Computerized Assessment Data Analysis Report

Prepared by Carol Cong Chen

July 2015

Data Analysis Report

Introduction

This report is designed to provide evidence of the usability of two computerized instructional modules in middle school science instruction in ecology. Specifically, this report aims to address two research questions:

1. How is the use of the computerized instructional modules related to student learning?
2. To what extent does evidence collected from the benchmark assessments support inferences about a student's proficiency (as determined by a standardized assessment)?

Methods

A quasi-experimental research design was applied in the study of using the computerized instructional modules in classrooms. Teachers had half of their classes assigned to a treatment group and the other half assigned to a control group. Classes in the treatment group took a standardized pre-test, followed by the ecology unit with the two instructional modules. After instruction, the students took a benchmark assessment (which is more proximal to the intervention) and a post-test, which was identical to the pre-test. Classes in the control group took the pre-test, the benchmark assessment, and the post-test, but had only the regular ecology unit, without the computerized instructional modules.

Data

The original data consisted of the performance of 615 students in 24 classes with six teachers on the two standardized assessments (pre- and post-tests) and one benchmark assessment. For both standardized assessments, a raw score and an IRT ability estimate were reported to each student. For the benchmark assessment, only a raw score was reported. Of the 615 students who participated in the study, 479 (78%) students had complete data for all three assessments, while 136 (22%) students had missing values; and 7 (1%) students chose to opt out of the assessment, which means that their data cannot be included in research. In this report, student data with missing values and opt out choices were removed, leaving a final data of 473

students. Of these, 303 (64%) students were in the treatment group and 170 (36%) students in the control group.

Data Analysis

Preliminary data analysis was conducted to provide a sense of how students in different groups perform on the three assessments using descriptive statistics, measures of central tendency and the two sample t-test. To address the first research question, two statistical analyses were conducted. 1) An independent sample t-test on the student learning gain score was performed to test whether students had made significant improvement on the post-test. 2) Analysis of covariance (ANCOVA) on the post-test measure, with the pre-test measure as the covariate, was performed to investigate the effects of using the computerized instructional modules on student learning after controlling for the effects of students' pre-test scores. To address the second research question, two statistical analyses were conducted. 1) Pearson correlation coefficients between benchmark assessment scores and post-test measures were calculated to analyses how they were related to each other. 2) A simple linear regression using the post-test measure as the dependent variable and the benchmark assessment score as the independent variable was conducted to assess to what extent the benchmark assessment related to a student proficiency's in standardized assessments.

All statistical analysis were computed by R-Studio software for Windows. The R script for running data analysis is also provided.

Results

Preliminary data analysis. Table 1 provides an overview of student performances on three assessments for both the treatment group and the control group, including descriptive statistics and two-sample t-test results. For all three assessments, the mean scores of the treatment group are higher than those of the control group while the standard deviations of the two groups are similar. Specifically, on average, the treatment group scored 2.86 higher than the control group on the pre-test, 4.45 higher on the benchmark assessment, and 2.97 higher on the post-test. The IRT mean ability estimate for the treatment group is 0.51 higher on the pre-test and 0.63 higher on the post-test than those of the control group. Further, a two-sample t-test was performed to compare whether the mean scores of the treatment group are significantly different from the

control group. The null hypothesis for this t-test is that the mean scores of the two groups are equal, assuming the data is normal distribution and the variances are the same. Table 1 shows that the two groups are significantly different on the pre-test, the benchmark assessment and the post-test at the .01 level.

Table 1

Group descriptive statistics and two Samples t-test on three assessment by groups

Student performance	**Group**	**N**	**Mean**	**SD**	**SE**	**Mean difference**	**df**	**t**	**P-value**
Pre-test score	Treatment	303	15.05	6.32	0.36	2.86	471	-5.05	<0.01**
	Control	170	12.19	5.05	0.39				
Pre-test IRT ability estimates	Treatment	303	0.05	1.17	0.07	0.51	471	-4.99	<0.01**
	Control	170	-0.46	0.89	0.07				
Benchmark assessment score	Treatment	303	29.69	6.36	0.37	4.45	471	-7.13	<0.01**
	Control	170	25.24	6.77	0.52				
Post-test score	Treatment	303	17.03	6.93	0.40	2.97	471	-4.67	<0.01**
	Control	170	14.06	6.01	0.46				
Post-test IRT ability estimate	Treatment	303	0.48	1.49	0.09	0.63	471	-4.85	<0.01**
	Control	170	-0.15	1.10	0.08				

**significantly different at the p<0.01 level

Research question one (How is the use of related to student learning?): First, an independent sample t-test was conducted to test whether students had made significant improvement on the post-test in terms of gains scores. The gain score was calculated by subtracting a student's pre-

test score/IRT ability estimate from those of the same student. Table 2 shows that students in the both groups have made significant improvement on their post-test ($p<0.01$). Specifically, students in the treatment group, on average, had gained 1.98 raw score and 0.43 IRT ability estimate, while students in the control group, on average, had gained 1.87 raw score and 0.31 IRT ability estimate. Then, a two-sample t-test was conducted on the mean gain scores of the two groups. Table 3 shows that the differences between the two mean gain scores is 0.11 ($t=-0.274$, $p=0.78$) for the raw score and is 0.12 ($t=-1.46$, $p=0.14$) for the IRT ability estimate, which means that the mean gain scores of the two groups are not significantly different.

