.

IPISD AND ADDIE

In the early 1970's, after reviewing its training methodologies, the United States Army decided to create a comprehensive instructional system to train and educate its personnel (Branson, 1973; Dick & Reiser, 1989; Logan, 1982). This instructional system was named IPISD (Interservice Procedures for Instructional Systems Development). It was so effective in training and educating the Army, that all American military branches soon adopted it. Today, it is one of the "most highly detailed models if the ID [Instructional Design] process generally available" (Gustafson & Branch, 1997, p. 62). Large corporations frequently rely upon this procedure when implementing training and development projects (Rossett, 1987, 2001). The entire model can be purchased as a four volume set from Educational Resources Information Center ([ERIC] http://www.ericfacility.net/extra/index.html).

At the heart of IPISD is a five-phase procedure often termed the *big box model* (Rossett, 1987). Each of the phases or boxes − analysis, design, development, implementation, and evaluation (ADDIE) − defines an integral procedure for creating an instructional system. Originally, the IPISD utilized the acronym ADDIC, where the "C" represented *control;* or the internal and external *evaluation* of the instructional system.

The first phase, analysis or needs assessment, allows researchers to analyze the educational and instructional procedures currently in place and determine if there are any gaps between learners' knowledge and educators' or other concerned parties' "visions of desired knowledge or performance" (Rossett, 1987, p. 15). This can be accomplished by assessing a learner's actions or other outcome indicators and comparing them to management interviews or a review of extant data (Babbie, 2003). If results from the needs assessment suggest that there is a gap and that implementing an instructional system could close it, subsequent steps in the big box model can then be taken.

The design phase allows instructional designers to create a plan of action for closing the gap described in the analysis phase. This typically involves outlining and describing the objectives, strategies, goals, and technologies of the instructional system.

Relying upon ADDIE, instructional designers can develop an instructional system based on the aforementioned design phase. The objectives, strategies, and goals are used as guides in an effort to ensure that the instructional system focuses on resolving current instructional inefficiencies. At this point "educators select methods, technologies, sequence, materials, practices, etc." to develop the instructional system (Rossett, 1987, p. 11).

The implementation phase allows instructional designers to implement or try out their instructional system. At this point, outcome data can be collected to use during the final phase of ADDIE.

The evaluation phase is used to ascertain if the instructional system has filled the gap described in the needs assessment. This phase also provides instructional designers with information to determine the "worth of the training effort" and if the goals created in the first phase were achieved (Rossett, 1978, p. 10).

This paper focus on the last phase of the ADDIE model. Next, I will explain how to develop a model of evaluation to ascertain the effectiveness of an online course based on the cognitive abilities that it increases in learners. To do this, I will rely upon the oft used *Bloom's Taxonomy of Educational Objectives*.

BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES

Bloom's Taxonomy of Educational Objectives was developed in the late 1950's as a tool to assist evaluators in classifying test items and testing outcomes (Bloom et al., 1956). The taxonomy is broken down into three domains of behavior: cognitive, affective, and psychomotor. Cognitive behaviors represent the use of knowledge or intellectual ability. Bloom believed that the majority of educational outcomes come from this domain.

As most educators know, Bloom et al. broke down the cognitive domain into major classifications. These major classifications are hierarchically organized from simple to complex levels of cognitive behavior. Knowledge, comprehension, and application are low-level cognitive abilities. Learners typically rely upon inert knowledge − such as a single memory, interpretation, or rule − to evoke these types of responses.

Analysis, synthesis and evaluation are high-level cognitive abilities. Questions framed by one of these classifications require the learner to invoke multiple memories, interpretations, and rules. Further, learners have to place these thoughts into a new contextual environment. Therefore, these cognitive abilities require knowledge that is dynamic.

By developing an instrument based on Bloom's Taxonomy of Educational Objectives that assesses a student's cognitive abilities, one is able to isolate specific learning outcomes. Specifically, the instrument can be used to assess the effect of instruction on learners' inert and dynamic knowledge and to assess the effect on cognitive abilities classified under this taxonomy.

CREATION OF THE INSTRUMENT

To evaluate the effectiveness of an online course, an instructional designer can create an assessment instrument consisting of six sets of five multiple-choice questions (30 questions in all). Each question set is oriented towards the student using a particular classification in the taxonomy as the highest cognitive ability needed to answer the question. The first set of five questions is directly related to the knowledge classification described in the taxonomy. This is the lowest classification in the taxonomy and assesses learners' recollection of facts and information pertaining to an educational outcome. Each subsequent set of questions follows the taxonomy's classifications, ending with questions 26-30, which relate to problems that require learners to invoke the cognitive behavior of evaluation. Question sets should be presented to the learner in the same order that they were presented in the taxonomy. Allowing learners to answer relatively easy questions first and gradually increasing the complexity of the questions minimizes the anxiety of learners. This technique can be observed in many standardized tests such as the SAT and GMAT (Educational Testing Service, 2003).

The key to developing the 30 multiple-choice questions is to identify important material in the online instruction, create questions, and effectively classify them. As in all assessment design, the creation of these questions should be done with learning targets in mind (Nitko, 2000). Most times, new topic terminology can be used to create inert questions while problems can be classified as dynamic questions. Because the question set is the crux of the assessment, I strongly recommend that instructional designers review Blooms initial taxonomy or the recent revision (Anderson, et al., 2001) to see numerous question examples that have been correctly classified.

After the questions have been created, the instrument must be validated. First, the content of the questions must be evaluated. The stem of the question and the keyed answer must be reasonably proven to be correct. The distractor responses must also be proven to be incorrect. One way to do this is to gather 20 or more experts on the subject matter and ask them to review the instrument for content integrity. This can be done through a Delphi-type process or an informal meeting. The objective is to ensure that the experts agree on the content validity of the instrument.

Next, the evaluator must ensure that each set of questions fall within the ascribed cognitive classification and thereby assist in minimizing the possibility of confounding in the study. A process that has worked is to randomly order the 30 questions so