Table 2

Independent Samples t-test: student gain scores by group

Group	**Measure**	**N**	**Mean Gain Score**	**df**	**t**	**P-value**
Treatment group	Raw score	303	1.98	302	8.69	<0.01**
	IRT ability estimates	303	0.43	302	8.54	<0.01**
Control group	Raw score	170	1.87	169	5.80	<0.01**
	IRT ability estimates	170	0.31	169	5.16	<0.01**

**significantly different at the $p<0.01$ level

Second, analysis of covariance (ANCOVA) was performed to control for the effects of student pre-test scores. Since the results of preliminary data analysis showed that the pre-test measures of the treatment group are significant higher than those of the control group ($p< 0.01$), students in the two groups are not in the same level before using the computerized instructional modules. Thus, ANCOVA, which can control the effects of student per-test scores is a better choice for analysis. Specifically, the post-test measures were used as dependent variables, the student group as the design factor and the pre-test measures as the covariate.

Table 3

Two Samples t-test: mean gain scores of treatment group and control group

Measure	**Treatment group mean gain score**	**Control group mean gain score**	**df**	**t**	**P-value**
Raw score	1.98	1.87	471	-0.274	0.78
IRT ability estimates	0.43	0.31	471	-1.46	0.14

Table 4

Analysis of covariance (ANCOVA): Post-test as dependent variable and Pre-test as covariate

Measure	**Source**	**df**	**Sum of square**	**Mean of square**	**F**	**P-value**
Raw score	Student group	1	955.2	955.2	59.63	<0.01**
	Student pre-test score	1	13071.3	13071.3	815.99	<0.01**
	Residual	470	7528.8	16		
IRT ability estimate	Student group	1	43.77	43.77	61.076	<0.01**
	Student pre-test score	1	537.94	537.94	750.70	<0.01**
	Residual	470	336.79	0.72		

**significantly different at the p<0.01 level

The results in Table 4 show that the covariate, student pre-test scores, significantly predicts student post-test scores (F=815.99, p <0.01), indicating that student post-test scores are influenced by their pre-test scores. Most importantly, there were significant differences between

the treatment group and the control group on student post-test scores after controlling for the effects of student pre-test scores ($F=59.63$, $p<0.01$). The results are similar when using the IRT ability estimate as the measure. The ANOVA results suggest that a student's improvement on the post-test is related to the group he or she belongs, and the treatment group outperformed the control group on the post-test given the same pre-test score. In other words, the use of the computerized instructional modules has a significant positive effect on student learning in ecology.

Research question two (To what extent does evidence collected from the benchmark assessments support inferences about a student's proficiency (as determined by a standardized assessment)? First, Pearson correlation coefficients between benchmark assessment scores and post-test measures were calculated. As shown in Table 5, student benchmark assessment scores and post-test raw scores were significantly correlated, $r=0.69$, $p< .05$; student benchmark assessment scores and post-test IRT ability estimates were also significantly correlated, $r=0.66$, $p< .05$.

Table 5

Pearson Correlation between Benchmark assessment scores and post-test measures

	Benchmark assessment score	Post-test raw score	Post-test IRT ability estimate
Benchmark assessment score	1.00	0.69*	0.66*
Post-test raw score		1.00	0.98*
Post-test IRT ability estimate			1.00

*significantly different at the $p<0.05$ level

Second, a simple linear regression analysis using the post-test measure as the dependent variables and the benchmark score as the independent variable was performed. Table 6 shows that the benchmark assessment scores significantly predicted student's post-test scores, $\beta=0.69$ ($t=20.90$, $p<.01$). The benchmark assessment scores also explained a significant proportion of variance in students' proficiency in the post-test, $R^2=0.48$, $p<.01$. The results were similar when using IRT ability estimate as the independent variable.

Table 6

Simple linear regression: mean gain scores of treatment group and control group

Measure	**Coefficients**	**Estimate**	**SE**	**t**	**P-value**	**R square**
	(Intercept)	-3.28	0.95	-3.46	<0.01**	0.48**
Raw score	Benchmark assessment score	0.69	0.03	20.90	<0.01**	
	(Intercept)	-3.51	0.20	-17.19	<0.01**	0.43**
IRT ability estimate	Benchmark assessment score	0.13	0.007	18.98	<0.01**	

**significantly different at the p<0.01 level

Conclusion

In summary, two conclusions were drawn from this data analysis report:

First, the use of the computerized instructional modules in middle school science instruction in ecology has a positive effect on student learning. Specifically, students who used the two computerized instruction modules performed significantly better on the pre-test, the post-test and the benchmark assessment than those who did not use these two computerized instruction modules.

Second, the benchmark assessment and the post-test are moderately correlated, and the benchmark assessment significantly predicted a student's proficiency on the post-test. Thus, the benchmark assessment could be considered as a part of the assessment system to support inferences about a student's proficiency.