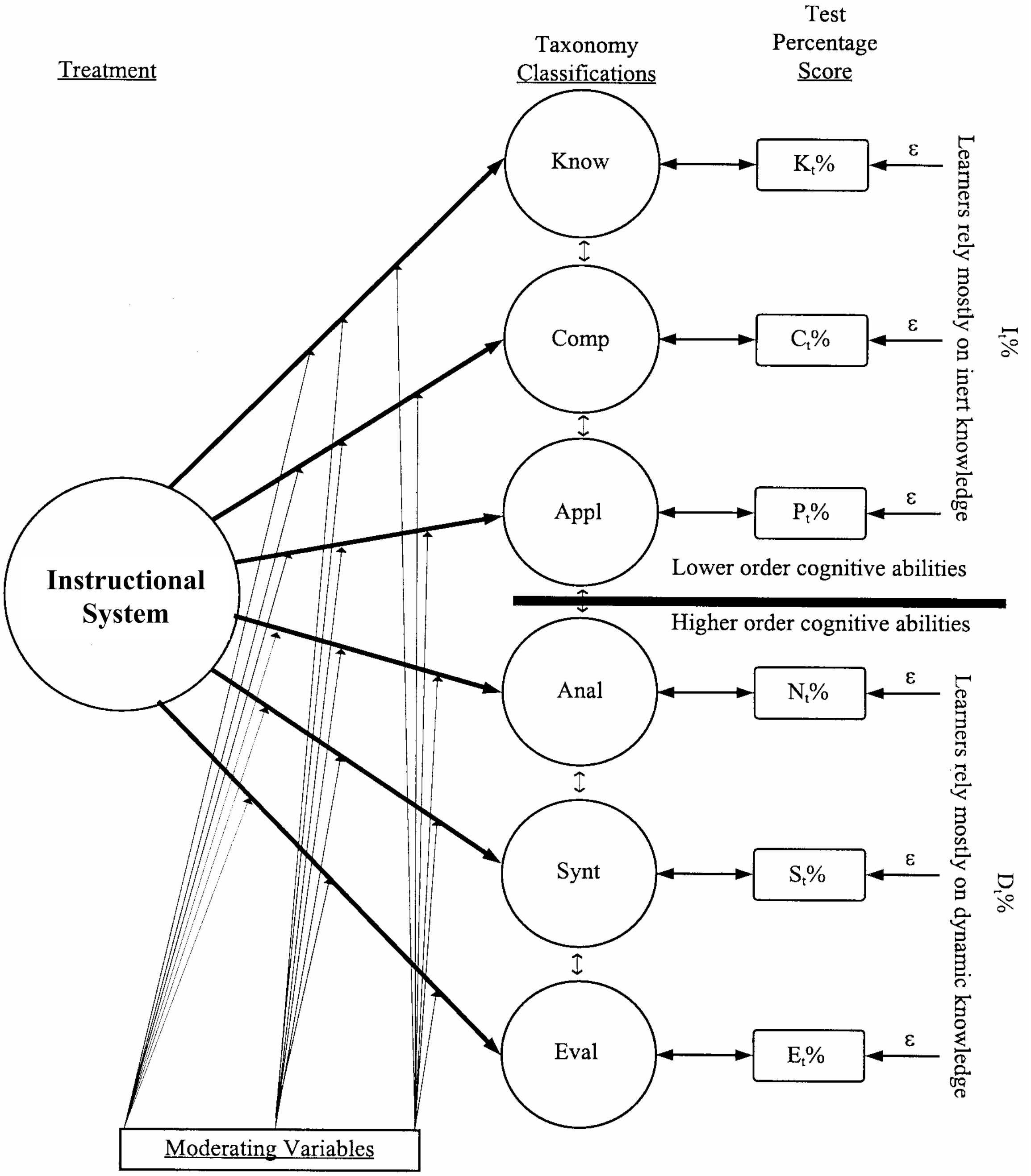

Figure 1. Conceptual Model of the Instructional System Assessment

as to mix the sets of classified questions. A panel of instructional designers very familiar with Blooms Taxonomy can then review the instrument and classify each question by placing the first two letters of the classification heading next to the associated question. In cases where there is disagreement, questions can then be discussed and modified by the group until there is unanimous agreement about the question's classification.

EXPERIMENTAL DESIGN

After the instrument has been created, the experiment must be developed. Score differences on exams can be evaluated by creating a classic quasi-experimental design (Babbie, 2003; Cook & Campbell 1979). As is the case in this model, this type of experiment often requires a repeated-measures, pretest-posttest, control-experiment groups design. In essence, learners need to be randomly assigned to one of two groups: control group (C) or experiment group (E). Groups C and E first take the instrument prior to the online instruction. After which, group E participates in the online instruction; for however long that may take. Group C continues with their normal activities but does not participate in the online instruction. After the online instruction, both groups retake the instrument.

In a well-controlled experiment, group C functions exactly as group E, except for the new instruction. For instance, if an online course on customer problem resolution was provided to employees, the control group would continue with their daily job duties, while the experiment group would participate in the online instruction.

The design of the experiment is structured so that there are two clearly defined sets of variables − independent and dependent. The independent variables include two bimodal variables − pretest/posttest (test) and control/experiment (group). An individual student variable (student) is used to identify each participant. Moderating variables − representing demographics of the participants – can also be included in the study as independent variables. Dependent variables consist of eight percentage scores acquired from the assessment instrument. Percentage scores are the participants' percentage of correct responses for each of the six classifications and a combined inert knowledge score and a combined dynamic knowledge score (Figure 1).

ANALYSIS

Analysis is broken into two phases. Phase one is used to determine if any of the groups or tests are significantly different from each other on any of the eight percentage scores. This is accomplished by analyzing the data using an unbalanced, repeated-measures ANOVA at the significance level of $\alpha = 0.05$. Figure 2 depicts the model to be utilized in phase one.

$$\mathbf{Y_{ijk} = \mu + Group_i + Test_j + Group*Test_{ij} + Student(Group)_{k(i)} + \varepsilon_{k(ij)}}$$

Where:

Y_{ijk} = Response for ijk - th individual

μ = Overall Mean

$Group_i$ = Fixed Effect, i = 0,1 (Control, Experiment)

$Test_j$ = Fixed Effect, j = 0,1 (Pretest, Posttest)

$Student_k$ = Random Effect, k = 0,1,2…n (Control Participants),
k = 0,1,2…n (Experiment Participants)

$\varepsilon_{k(ij)}$ = Error Term = Student (Group * Test) $_{k(ij)}$

Figure 2. Model Used for Phase One Analysis

Phase two of the analysis focuses on determining if any moderating variables might explain participants' gains in cognitive abilities from participating in the online instruction. Phase two can also be used to determine whether any demographic characteristics might be able to identify homogenous traits in some of the participants within their respective group. The data in phase two is analyzed similarly to phase one, however the repeated measures component is removed. This is done by creating gain scores through the subtraction of an individual's pretest scores from their posttest scores. Figure 3 depicts the model to be utilized in phase two.

$$\mathbf{Y_{ijklmnopqrstu} = \mu + Group_i + Moderating1_j + Moderating2_k + Moderating(n)_{l+n}\ Group*Moderating1_{ij} + Group*Modearting2_{ik} + Group*Moderating(n)_{i(l+n)} + \varepsilon}$$

Where:

$Y_{ijklmnopqrstu}$ = Response for ijk(l+n) - th individual

μ = Overall Mean

$Group_i$ = Fixed Effect, i = 0,1 (Control, Experiment)

$Moderating1_j$ = Fixed Effect, categories based on variable

$Modearaing2_k$ = Fixed Effect, categories based on variable

$Modearting(n)_{l+n}$ = Fixed Effect, last moderating variable, categories based on variable

ε = Error Term = All three-way and higher interactions

Figure 3. Full Model Used for Phase Two Analysis

Moderating variables should be selected based on previous research on the content area of the online instruction or by identifying exploratory variables that might have an impact on learning outcomes. Common choices include level of educational attainment, gender, age, particular levels of experience, etc. Take into consideration that the more moderating variables that you include into the study, the larger the sample will need to be. Sample size should be calculated by either using the power or estimation approach to sample size planning. For further discussion of sample size estimations, I suggest that you consult the seminal reference on linear models (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

RESPONSE VARIABLE TRANSFORMATION

Dependent variables will first have to be converted from proportional to continuous data. This is done because the eight percentage scores from the assessment instrument are constrained between -1 and 1 (100% being a perfect score). One of the basic assumptions of ANOVA is that errors are normally distributed. Therefore, the proportional data must be converted to continuous data using a transformation calculation.

Two transformations should be evaluated: arc sine and logit. The logit transformation requires numbers to be greater than 0. Therefore, when percentage scores are 0, 0.005 will need to be added. The logit transformation calculation is shown in Figure 4 and the arc sine calculation is shown in Figure 5.

$$\boxed{\log\left(\frac{\dfrac{\text{percentage score}}{100}}{1-\left(\dfrac{\text{percentage score}}{100}\right)}\right)}$$

Figure 4. Logit Transformation Calculation Used on Inert and Dynamic Knowledge Percentage Scores

$$\boxed{\arcsin\sqrt{\frac{\text{percentage score}}{100}}}$$

Figure 5. Arc Sine Transformation Calculation Used on Inert and Dynamic Knowledge Percentage Scores

To determine which transformation yields a higher consistency of error variance and a greater number of normal distributions of errors, data should be transformed using both methods. Transformed data representing D_t% and I_t% should be used as response variables to analyze the logit transformed data and the arc sine transformed data. The input variables for the models should be group, test, interaction between group and test, and group nested within student (as shown in Figure 2).

Standardized residuals of each of these four models are then potted in three ways: residuals versus fitted values are plotted to determine if a pattern existed; a normal probability plot of residuals is done to see if a diagonal positively sloping line exists; and a histogram of residuals is charted to determine if a bell-shaped curve is apparent. All of these measures are undertaken to determine which transformation produces a greater number of consistent error variances and normal error distributions. For further information on assessing transformations of linear models, I again suggest consulting the seminal reference on linear models (Neter, Kutner, Nachtsheim, & Wasserman, 1996). After carefully reviewing the graphs, determine which transformation yields the best results and transform all dependent variables.

INTERPRETATION OF DATA

Depending on the statistical package that you use (I use Minitab and SPSS), phase one can be run as a General Linear Model (GLM). The benefits of a GLM over ANOVA are that you can include both metric and nonmetric independent variables and nested variables. GLM also takes into consideration an unbalanced design if you do not have the same number of participants in each group.

Phase one analysis will provide summary results (in ANOVA tables) for each of the eight percentage scores from the control and experiment participant group's pretests and posttests. The analysis will determine if any one of these four cells (control pretest, control posttest, experiment pretest, experiment posttest) are significantly different from one another in any of the eight models. A group nested within student variable is assessed in each model to determine if the groups consisted of homogenous percentage scores. If the nesting is determined to be significant in any one of the eight models, the variability suggests that each group consists of a heterogeneous set of participants. These findings would support the creation of analysis phase two.

For a particular online instruction to increase the cognitive ability in learners, phase one would first have to identify that a significant difference exists in one of the models. Next, a post-hoc analysis (such as Tukey's) would need to be done to determine where that change occurred. The best identification that a significant change in cognitive ability has occurred is when the interaction between test and group is significant. This would indicate that there are differences between the groups and their scores on the exams. An example of this significant interaction is provided in Tables 1 and 2 and Figure 6.

Table 1
Example of a Factorial ANOVA Table for Arc Sine Transformed Dynamic Knowledge Scores

Source	Df	SS	MS	F	P
Group	1	2.4E-02	2.4E-02	40.75	0.000
Test	1	3.3E-02	3.3E-02	85.95	0.000
Group by Test	1	3.6E-02	3.6E-02	94.05	0.000
Group nested within Student	241	1.4E-01	5.8E-04	1.52	0.001
Error	241	9.3E-02	3.9E-04		
Total	485	3.3E-01			

Table 2
Example of Means and Standard Deviations for Dynamic Knowledge Scores by Group and Test

			Test			
Group	Pre			Post		
	n	X	S.D.	n	X	S.D.
Control	121	.32	.220	121	.30	.192
Experiment	122	.28	.183	122	.67	.212
Total	243	.30	.203			

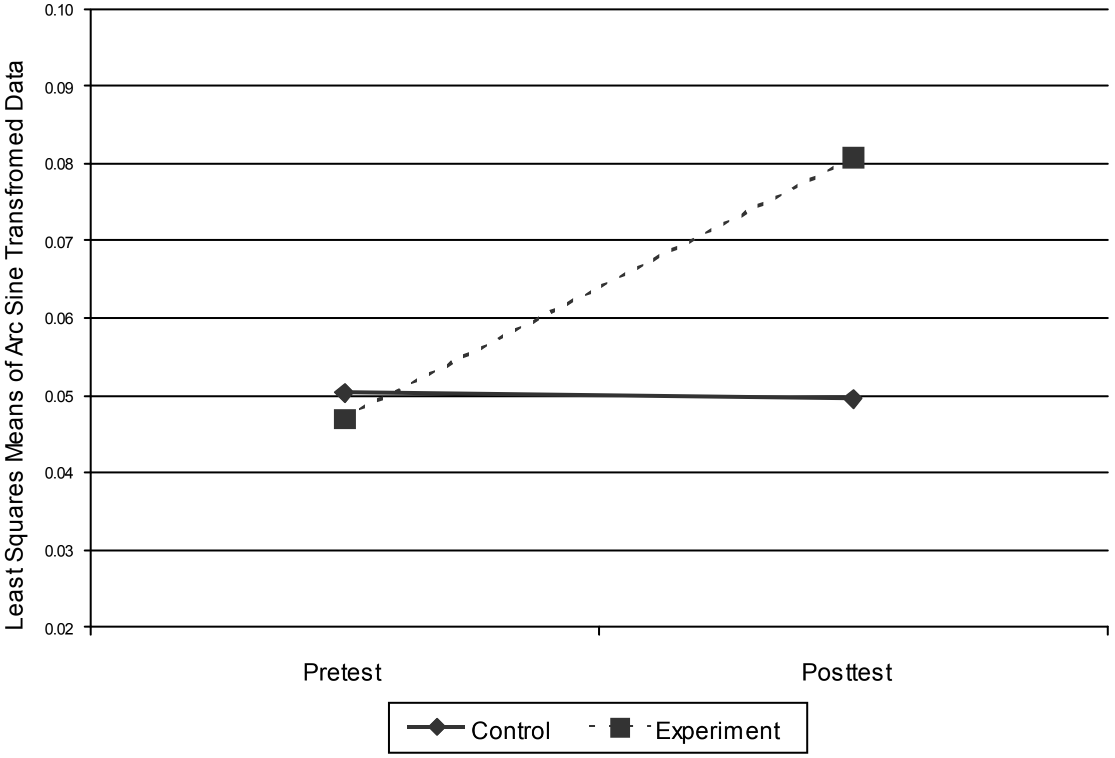

Figure 6. **Example of Disordinal Interaction of Group and Test on Least Square Means of Arc Sine Transformed Dynamic Knowledge Percentage Scores**

In this example, group, test, and their interaction are significant when addressing Dynamic Knowledge scores. When these variables are plotted, it is easy to visualize that although the control group's dynamic knowledge scores have remained relatively constant, the experiment group's (the folks who received the online instruction) scores increased significantly. A post hoc test would confirm the significant difference between the experiment group's post test and the other three cells of data.

Phase two analysis provides insight into whether any moderating variables might be influencing the cognitive effect of the online instruction. Using the model depicted in Figure 3, time (the variable "Test") is removed, and test scores are condensed into gain scores as described earlier. Next, demographic and other potential moderating variables are included. It is important to run correlation analysis (such as Pearson's) on all moderating variables to assess the possibility of multicollinearity. GLM is the preferred choice here again, particularly if some of independent variables are metric.

Interpreting the results are similar to phase one. First identify any significant variables (particularly interactions), perform a post hoc analysis, and plot significant variables. An example of the analysis and graphing is provided in Tables 3 and 4 and Figure 7.

Table 3

Factorial ANOVA Table for Arc Sine Transformed Dynamic Knowledge Gain Scores

Source	Df	SS	MS	F	P
Group	1	5.9E-02	9.4E-03	29.17	0.000
Age	3	3.7E-03	1.1E-03	3.38	0.019
Error	221	7.2E-02	3.2E-04		
Total	225	1.4E-01			

Table 4

Means and Standard Deviations for Dynamic Knowledge Gain Scores by Age and Group

Age

Group	<21			21 to 24			25 to 28			≥ 29		
	n	x	S.D.	n	x	S.D.	n	X	S.D.	n	x	S.D.
Control	29	0.1	0.07	77	0	0.06	2	0	0	2	0	0.05
Experiment	36	0.3	0.16	81	0.2	0.14	2	0.2	0.14	3	0.3	0.04
Total	65	0.2	0.17	158	0.1	0.14	4	0.1	0.13	5	0.2	0.16

CONCLUSION

Results from the model provide clear findings for interpretation. If the examples described in this paper were used in a real study, they would provide a researcher with a rich description of not only the cognitive abilities that were increased in learners but also which participant characteristics might be significantly moderating these increases.

This methodology for evaluating the effectiveness of online instruction is a departure from the numerous studies that have been done using comparative measurement methods. As stated at the beginning of the article, a comprehensive website documents hundreds of these studies and the challenges associated with them.

The benefits of performing an evaluation of online instruction in the manner described herein are numerous: The questionnaire is implemented as an objective multiple-choice format which is very convenient in current online instructional environments – such as WebCT and Blackboard; it can be scored automatically and thus minimizes data collection errors and collection time; and the step-by-step procedure and robust statistical models provide an easy to use and effective methodology.

This methodology has been evaluated by the Statistical Consulting Center at The Pennsylvania State University and implemented at two major research universities. I hope that this model will assist instructional system designers develop rubrics of assessment for their online courses and contribute to the body of literature on instructional systems assessment.

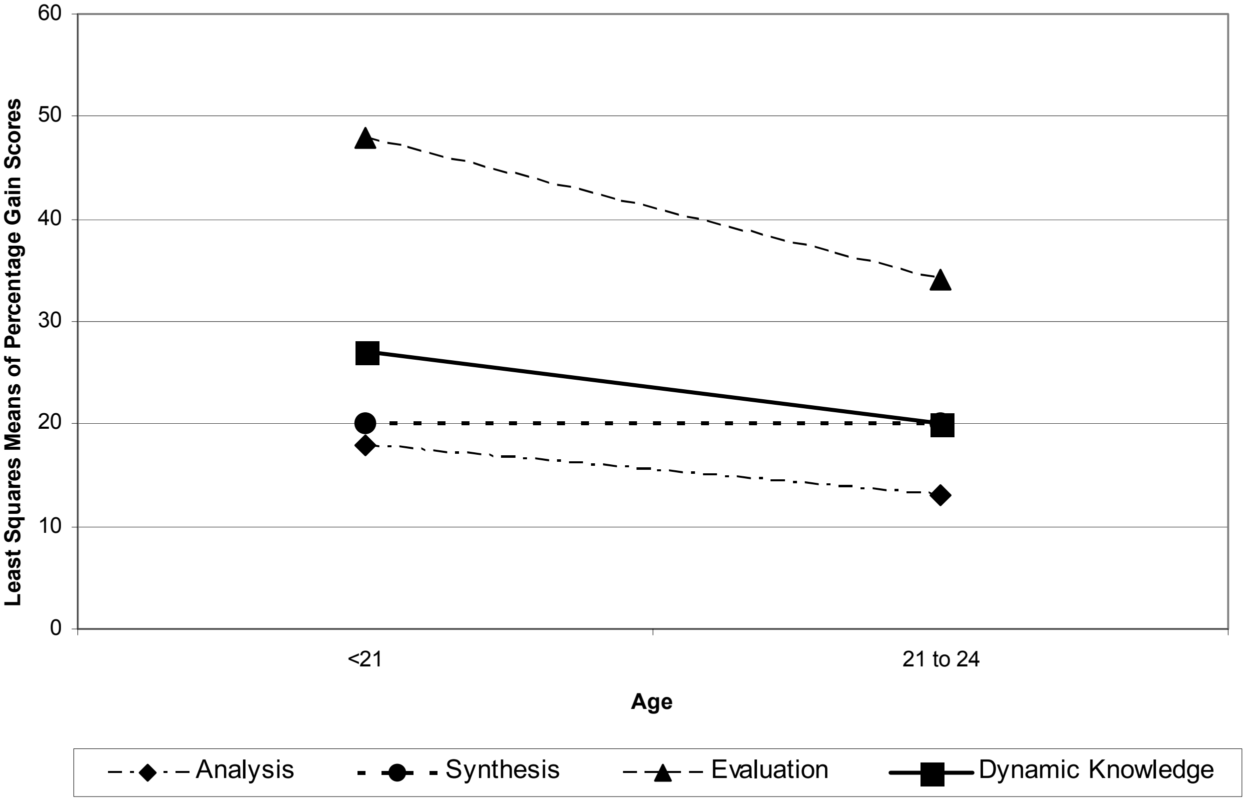

Figure 7. Age and Higher Order Cognitive Abilities on Least Squares Means Percentage Gain Scores

REFERENCES

Anderson, L.W., Krathwohl D.R., Airasian P.W., Cruikshank K.A., Mayer R.E., Pintrich P.R., Raths J., Wittrock M.C. (2001). *Taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Boston, MA: Pearson Allyn & Bacon

Babbie, E.R. (2003). *The practice of social research (10th Ed.).* Belmont, CA: Wadsworth.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives: Book one - cognitive domain*. New York, NY: Longman.

Branson, R.K. (1973). *Analysis and assessment of the state of the art in instructional technology*. Fort Monroe, VA: Army Training and Doctrine Command. (NTIS Document Reproduction Service No. AD A010 394; ERIC Document Reproduction Service No. ED 088 436).

Cook, T.D. & Campbell, D.T. (1979*). Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Dick, W. & Reiser, R.A. (1989). *Planning effective instruction*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Educational Testing Service (2003), *The official guide for GMAT review (10th Ed.)*. Princeton, NJ: Warner Books.

Gustafson, K. L., & Branch, R. (1997). Survey *of instructional development models, (3rd Ed.).* Syracuse, New York: ERIC Clearinghouse on Information and Technology, Syracuse University. (IR - 103).

Logan, R.S. (1982). *Instructional systems development: An international view of theory and practice*. Chicago, IL: Academic Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models. Chicago, IL: Irwin.*

Nikto, A. J. (2000). *Educational Assessment of students, (3rd Ed.*). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Russell. T.L. (1999). The *No Significant Difference Phenomenon.* Raleigh, NC: Office of Instructional Telecommunications, North Carolina State University.

Rossett, A. (1987). *Techniques in training and performance development series: Training needs assessment.* Englewood Cliffs, NJ: Educational Technology Publications.

Rossett, A. (2001). *The ASTD e-learning handbook : Best practices, strategies, and case studies for an emerging field.* New York, NY: McGraw-Hill.